High-Volume Hypothesis Testing for Large-Scale Web Log Analysis

Sana Malik
Human-Computer Interaction Lab
University of Maryland
College Park, MD 20742, USA
maliks@cs.umd.edu

Eunyee Koh
Adobe Research
San Jose, CA 95113, USA
eunyee@adobe.com

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892487

Abstract
Time-stamped event sequence data is being generated across many domains: shopping transactions, web traffic logs, medical histories, etc. Oftentimes, analysts are interested in comparing the similarities and differences between two or more groups of event sequences to better understand processes that lead to different outcomes (e.g., a customer did or did not make a purchase). CoCo is a visual analytics tool for Cohort Comparison that combines automated high-volume hypothesis testing (HVHT) with and interactive visualization and user interface for improved exploratory data analysis. This paper covers the first case study of CoCo for large-scale web log analysis and the challenges that arise when scaling a visual analytics tool to large datasets. The direct contributions of this paper are: (1) solutions to 7 challenges of scaling a visual analytics tool to larger datasets, and (2) a case study with three real-world analysts with these solutions implemented.

Author Keywords
Hypothesis testing; visual analytics; time-stamped event sequences; cohort comparison

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User interfaces

Key Terms

We define some key terms in the context of event sequence analysis.

Event type. The category of a time-stamped occurrence.

Event. A specific instance of an event type associated with a timestamp.

Record. All events in a single user's history.

Sequence. Two or more events.

Consecutive. A sequence of events uninterrupted by other events.

Concurrent. Two or more events that occur at exactly the same timestamp.

Cohort. A group of records.

Prevalence. The percent of records containing an event type or sequence.

Frequency. The number of times per record an event type or sequence occurs.

Introduction

Time-stamped event sequence data is generated across many domains. Business analysts use customer transaction histories to predict future purchases. Medical researchers and doctors use Electronic Health Records (EHRs) to find patterns in side effects among a group of patients. Oftentimes, analysts are interested in comparing two or more groups of event sequences to better understand the similarities and differences between them and how these histories affect outcomes.

With more complex data in larger volumes than ever, exploratory data analysis is becoming prevalent and visual analytics are increasingly important. Visualization plays an important role in allowing analysts to explore their data more organically, allowing them to discover things they did not know existed, easily identify anomalies, and generate hypotheses.

CoCo, for "Cohort Comparison," is a visual analytics tool for exploring results of high-volume hypothesis tests with an interactive interface, curated sorted and filtering, and a pre-defined taxonomy of metrics. Previous work introduced CoCo and evaluated its efficacy in the medical domain through case studies [3,10]. This paper covers the first case study of CoCo to demonstrate the value of high-volume hypothesis testing when paired with an interactive user interface in the web log analysis domain.

The direct contributions of this paper are:

1. Proposed solutions to 7 challenges when extending HVHT to larger datasets, and
2. A case study with real-world analysts using the solutions implemented in a visual analytics tool.

We begin by reviewing related work in event sequence analysis and visualization. We briefly describe CoCo. From there, we describe the challenges of scaling high-volume hypothesis testing for large-scale web logs and our iterative approach in solving these issues. Lastly, we implement these changes in CoCo and evaluate their efficacy in a case study with three analysts.

Related Work

Temporal Data Mining

Automated hypothesis testing is closely related to data mining. Traditional data mining algorithms include frequent sequence mining [7,9] and association rule (itemset) mining [1]. Most of these techniques mine sequences in a single dataset rather than compare differences and similarities across two datasets. While a data mining technique can be used in tandem to facilitate similar comparisons (e.g. comparing frequent sequence results between two datasets), more specialized methods are needed to answer questions about which sequences occur significantly differently between datasets? Bay and Pazzani introduce *contrast mining sets* [2], an algorithm for detecting differences between groups based on record attributes (i.e., age, gender, or occupation). In addition to record attributes, we look at differences in the event sequences themselves, based on both occurrence and timestamps.

