탐색적 데이터 분석과정의 복기 및 재사용을 위한 사용자 인터랙션 프로비넌스 시각화

김영택[01] 김이은[2] 송현주[3] 서진욱[1]
[1]서울대학교 컴퓨터공학부, [2]삼성전자, [3]숭실대학교 컴퓨터학부
ytaek.kim@hcil.snu.ac.kr, yieun12.kim@samsung, hsong@ssu.ac.kr, jseo@snu.ac.kr

Interactive Provenance Visualization for Recovery and Reuse of User Interaction during Exploratory Data Analysis

Youngtaek Kim[01] Yieun Kim[2] Hyunjoo Song[3] Jinwook Seo[1]
[1]Department of Computer Science and Engineering, Seoul National University
[2]Samsung Research, Samsung Electronics
[3]School of Computer Science and Engineering, Soongsil University

요 약

탐색적 데이터 분석의 결과를 명확히 이해하기 위해서는, 해당 결과에 도달하기까지의 분석 과정을 추론해 볼 필요가 있다. 시각화 도구를 활용한 분석 (Visual Analytics) 과정에서는, 사용자의 인터랙션 프로비넌스 (Provenance) 탐색을 통하여 결과의 통찰을 얻기까지의 과정을 확인해 볼 수 있다. 기존 시각화 툴들은 분석 히스토리를 선형적으로 탐색할 수 있도록 지원하고 있지만, 주로 원시적인 액션을 기반으로 하여 오버뷰를 얻거나, 그 의도를 파악하기에는 어려움이 있었다. 이 논문에서는 프로비넌스 데이터의 효율적인 탐색을 위한 기법 및 시각화 도구를 제안하였다. 먼저, 사용자 인터랙션을 구조적으로 인코딩하고, 추상화할 수 있도록 하였다. 그리고 이를 기반으로 그래프 형태로 프로비넌스를 복기하고, 재사용할 수 있는 웹 기반 프로토타입을 구현하였다.

1. 서 론

탐색적 분석의 과정을 이해하는 일은, 분석의 최종 결과를 이해하는 일만큼 중요하다. 이는 분석 과정 자체가 개별 분석자들이 발견한 통찰뿐만 아니라, 어떻게 해서 그 통찰에 도달했는지에 대한 추론 과정까지 포함하고 있기 때문이다. 시각화 도구를 활용한 분석 과정에서는, 사용자의 인터랙션 프로비넌스를 통하여 분석 과정을 살펴볼 수 있다.

널리 활용되는 대표적인 분석 툴인 Tableau, Spotfire는 분석 과정에서 발생하는 사용자의 인터랙션을 저장하여, 그 히스토리를 확인할 수 있도록 하고 있다. 또한, 히스토리의 revision을 저장하고 다시 불러올 수 있는 기능(Restore)을 추가적으로 제공하고 있다. 하지만 해당 히스토리는 단순히 시간순의 선형적인 탐색을 위주로 하며, 가장 원시적인 사용자 액션을 저장하기 때문에, 프로비넌스를 효율적으로 탐색하고, 해석하기에 어려움이 존재한다. Heer [6]는 분석 히스토리를 그래프 형태로 나타냈고, 이를 사용자 액션의 타입과 타이밍에 따라 chunk시켜 표현하였다. 하지만 이 역시 사용자의 원시적 액션 단위로 프로비넌스를 분석해야 하며, 오버뷰를 얻기가 용이하지 않다는 단점이 존재한다.

프로비넌스의 효율적인 분석을 위해서는 사용자의 인터랙션을 추상화하는 단계들이 필요하다. M. Pohl은 사용자 인터랙션의 시퀀스를 분석하여 select, explore, reconfigure, encode, abstract/elaborate, filter and connect의 카테고리로 분류하였다 [3]. 그리고, Bors는 인지적 업무 분석을 이용하여 프로비넌스를 고레벨과 저레벨 2단계로 나눠서 분석할 수 있는 기법을 제안하였다 [4].

이 논문에서는 프로비넌스 데이터의 효율적인 분석을 위한 기법 및 시각화 도구를 제안한다. 먼저, 사용자 인터랙션을 구조적으로 인코딩하고, 추상화할 수 있도록 하였다. 그리고 이를 기반으로 그래프 형태로 프로비넌스를 복기하고, 재사용할 수 있는 웹 기반 프로토타입 VRR을 구현하였다.

* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2019R1A2C2089062).

A

...
열, 행에 데이터 삽입
time series 변경
select
select
back
열 email 행 id(count)
정렬
색상 변경 (Mark)
레이블 달기
연도별 필터링 결고
시트 2개 병치해서 보기
다 지우고 다시 시작
Top 3 선택
열에 연도, 행에 카운트
...

B

pid	id	category	interaction	parameters
0	1	Arrange:Chart	barchart	x=email, y=count(id)
1	2	Arrange:Order	orderBy	count(id)
2	3	Map:Color	changeColor	blue
3	4	Map:Label	addLabel	label=data
4	5	Manipulate:Select	brush	email=["mbostock@gmail.com", "jason@jasondavies.com"]
5	6	Manipulate:Select	filterRange	year=2011~2012
6	7	Facet:Juxtapose	copyAndJuxtapose	horizontal
7	8	Manipulate:Select	filterRange	year=2013~2014
5	9	Manipulate:Select	filterAscendingTop	3
9	10	Arrange:Chart	Linechart	x=year(author_date), y=filtered(count(id)), color=filtered_previously
10	11	Map:Color	changeColor	range=colorbrewer2.xxx
11	12	Arrange:Express	addAverageLine	year

Figure 1. (a) 사용자가 직접 작성한 인터랙션 로그 (b) 변환된 Intermediate Representation

C. North는 프로비넌스 연구를 perceive, capture, encode, recover, reuse의 다섯 단계로 나누고 있다 [1]. 본 연구에서는 캡처한 사용자 액션을 인코딩 하고 (encode), 프로비넌스를 이용하여 사용자의 분석 과정을 복기하며 (recover), 이 분석과정을 다시 다른 데이터셋에 적용하여 재사용할 수 있는 단계 (reuse)를 중점적으로 다룬다.

2. VRR(Visual analytics tool for Recover&Reuse) 시스템

2.1 사용자 인터랙션 데이터의 수집 및 인코딩

본 연구에서는 여러 개발 관련 시스템(ex. 버전 관리 시스템, 이슈/버그 관리 시스템 등)에서 수집한 소프트웨어 엔지니어링 관련 도메인의 데이터를 활용하였다. 우선 랩 스터디를 진행하여 2명의 지원자들로부터 Think aloud기법을 활용하여 인터랙션 프로비넌스 데이터를 얻었다. 지원자들은 데이터의 분석 과정에서 발생하는 인터랙션들을 말로 표현하면서, 동시에 그 내용을 서술하도록 하였다. 액션의 예제는 Figure 1(a)와 같다.

위와 같이 매뉴얼하게 작성된 사용자 인터랙션 데이터는 VRR시스템이 인식할 수 있는 포맷으로 인코딩이 필요하다. 우리는 이를 컴파일러 기술들이 채택하는 방식[8]과 유사하게, *Intermediate Representation(IR)* 형태로 변환하였다. 이는 개개인의 분석 서술방식을 공통적인 언어로 변환하는 작업이 필요했기 때문으로, 차후 여러 종류의 시각적 분석 툴에 관계없이 인터랙션 데이터를 수집할 수 있는 기회를 제공할 수 있다.

1	컨택스트	오름(내림)차순 Top 3만 표현(필터링)
	기존 IRs	오름(내림)차순 정렬 ➔ 왼쪽 3개의 엘리먼트들 브러싱 ➔ 셀렉션의 필터링
2	컨택스트	평균 레퍼런스 라인 추가
	기존 IRs	Y축 데이터의 평균값을 구한다. ➔ 평균값을 기준으로 수평선을 그린다.

Table 1. 컨택스트 레벨 IR의 예

IR은 기본적으로 Vega [7]및 Tamara의 visualization interaction idiom [5]을 기반으로 작성되었다. Figure 1(b)에서 나온 것과 같이, IR의 데이터는 인덱스, 그래프 형태를 보존하기 위한 부모의 인덱스, 카테고리, 인터랙션, 그리고 파라미터들로 구성된다. 이런 단위 인터랙션을 표현하는 IR은 기존 툴들의 히스토리 로그와 유사하게 세부적인 사용자 인터랙션을 표현할 수 있으나, 이것만으로는 사용자의 의도나 통찰을 쉽게 파악하기가 쉽지 않다. VRR에서는 이런 인터랙션을 사용자의 의도를 파악할 수 있도록 추상화하여 표현하기 위해, 컨택스트 기반 레벨을 더하였다. 컨택스트 레벨 IR은 Table 1에서 볼 수 있듯이, 기존의 IR의 모음을 컨택스트 기반으로 해석한 후, 추상화한 표현 방식이다.

