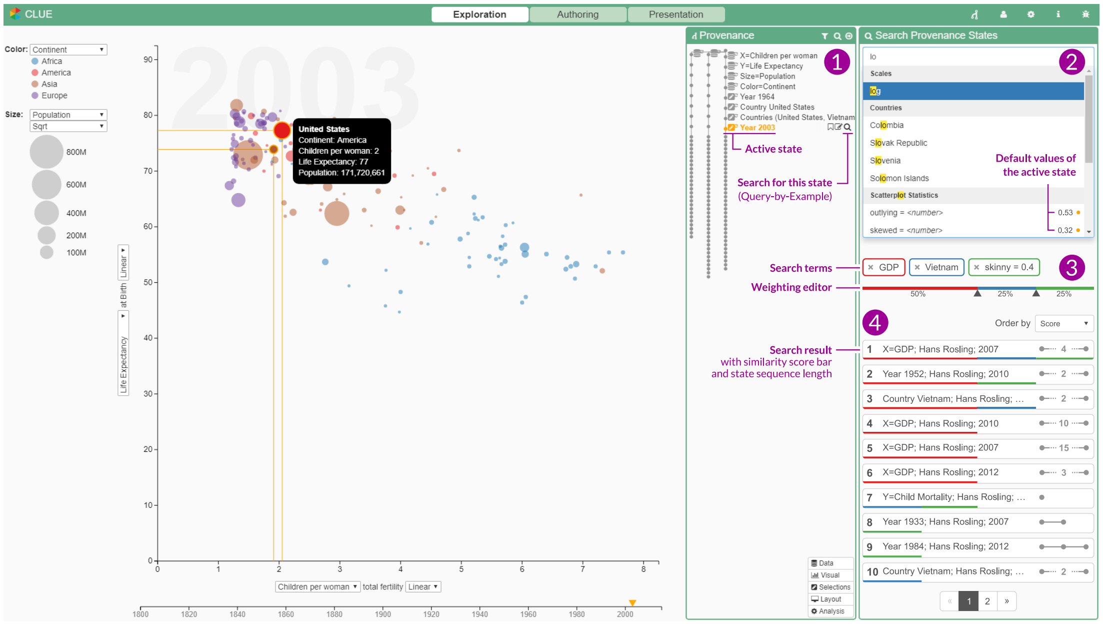
Provenance-Based Visualization Retrieval

Holger Stitz[*]
Johannes Kepler University Linz

Samuel Gratzl[†]
Johannes Kepler University Linz

Harald Piringer[‡]
VRVis

Marc Streit[§]
Johannes Kepler University Linz



United States
Continent: America
Children per woman: 2
Life Expectancy: 77
Population: 171,720,661

1 X=GDP; Hans Rosling; 2007 4
2 Year 1952; Hans Rosling; 2010 2
3 Country Vietnam; Hans Rosling; … 2
4 X=GDP; Hans Rosling; 2010 10
5 X=GDP; Hans Rosling; 2007 15
6 X=GDP; Hans Rosling; 2012 3
7 Y=Child Mortality; Hans Rosling; …
8 Year 1933; Hans Rosling; 2007
9 Year 1984; Hans Rosling; 2012
10 Country Vietnam; Hans Rosling; … 2

Figure 1: Mockup of a Gapminder-inspired prototype with (1) a captured provenance graph of three stories from Hans Rosling's presentations. The user can query the provenance graph for visualization states using (2) the search field, which suggests visualization properties while typing. Properties of the active state are marked and provide a default value, if available. (3) The selected search terms can be weighted based on the user's interest. (4) The search results are ranked by the state's similarity score.

Abstract

Storing interaction provenance generates a knowledge base with a large potential for recalling previous results and guiding the user in future analyses. However, search and retrieval of analysis states can become tedious without extensive creation of meta-information by the user. In this work we present an approach for an efficient retrieval of analysis states which are structured as provenance graphs of automatically recorded user interactions and visualizations. As a core component, we describe a visual interface for querying and exploring analysis states based on their similarity to a partial definition of the requested analysis state. Depending on the use case, this definition may be provided explicitly by the user or inferred from a reference state. We explain the definition by means of a Gapminder-inspired prototype and discuss our implementation for an effective retrieval of previous states.