Data mining algorithms are often a blackbox, allowing little user involvement during the process. Recent work has been done on interactive sequence mining [5,8,12,14], though these system focus primarily on mining frequent patterns in a single dataset, not differences between two datasets.

Overview of CoCo

CoCo is comprised of five main panels that enable users to systematically explore results of high-volume hypothesis tests: (1) A scattergram that displays all sequences found in the dataset and their occurrence in each of the cohorts provides a way to control sample size, (2) an aggregated overview of each of the cohorts allows users to see broad trends, (3) flexible methods for filtering and sorting the result set based on p-value or metric enable users to easily highlight potentially meaningful results, (4) a visualization to easily scan the result set, and (5) details-on-demand allow users to explore a selected result in more detail.

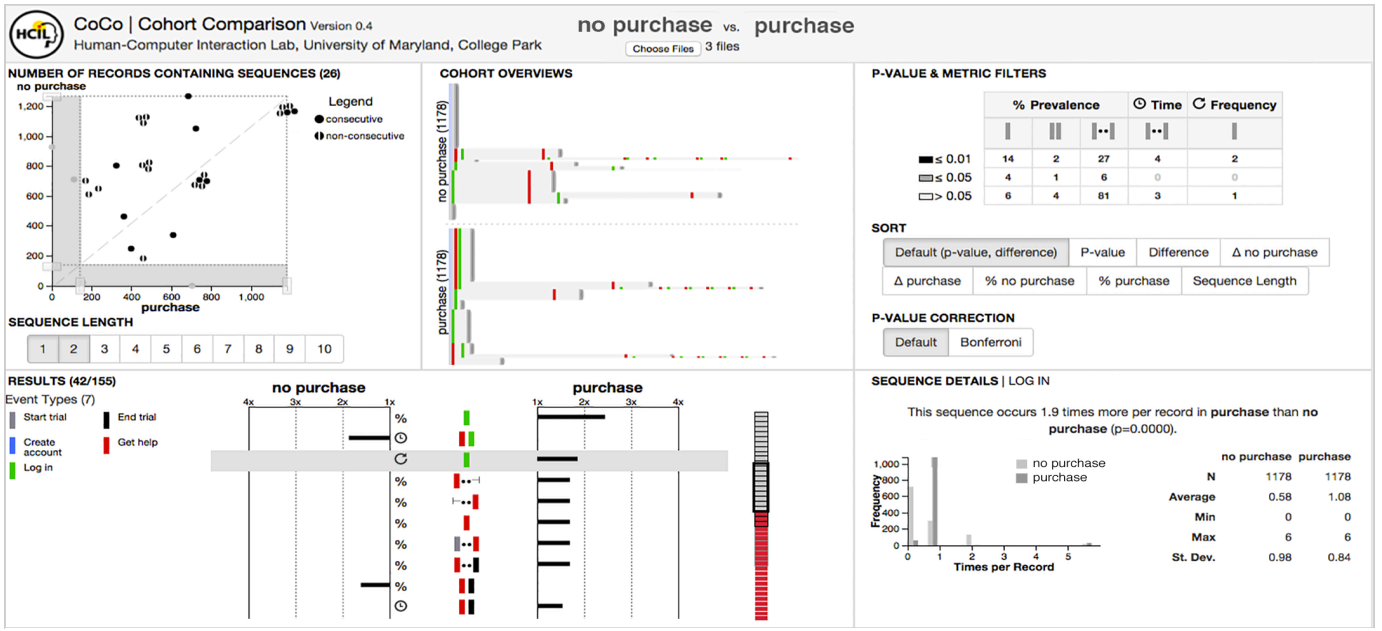

Figure 1. All data shown is synthetic. An extended description is given in the sidebar.

Event Sequence Visualization & Comparison

Gleicher et al. [6] provide an extensive survey of visual comparison techniques categorized into three metods: juxtaposition, superposition, and explicit encoding. Visualization tools have been designed for event sequence data [11,13], however there has been little research on visualizing event sequence comparisons.