2.2 그래프 표현

IR을 그래프 형태로 표현하는 과정에서, 인터랙션의 로그들이 매우 길거나 복잡할 경우에 분석 과정이 복잡해질 수 있다. 이에 우리는 시퀀셜한 IR을 두 레벨의 계층으로 나누어 표현하였다. Anchor는 상위 계층으로, Detail로 표현이 되는 IR들의 그룹의 첫번째 IR을 뜻한다. Figure 2에서 보듯이, 하나의 Anchor IR들이 여러 개의 Detail IR들을 포함하는 형태로 표현이 된다. IR이 Anchor가 되는 것은 아래이 차트의 표현이나 데이터에 변경이 있을 경우이며, 아래와 같다.

a. 차트의 생성 혹은 차트 종류의 변경
b. Facet의 변경
c. 표현되는 데이터의 종류 혹은 범위의 변경

예를 들어, IR이 Bar 차트가 라인차트로 바뀐다거나 (a), 데이터의 특정 부분만 필터링 시키는 경우 (c)에, 해당 IR은 Anchor가 되고, 나머지 Detail IR들은 각각이 포함되는 Anchor IR쪽으로 그룹핑 된다.

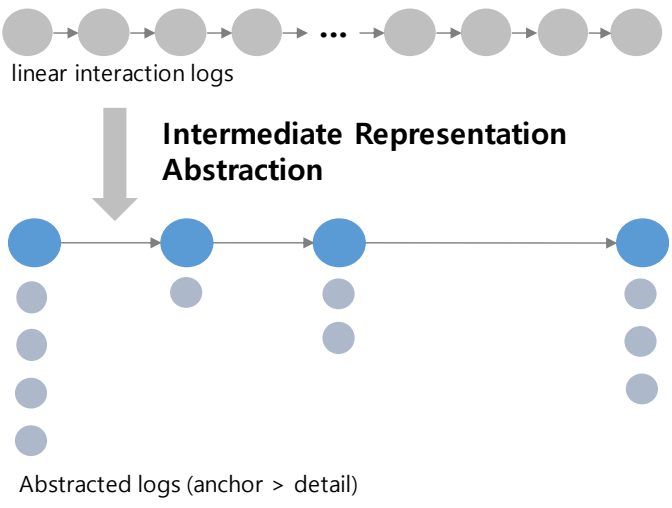

Figure 2. IR의 추상화 (hierarchy). 파란색 원이 Anchor 이며, 그 아래로 Detail들이 그루핑된다

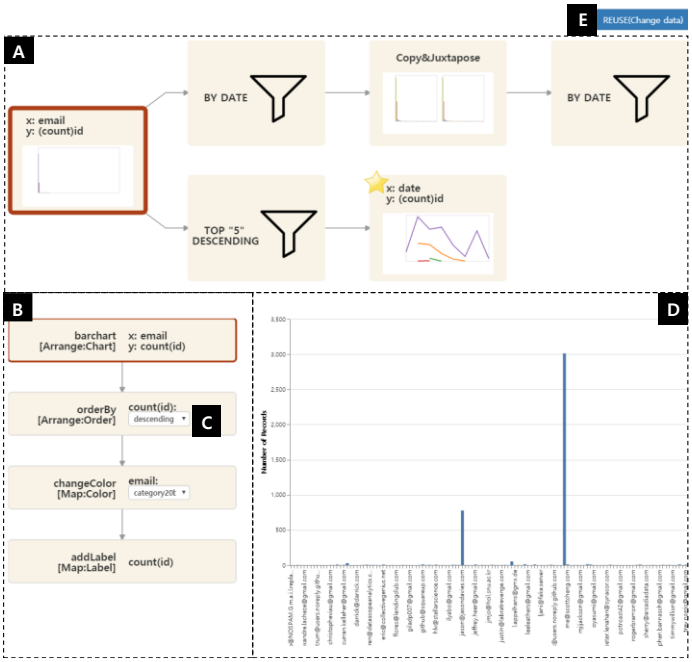

Figure 3. VRR 화면. (a) overview (b) detail view (c) controllable interaction parameter (d) chart view (e) reuse menu

