Open-World COVID-19 Data Visualization [Extended Abstract]

Hyunseung Hwang and Steven Euijong Whang(✉)

Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{aguno,swhang}@kaist.ac.kr

Abstract. As COVID-19 becomes a dangerous pandemic worldwide, there is an urgent need to understand all aspects of it through data visualization. As part of a larger COVID-19 response by KAIST, we have worked with students on generating interesting COVID-19 visualizations including demographic trends, patient behaviors, and effects of mitigation policies. A major challenge we experienced is that, in an *open world setting* where it is not even clear which datasets are available and useful, generating the right visualizations becomes an extremely tedious process. Traditional data visualization recommendation systems usually assume that the datasets are given, and that the visualizations have a clear objective. We contend that such assumptions do not hold in a COVID-19 setting where one needs to iteratively adjust two moving targets: deciding which datasets to use, and generating useful visualizations with the selected datasets. We thus propose interesting research challenges that can help automate this process.

1 Introduction

The COVID-19 pandemic is widely considered as one of the world's biggest challenges since World War II. Even as some countries are successful in overcoming the first wave of this pandemic, they are bracing for a second wave that is likely to come towards the end of the year. Hence, there is an urgent need to understand all aspects of COVID-19. We not only need to develop vaccines, but also understand how the disease spreads, how people are reacting, how to mitigate COVID-19, and more. Data visualization is commonly used to provide decision support for policy making.

Public health 2.0 has never been more important where we can utilize a growing list of public and restricted, but accessible data sources. Some well-known sources include the WHO website, Johns Hopkins Coronavirus Resource Center, the Centers for Disease Control and Prevention (CDC), and Kaggle. South Korea is one of the leading countries for combating COVID-19, and there is a Korean version of the CDC (KCDC) that shows various statistics about COVID-19 patients. In addition, the Korean government supports access to various datasets through data safe zones (e.g., Korea Telecom mobile data) and

Supported by a Google AI Focused Research Award.

V. Gadepally et al. (Eds.): Poly 2020/DMAH 2020, LNCS 12633, pp. 81–84, 2021.
https://doi.org/10.1007/978-3-030-71055-2_8

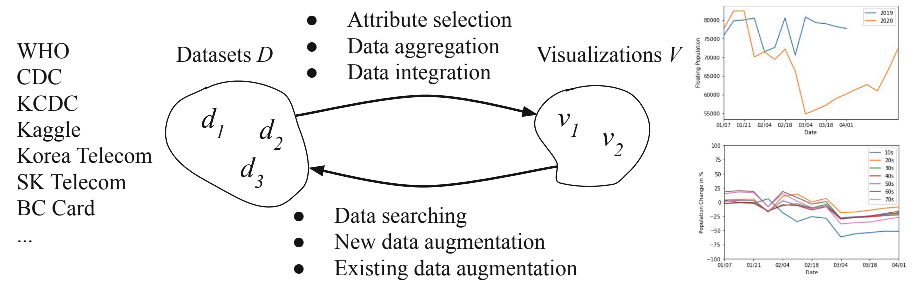


Fig. 1. In open-world data visualization, finding the right datasets and visualizations is repeated until the visualizations are considered insightful to the analyst.

funds data voucher projects where companies process and sell datasets that help COVID-19 analysis (e.g., BC Card credit card data). There are also various data visualization competitions for COVID-19.

Ironically, all these data sources make data visualization extremely challenging, as there is a large search space of combinations of datasets and possible visualizations on them. Traditional data visualization systems are not sufficient because they usually assume the input datasets are fixed, and the visualizations have specific performance goals. In reality, we are in an *open-world setting* where datasets are continuously evolving (e.g., KCDC may update its dataset or introduce a new one) and the helpful visualizations may change as new events occur (e.g., a new policy for social distancing). Hence, open-world data visualization involves multiple iterations of two moving targets: searching and integrating potentially useful datasets [1] and generating the right visualizations on them [3]. Both topics have been studied individually under various names, but have seldom been studied together within the same workflow.

We suggest research challenges that can help automate open-world data visualization as illustrated in Fig. 1. Given datasets, generating visualizations involves selecting attributes, aggregating data, and integrating data. Once visualizations have been generated, the data analyst may improve them by searching for new datasets and augmenting the existing ones. We also provide case studies of working with independent study students to generate useful COVID-19 visualizations. While we were able to obtain a number of interesting visualizations, the iterative trial and error with students literally took weeks. In hindsight, this process was tedious and not fast enough to generate visualizations to be useful during the first wave of the pandemic. For example, Korea initially experienced a shortage of face masks, but by the time we figured out how to estimate the supply and demand of masks per region, the problem was largely solved. However, our efforts are expected to be useful for the possible second wave.

2 Research Challenges

The goal of open-world data visualization is to choose a set of datasets $D = \{d_1, d_2, \dots d_n\}$ and a set of visualizations $V = \{v_1, v_2, \dots v_m\}$ that provide the best insights of COVID-19. Whether a visualization is insightful can be determined manually by the data analyst. If the analyst is not satisfied, she can generate new visualizations on the current datasets or search for more datasets that complement the current ones. Even if the visualizations look good now, there may be new events that prompt the analyst to generate better ones later.