Keywords: Provenance, Retrieval, Visual Analytics

1 Introduction

Provenance information plays a vital role to ensure reproducibility of analysis results. Over time, this information generates a knowledge base that can be used for recalling previous results or guiding the user in future analyses. In order to fulfill both tasks, an efficient search and retrieval of previous visualization states is necessary.

We started to develop the retrieval approach to target real-world use cases from our collaboration partners, who are researchers at a pharmaceutical company. They aim to discover cancer genes that can be targeted with future drugs. For the data-driven drug discovery, the analysts use a specialized visual analysis software that records all user interactions as provenance graphs. In such scenarios, analysts work in distributed teams and perform analyses using the same software. Hence, it is likely that an analyst wants to know if she or a colleague has already investigated the same or a similar set of genes in earlier analyses. Using our approach, the analyst can find similar visualization states by performing a retrieval based on her colleagues' as well as her own provenance information. The search query can be either built by definition (i.e., data attributes, selected items, etc.) or by example (i.e., from an active visualization state). Further, all query aspects can be weighted based on the user's interest. We argue that this approach has potential to minimize redundant work as well as to accelerate the analysis process.

Our primary contribution is a novel provenance-based retrieval process that is complemented by a set of visual interfaces. In the following, we use the Gapminder-inspired prototype from Gratzl et. al. [1] as guiding example. The loaded provenance graph contains all interactions Hans Rosling performed with the original Gapminder software during three different presentations [2–4].

2 Related Work

Many research efforts focus on the creation, analysis, and visualization of provenance graphs. Heer et al. [5] employ the provenance

[*] e-mail: holger.stitz@jku.at
[†] e-mail: samuel.gratzl@jku.at
[‡] e-mail: hp@vrvis.at
[§] e-mail: marc.streit@jku.at

graph ("worksheet history") of *Tableau* for retrieval tasks. The analyst can search for visualizations and filter results by visualization type (e.g., bar chart, line chart) or data attributes. A limitation of this approach is the Boolean search for visualizations where a state is found only if all attributes match. In contrast, our retrieval approach provides a fuzzy search for multiple terms and allows weighting of different search aspects based on the user's interest. We also employ the metadata, such as author and timestamp, to retrieve states from earlier analysis sessions contained in the provenance graph.

3 Provenance-Based Retrieval Approach

Recording all user interactions during a visual exploration session results in an interaction provenance graph that contains a list of actions as edges and the corresponding visualization states as nodes. Each visualization state is defined by properties that can be distinguished by their data type. (1) **Categorical properties** include, for example, displayed data attributes (e.g., GDP, population) and categorical visualization settings, such as axis scales (linear, logarithmic). (2) **Numerical properties** include derived visualization metrics, e.g., *scagnostics* [7] for scatter plots. (3) **Set-typed properties** typically refer to selections of data items, e.g., to search for analysis states where particular countries have been brushed.

3.1 Index and Retrieval Mechanism

To discover similar states, we compare each search term to the properties of the visualization state, resulting in a comparison score. Subsequently, we calculate a weighted sum that is used as a similarity score describing each state. This ensures that states matching all search terms result in a higher rank. Since properties might remain unchanged for multiple subsequent visualization states, we cluster them into state sequences to provide more meaningful search results.

Based on the property type, we apply different index and retrieval mechanisms. For categorical properties, we index the values using a *term frequency–inverse document frequency* (*tf–idf*) approach [6]. For retrieval, the *tf–idf* score for the search term is used as comparison score. For retrieving numerical properties, the absolute delta between the query value and the state value is used as comparison score. For set-typed properties, we use the *Jaccard index* between the query set and the state set as comparison score.