2.3 VRR 시스템 구현

VRR은 자바스크립트 및 d3.js를 기반으로 하는 프론트엔드 어플리케이션 형태로 구현되었다. 우리는 Overview를 유지하면서 context를 보여줄 있는 Focus-plus-context를 기반으로 하여 Figure 3과 같이 레이아웃을 구성하였다. Overview에서는 Anchor들로 구성된 단순화된 Path를 이용하여 분석자의 추론 프로세스를 보여주고 있다 (a). Detail view는 overview에서 선택된 anchor에 속한 detail에 대한 리스트를 확인할 수 있다 (b). Chart view에서는 Detail view에서 선택한 부분까지 적용된 차트 결과를 확인할 수 있다 (d). 여기서 사용자는 필요에 따라, Detail view의 옵션을 조절함으로써 프로비넌스 데이터를 제한적으로 커스터마이제이션 할 수 있다. 예를 들어 (c)에서는 정렬 방식을 오름 혹은 내림 차순으로 변경할 수 있다.

또한, VRR은 (e)를 클릭함으로써, 기존 데이터와 유사한 형태의 새로운 데이터가 있을 경우, 해당 프로비넌스 정보를 새 데이터에 적용시킬 수 있다.

3. 결론 및 토론

이 논문에서는 프로비넌스 데이터의 효율적인 분석을 위한 기법 및 시각화 도구를 제안하였다. 먼저 사용자 인터랙션 로그를 구조적으로 인코딩하고, 컨택스트를 고려한 추상화를 할 수 있는 IR형태로 변환하였다. 시각화 과정에서는 Anchor와 Detail의 두 계층을 두어 기존 선형적인 분석을 보완할 수 있는 오버뷰를 가능하게 하는 기법을 제안하였다. 마지막으로 이를 기반으로 그래프 형태로 프로비넌스를 복기하고, 재사용할 수 있는 웹 기반 프로토타입을 구현하였다.

본 연구는 프로비넌스의 capture단계에 대한 연구가 메인이 아닌 관계로, 사용자 액션을 매뉴얼하게 수집하는 think aloud기법을 활용하였다. 프로비넌스 데이터의 대량 수집 및 처리를 위해서는 이에 대한 자동화 기법들을 고려해야 할 필요가 있다.

그리고 본문에서 제안한 컨택스트 레벨 IR의 케이스를 강화할 필요가 있다. 이는 자연어 기반 혹은 멀티모달 환경에서의 시각화 쿼리를 보면, 추상화된 쿼리들의 종류를 확인할 수 있다 [2]. 이를 활용하여 컨택스트 레벨 IR의 종류를 늘릴 수 있을 것이다. 그리고, 다수 분석가의 분석 패턴들을 확인할 수 있는 기능들로 확장 역시 필요하다.

참 고 문 헌

[1] C. North et al., “Analytic Provenance: Process + Interaction + Insight,” 29th Annu. CHI Conf. Hum. Factors Comput. Syst. CHI 2011, pp. 33–36, 2011.

[2] Setlur, Vidya, et al. "Eviza: A natural language interface for visual analysis." Proceedings of the 29th Annual Symposium on User Interface Software and Technology. 2016.

[3] Pohl, M., Wallner, G., & Kriglstein, S. (2016). Using lag-sequential analysis for understanding interaction sequences in visualizations. International Journal of Human Computer Studies, 96, 54–66.

[4] Bors, Christian, et al. "A provenance task abstraction framework." IEEE computer graphics and applications 39.6 (2019): 46-60.

[5] Visualization Analysis and Design by Tamara Muzner

[6] Heer, Jeffrey, et al. "Graphical histories for visualization: Supporting analysis, communication, and evaluation." IEEE transactions on visualization and computer graphics 14.6 (2008): 1189-1196.

[7] Satyanarayan, Arvind, et al. "Reactive vega: A streaming dataflow architecture for declarative interactive visualization." IEEE transactions on visualization and computer graphics 22.1 (2015): 659-668.

[8] LLVM (https://llvm.org/)