Zhao et al. [16] design MatrixWave to compare the flow of users in clickstream datasets. MatrixWave focuses on differences in the occurrence of immediate, pairwise steps in the event stream, whereas we generalize to differences in sequences of any length, as well as differences dealing with time. Vrotsou et al. [14] introduce a set of event sequence similarity metrics; we focus on difference metrics, but believe our work can applied to these metrics as well. Towards large-scale event analytics and visualization, Wongsuphasawat and Lin [15] develop two interactive visualizations for common analysis tasks and Du et al. [4] propose 14 strategies for coping with volume and variety in large event sequence datasets.

Overview of System

In this section, we provide a brief overview of the CoCo's metrics and interface (Figure 1). A complete taxonomy is discussed in previous work [10], in addition to the motivation for the interface design.

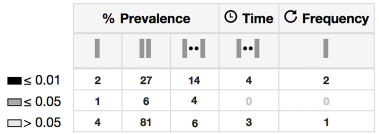

Figure 2. A table of the p-value distribution across metrics provides a way for analysts to understand if the dataset contains more than expected false positives. The table also serves as a filter, where users can click to remove a p-value group or a metric group.

CoCo includes five metrics in its current version: prevalence of events, sequences, and non-consecutive event pairs; duration of consecutive event pairs; and frequency of events.

Interface
Top, left: A scatterplot of all sequences, colored by consecutive or non-consecutive. Each axis is the number of records that contain the record in alpha cohort and beta cohort.
Top, center: Aggregated EventFlow [11] overviews.
Top, right: Methods for filtering and sorting the result set. Filtering is possible by event type, record coverage, significance, metric type, or sequence type. Users can sort results based by ratio and significance, significance only, ratio only, alpha value, or beta value. Users may use the default p-values or apply a correction.
Bottom, left: The results are displayed in a ranked list. Each row represents a result, and the hypothesis is shown in the middle. Icons to the left indicate metric type: Percent (%) for record coverage, clock for time, and round arrow for frequency. The sequence is shown to the right of the metric. Each bar represents an event, colored according to the legend (left). A bar grows to the left or right indicating the ratio of values between the two groups. The bars are colored by significance (black is $p < 0.01$).
Bottom, right: Details-on-demand of the result are shown to the right when a user clicks a specific event. For average metrics (e.g., time or frequency), users can see a distribution of the value in each cohort.

Scaling HVHT to Larger Datasets

Previous versions of CoCo had only been used on relatively small datasets of up to 2,000 records per cohort and up to 50 event types. Web log datasets, on the other hand, record millions of users who access the website per day and hundreds of clickstream events per user. The increased volume of records and variety of event types presented new challenges for CoCo on both the front- and back-ends. Through an iterative process over six weeks, analysts highlighted challenges they faced with CoCo's current implementation, we proposed solutions, implemented them into CoCo, and received further feedback from analysts about these solutions and additional challenges. We cover the seven major challenges and their solutions, divided into two main areas: efficiency on the backend (1—3) and the analytical process on the frontend (4—7).

Challenge 1: Long Wait Times for Computation

Long wait times can cause an analyst to lose concentration and incur more time recalling their task. In an effort to minimize long waits, we implemented two timesaving changes.

The original version of CoCo counted every sequence that appeared in the loaded datasets. However, in the clickstream data, there were some records that had as many as 320 events. Based on observations of the previous three case studies, we found the analysts often did not look at results of sequences of length more than four. Typically, longer sequences were more obscure and analysts were not able to derive meaningful insights from them. We chose to implement a sequence length limit of 10, to be adequately long enough for the longer sequences found in clickstreams. In the use of this new limited version, analysts still looked mostly at sequences of length 4—5 at most, so there was no need to extend the range beyond length 10. We did not further reduce the length limit because performance at this stage was reasonable. Limiting the

sequence length offered a speed up of about 15x. If future datasets would benefit from a shorter limit, we will leave determining the ideal limit for future work.