We identify research challenges that occur when exploring visualizations and searching datasets. Some of the challenges are not new in data analytics, but we would like to tailor them to an open-world visualization setting.

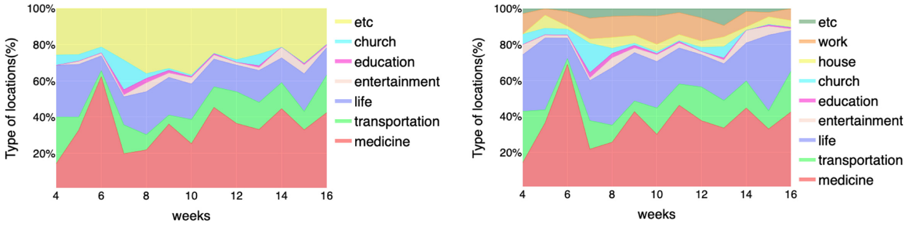
1. Exploring visualization candidates given a set of datasets ($D \rightarrow V$)
   (a) *Attribute selection*: Finding the right attributes to show in visualizations is a time-consuming process where the analyst has to identify which combinations of attributes produce interesting visualizations.
   (b) *Data aggregation*: Determining the granularity (e.g., daily/weekly/ monthly) of analyses and deciding which attributes (e.g., age, region) to group on. This process can be manual and repetitive.
   (c) *Data integration*: Aligning possibly-inconsistent attributes so the datasets can be joined. The integration may be repeated as the analyst finds new datasets to augment existing visualizations.
2. Searching new datasets to improve the visualizations ($V \rightarrow D$)
   (a) *Data searching*: The analyst may realize the given datasets are unsuitable for generating the desired visualization and may need to search new datasets from scratch. While there are existing dataset search tools, they do not necessarily cover recently-added COVID-19 data sources.
   (b) *New data augmentation*: The analyst may want to add new visualizations on top of existing ones and search for new datasets that contain the needed information. Unlike searching from scratch, the new datasets may need to be integrated with the existing ones.
   (c) *Existing data augmentation*: This case is similar to new data augmentation, but the emphasis is more on supplementing any missing values of existing datasets and visualizations.

3 Case Studies

We highlight the data visualization experiences of two analysts[1] among others. Both analysts ran into most of the research challenges above, and we specify when the new or existing data augmentations occur below.

Analyst A wanted to visualize types of locations visited by COVID-19 patients over time. Initially, Analyst A used a patient route dataset provided by KCDC. This dataset contains for each patient a list of locations where each location has a latitude, longitude, and location type. After the initial visualization, however, Analyst A noticed that many location types were not categorized and had "etc." values as shown in Fig. 2a. Analyst A then decided to utilize a public building information database to fill in the missing categories (i.e., existing data

[1] Credits go to Beomsik Park and Jaeyoung Park.



(a) Before augmentation (b) After augmentation

Fig. 2. (a) Types of locations visited by COVID-19 patients over time. (b) After augmenting the "etc." values (yellow region) using additional building information. (Color figure online)

augmentation) as shown in Fig. 2b. This operation required a non-trivial conversion from latitude/longitude values to building addresses using Naver Maps before joining with the building information database.

Analyst B wanted to see how significant COVID-19 events influence people's behaviors. Initially, Analyst B focused on subway usage data provided from the Seoul Open Data Plaza website. The subway population significantly decreases after two events: right after the first COVID-19 patients were diagnosed in Korea and after a major outbreak by members of Shincheonji. Analyst B then wanted to compare this visualization with the more general floating population data by obtaining a new dataset provided by SK Telecom (i.e., new data augmentation). As a result, while the floating population also decreases after the first COVID-19 patient diagnosis, it does not decrease much after the outbreak. We suspect that people learned how to cope with COVID-19 better and, instead of taking the subway, drove themselves during the outbreak.

4 Discussion

COVID-19 is here to stay, and understanding it through data visualizations will only become more important. In addition to the existing approaches for closed-world data visualization, we hope the research community will tackle the novel problem of accelerating open-world data visualization. Recently, deep learning approaches [2] have been used to automate data visualizations by determining which plot types are suitable for visualizing which attributes. An interesting direction is to expand this approach to jointly perform dataset searching as well.

References

1. Brickley, D., Burgess, M., Noy, N.F.: Google dataset search: building a search engine for datasets in an open web ecosystem. In: WWW, pp. 1365–1375. ACM (2019)
2. Hu, K.Z., Bakker, M.A., Li, S., Kraska, T., Hidalgo, C.A.: Vizml: a machine learning approach to visualization recommendation. In: CHI, p. 128 (2019)
3. Wongsuphasawat, K., et al.: Voyager 2: augmenting visual analysis with partial view specifications. In: CHI, pp. 2648–2659. ACM (2017)