3.2 Visualizations and Interaction Design

The user interface (see Fig. 1) consists of two views: The *provenance view* (1) provides a scalable visualization of all recorded states [1] while the *search view* contains a search field (2), selected search terms as query, a weighting editor (3), and a list of search results (4).

The search field supports the user in the selection of search terms by highlighting matching property names (e.g., "data attributes", "scagnostics") and property values (e.g., "GDP", "clumpy") while typing. Properties that do not match the current search term are hidden. We also support query-by-example in two ways: First, the user can add all properties of the active state (e.g., all displayed data attributes) as search terms by clicking the search button next to the state in the provenance view. Second, properties of the active state are marked in the search suggestions, where they can be spotted easily. In case of numeric property, the default value is equal to the one of the active state, but can be overridden by the user. This approach allows the user to find either exact or similar visualization states with respect to the active state.

Added search terms can be weighted using the editor. By default, the weights are equally distributed among all search terms. The user can flexibly distribute the weight by dragging the sliders.

The search results are updated automatically and can be ranked by the similarity score or the state sequence length. Each result item shows the rank, the similarity scores of matching search terms as bars, metadata, such as author and time of the first match, and the state sequence length. When hovering over a search result, the sequence of states is highlighted in the provenance view. Selecting a search result loads the first state of the sequence.

4 Discussion

Provenance Graph Characteristics Each interaction leads to a new visualization state in the provenance graph. Due to temporal coherence, adjacent states of the same interaction sequence are typically very similar. In case of a plain retrieval, all these similar states would be listed as separate search results, cluttering the result list. We address this issue by clustering the search results and presenting the first state of a sequence with the number of subsequent states that match the search terms.

Relationship of Visualization Properties The *tf–idf* retrieval does not consider the relationship among search terms. For example, searching for a state with an axis that contains both GDP and logarithmic scale as attributes is not possible. Instead, it finds states that contain the terms in arbitrary order (i.e., *bag of words*). Building an *n-gram* model, which stores the spatial information of terms, might be a possible solution. However, defining the correct word order is challenging (e.g., "GDP logarithmic" vs. "logarithmic GDP") and would increase the index size, due to the storage of all permutations of a tuple. A more sophisticated solution requires the definition of an ontology between possible terms (e.g., that an axis contains attributes and a scale) and an advanced notation of the search query to express these relationships.

Generalization The proposed retrieval approach can be applied to any interaction provenance graph and only requires the visualization properties stored in the visualization state of the provenance graph. However, it remains open and has to be evaluated if the retrieval approach also works for multi-coordinated view setups.

5 Conclusion and Future Work

In this poster we presented an approach to effectively search for visualization states in interaction provenance graphs by applying different index and retrieval mechanisms to the visualization properties and calculating a user-defined weighted sum for all query terms. In future work, we would like to extend our approach to provenance graphs of other visual analytic tools and perform a user study with our collaborators.

Acknowledgments

This work was supported in part by Boehringer Ingelheim Regional Center Vienna, the Austrian Science Fund (FWF P27975-NBL), and the State of Upper Austria (FFG 851460). The VRVis Forschungs-GmbH is funded by COMET – Competence Centers for Excellent Technologies (854174) by BMVIT, BMWFW, Styria, Styrian Business Promotion Agency – SFG and Vienna Business Agency. The COMET Programme is managed by FFG.

References

[1] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit. From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum*, 35(3):491–500, 2016.

[2] Hans Rosling. The best stats you've ever seen. `https://youtu.be/hVimVzgtD6w`, 2007. Accessed: 2017-06-08.

[3] Hans Rosling. 200 Countries, 200 Years, 4 Minutes. `https://youtu.be/jbkSRLYSojo`, 2010. Accessed: 2017-06-08.

[4] Hans Rosling. Religions and babies. `https://youtu.be/ezVk1ahRF78`, 2012. Accessed: 2017-06-08.

[5] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Trans. on Vis. and CG (InfoVis '08)*, 14(6):1189–1196, 2008.

[6] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[7] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization*, pp. 157–164, 2005.