Challenge 2: Browser data transfer Limitations

Larger datasets require more hypotheses to be tested, thus larger result sets to return to the browser. Due to some browser limitations, it is not possible to send data over a certain size. Thus, to reduce the volume of the result set, we automatically filter those sequences that occur in less than 1% of the records.

Challenge 3: Precision Loss with Corrections

When a statistical correction, such as Bonferroni Corrections, is applied, the p-values are divided by the number of hypotheses that are tested. This results in very small values. Previously, CoCo showed p-value precision up to two decimal points, but now shows four.

Challenge 4: Visualizing Both Cohorts

With a large dataset, an analyst may not know what his or her data looks like. We embedded EventFlow [11] displays for each cohort to provide this overview. EventFlow was chosen because many of our users are familiar with it, and its aggregate display provides an overview of the most frequent patterns across the cohorts in a compact view that will scale to large datasets without using more space.

Challenge 5: Chance of False Positives is Increased

We highlight the potential for false positives by providing the distribution of p-values to the user in a filterable table (Figure 2). Two statistical experts that we consulted with suggested this, because with any statistical test that is applied many times to a single dataset, there is some likelihood of false positives. By providing the user the distribution of the resulting p-values, the user can see if the actual distribution of p-values is what would be expected by random chance or if it is in fact affected by the content of the dataset.

Challenge 6: Large Number of Event Types

By default, CoCo starts by showing only the results for single event types (sequences length 1) so analysts can make informed decisions about which events occur frequently and which might be important to the analysis. After determining if any events can be dismissed, analysts can filter out those events that they deem unrelated or unimportant to their questions.

Challenge 7: Beginning Analysis is Daunting

With hundreds of thousands of hypothesis results, it might be daunting for analysts to know where to start with their analysis. To simplify this process, we suggest two methods.

First, we recommend a process model and arrange the layout to match this process. We rearranged the panels on CoCo to suggest the order in which analysts should explore their dataset. We first provide methods for seeing an overview of the all the data (scattergram and cohort overviews) on the top left, followed by more detailed views of the result set. Controls for filtering and sorting this list are prominently displayed on the top right.

Second, we provide default values for all filters and sorting methods. While these filters are customizable, the default values provide the simplest starting point for the users. It is important that the default values are carefully chosen. For example, for sequence length, we decided to start with length 1, since users are often

overwhelmed after looking for at the long results list. Starting with length 1 allows users to get a bearing on the events in their cohorts and allowing them to choose when they are ready to move onto the next result set.

Case Study

Three analysts used CoCo with a real-world event sequence dataset. One analyst was an experienced user of CoCo and two were novice users. The dataset contained users' events on a product website, such as viewing the display ads, signing up for promotions or free trials, and purchasing products, and all three analysts used the same dataset to compare the group of users who purchased the products without using trials versus with using product trials. In particular, the analysts explored the occurrence of the display ads and retargeting events (e.g., an ad for a product the user has already viewed).

By exploring events that are statistically significant in the result panel, analysts found one group viewed display ads more than the other group, and that group also contained more retargeting events than the other group. By investigating more on other events such as product trial and adoption using CoCo, analysts hypothesized that the first group, who viewed the display ads more, seemed fairly new to the websites' product offerings ("explorers") while the other group, who were exposed to fewer display ads and retargeting, seemed to have good knowledge about the websites' products and offerings ("experienced users").

Since the datasets contained many events (over 150), the analysts found the event filtering most helpful and they were able to focus the analysis on specific events. In addition, the reduced metric calculation time provided a much better user experience for data analysis, as the analysts did not need to wait for CoCo to load data and finish hypothesis testing before they could begin their explorations. Analysts all mentioned that the results were a bit linear, and they would prefer to have more freeform exploration and interactions to explore individual sequences.

Conclusions and Future Work

This work explores the applicability of high-volume hypothesis testing (HVHT) to large-scale web logs. Through an iterative design process, we identify 7 challenges for extending a visual analytics tool, CoCo, to larger datasets and proposed solutions. Though we apply these solutions to a single analytics tool, we feel this can be extended to any tool that involves a large number of computations and statistical uncertainty.

