Provenance-Based Visualization Retrieval

Holger Stitz, Samuel Gratzl, Harald Piringer, and Marc Streit

JOHANNES KEPLER UNIVERSITY LINZ

vrvis

Best Poster Award

Figure 1: *Mockup of a Gapminder-inspired prototype with (1) a captured provenance graph of three stories from Hans Rosling's presentations [1-3]. The user can query the provenance graph for visualization states using (2) the search field, which sugests 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.*

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. In the following, we use the Gapminder-inspired prototype from [4] as guiding example (see Fig. 1).

Motivation

Our collaboration partners, who are researchers at a pharmaceutical company, 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.

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 [4] 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 with highlight and autocomplete of property names and values. We also support query-by-example in two ways:

- 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.
- Properties of the active state are marked in the search suggestions, where they can be spotted easily.

Added search terms can be weighted interactively by dragging the sliders in the editor.

The search results are updated automatically. 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.

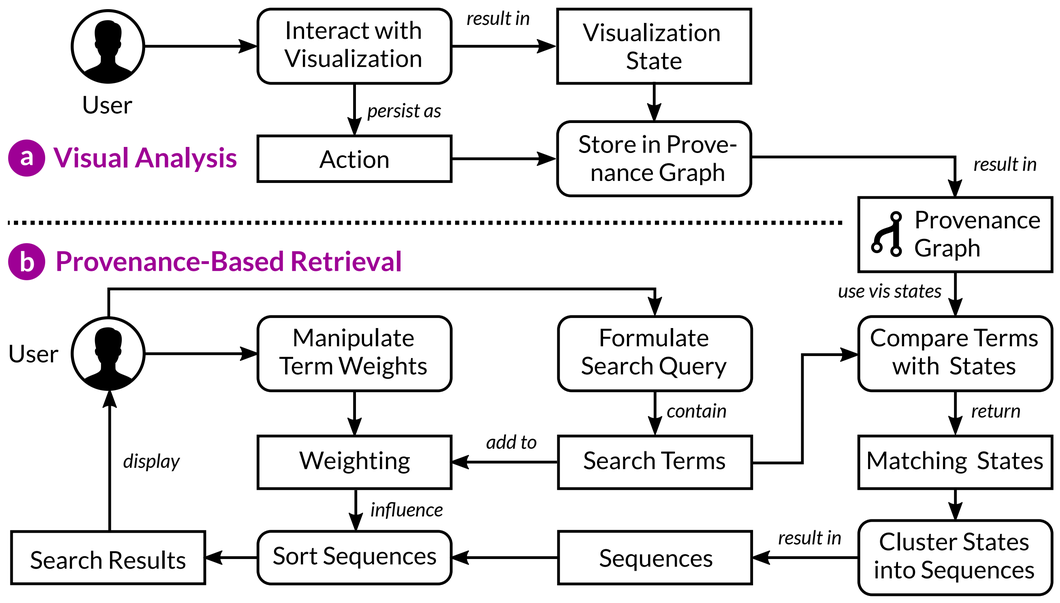

Figure 2: *The retrieval approach consists of two phases:*

(a) User interactions change the visualization and result in new visualization states, which are added to the provenance graph.

(b) The user formulates a search query that is compared to the prop-erties of visualization states from the provenance graph. Matching states are clustered into sequences and sorted according to the weighted sum of all search terms (i.e., the similarity score of the state).

a Visual Analysis

Recording all user interactions results in an interaction provenance graph that contains a list of actions as edges and the corresponding visualization states as nodes (see Fig. 2a). Each visualization state is defined by properties that can be distinguished by their data type:

- **Categorical properties** include data attributes (e.g., GDP) and categorical visualization settings (e.g., axis scales).
- **Numerical properties** include derived visualization metrics, e.g., scagnostics for scatter plots.
- **Set-typed properties** typically refer to selections of data items, e.g., to search for analysis states where particular countries have been brushed.

Implementation

The web application is based on the Phovea Platform.
Demo: **vistories.org/gapminder-retrieval-poster**

b Provenance-Based Retrieval

To discover similar states, we compare each search term to the properties of the visualization state (see Fig. 2b), 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:

- **Categorical values** are indexed using a *term frequency–inverse document frequency* (*tf–idf*) approach [5]. For retrieval, the *tf–idf* score for the search term is used.
- For **numerical properties** the absolute delta between the query value and the state value is used.
- **Set-typed properties** the *Jaccard index* between the query set and the state set is used.

[1] Hans Rosling. The best stats you've ever seen. https://youtu.be/hVimVzgtD6w, 2007.
[2] Hans Rosling. 200 Countries, 200 Years, 4 Minutes. https://youtu.be/jbkSRLYSojo, 2010.
[3] Hans Rosling. Religions and babies. https://youtu.be/ezVk1ahRF78, 2012.
[4] 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.
[5] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975.

Contact
holger.stitz@jku.at
caleydo.org

Acknowledgements
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.

FWF Der Wissenschaftsfonds. FFG CALEYDO