Providing more control over finding the sequences and allowing users to select sequences of interests would provide a more freeform analysis. Additionally, there are a limited amount of colors that can be used for the event types, which may cause ambiguity without the use or patterns, borders, or shapes. Lastly, we use a specific set of metrics for event sequence comparison, but the interface could be extended to display results of more traditional data mining techniques or metrics for datasets besides event sequences (e.g., networks).

Acknowledgements

The authors would like to thank Ben Shneiderman, Catherine Plaisant, and Fan Du for their feedback during the design and implementation of CoCo. We also gratefully acknowledge the support of Adobe, Oracle, and the University of Maryland's Center for Health-related Informatics and Bioimaging.

References

1. Rakesh Agrawal, Tomasz Imieli'nski, and Arun Swami. 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM International Conference on Management of Data*, ACM, 207–216.
2. Stephen D Bay and Michael J Pazzani. 2001. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery* 5, 3: 213–246.
3. M Bjarnadottir, S Malik, E Onukwugha, T Gooden, and C Plaisant. 2015. Understanding Adherence and Prescription Patterns Using Large Scale Claims Data. *PharmacoEconomics*.
4. Fan Du, Ben Shneiderman, Catherine Plaisant, Sana Malik, and Adam Perer. 2016. Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. *Visualization and Computer Graphics, IEEE Transactions on*.
5. Paolo Federico, Jürgen Unger, Albert Amor-Amorós, Lucia Sacchi, Denis Klimov, and Silvia Miksch. 2015. Gnaeus: Utilizing clinical guidelines for knowledge-assisted visualisation of EHR cohorts. *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA '15)*, The Eurographics Association. http://doi.org/10.2312/eurova.20151108
6. Michael Gleicher, Danielle Albers, Rick Walker, I. Jusufi, C. D. Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4: 289–309. http://doi.org/10.1177/1473871611416549
7. Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery* 15, 1: 55–86. http://doi.org/10.1007/s10618-006-0059-1
8. Tim Lammarsch, Wolfgang Aigner, Alessio Bertone, Silvia Miksch, and Alexander Rind. 2014. Special Section on Visual Analytics: Mind the time: Unleashing temporal aspects in pattern discovery. *Computer Graphics* 38: 38–50. http://doi.org/10.1016/j.cag.2013.10.007
9. Nizar R Mabroukeh and Christie I Ezeife. 2010. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys* 43, 1: 3:1–3:41. http://doi.org/10.1145/1824795.1824798
10. Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. 2015. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM, 38–49. http://doi.org/10.1145/2678025.2701407
11. Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. 2013. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics* 19, 12: 2227–2236.
12. Adam Perer and Fei Wang. 2014. Frequence: Interactive mining and visualization of temporal frequent event sequences. *Proceedings of the 19th International Conference on Intelligent User Interfaces*, ACM, 153–162. http://doi.org/10.1145/2557500.2557508
13. Charles D Stolper, Adam Perer, and David Gotz. 2014. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *{IEEE} Trans. Vis. Comput. Graph.*, 1653–1662. http://doi.org/10.1109/TVCG.2014.2346574
14. Katerina Vrotsou and Aida Nordman. 2014. Interactive visual sequence mining based on pattern-growth. *2014 IEEE Conference on Visual*

Analytics Science and Technology (VAST '14), 285–286. http://doi.org/10.1109/VAST.2014.7042532

15. Krist Wongsuphasawat and Jimmy Lin. 2014. Using Visualizations to Monitor Changes and Harvest Insights from a Global-Scale Logging Infrastructure at Twitter. *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*: 113–122.

16. Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. 2015. MatrixWave: Visual comparison of event sequence data. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 259–268. http://doi.org/10.1145/2702123.2702419

For more information about CoCo, please visit: http://hcil.umd.edu/coco.