CovidLens: Visually Understanding the Covid-19 Indicators through the Lens of Mobility Data

Mohamed A. Sharaf
Department of Computer Science
United Arab Emirates University
msharaf@uaeu.ac.ae

Xiaozhong Zhang
Department of Computer Science
University of Pittsburgh
xiz151@pitt.edu

Panos K. Chrysanthis
Department of Computer Science
University of Pittsburgh
panos@cs.pitt.edu

Wadima Alsaedi, Maitha Alkalbani, Heba Helal, Alyazia Aldhaheri
Department of Computer Science
United Arab Emirates University
{201701975,201709425,202090034,201701943}@uaeu.ac.ae

***Abstract*—Since the onset of the Covid-19 pandemic, an overwhelming amount of related data has been released. In an attempt to gain insights from that data, multiple public data visualization dashboards have been deployed. Differently from such dashboards, which mainly support basic data filtering and visualization of separate datasets, in this work, we propose CovidLens, which: 1) integrates various Covid-19 indicators and is centred around the *Google Community Mobility Report* dataset, 2) supports similarity search for finding similar and correlated patterns and trends across the integrated datasets, and 3) automatically recommends insightful visualizations that unlocks valuable insights into the pandemic effects. To that end, we will be presenting the employed dataset, together with the design, implementation, and multiple usage scenarios of our proposed CovidLens.**

***Index Terms*—data exploration, data visualization, visual dashboard, similarity search, visualization recommendation**

I. Introduction

Visual data exploration is ubiquitous in nearly every domain such as transportation, healthcare, education, manufacturing, finance, just to name a few. Hence, data-driven visual dashboards have become indispensable in supporting well-informed actionable decision making.

Essentially, dashboards are generated as a collection of summarized aggregate queries over the underlying data, providing a bird's eye view to facilitate discovering valuable insights. However, unlocking those insights has been anecdotally compared to *finding a needle in a haystack*. Particularly, analysts need to manually construct a prohibitively large number of aggregate queries and visually explore their results looking for insights, which is clearly an ad-hoc and labor-intensive process. Hence, in this work, we propose and demonstrate *CovidLens*, our visual analytics dashboard for assisting analysts into automatically unlocking insights from Covid-19 spatio-temporal data.

Since the outset of the Covid-19 pandemic, both the public and the experts have been equally monitoring an overwhelming amount of Covid-19 related data. For instance, *Johns Hopkins University CSSE* maintains a public repository, which collects and stores related data from allover the globe such as the number of Covid-19 daily cases, recoveries, vaccinations, etc. [1]. Similarly, *Google* has released and maintains its Google Community Mobility Reports dataset, which records how the community mobility habits have changed since the Covid-19 pandemic outbreak [2].

In an attempt to gain insights from such datasets mentioned above, multiple public data visualization dashboards have been deployed (e.g., [3], [4]). However, such dashboards mainly support basic data filtering operations and visualizations. For instance, [3] provides an outstanding interface that supports rudimentary visualizations of the Google Mobility Reports based on the user's search parameters. Moreover, existing dashboard typically focus on only one of the available datasets, where analysts will have to use different dashboards to explore the different available Covid-19 related data.

Accordingly, we have designed our proposed CovidLens to: 1) integrate multiple Covid-19 indicators that are centred around Google's mobility data, 2) support similarity search for finding correlated patterns and trends, and 3) provide automatic recommendation of insightful visualizations based on different well-studied metrics such as deviation and correlation (e.g., [5], [6]). Our design allows CovidLens to support analysts in automatically unlocking valuable hidden insights (e.g., countries with very similar/different approaches in implementing lockdown). Moreover, it also allows understanding the relationships between different Covid-19 indicators (e.g., the relationship between number of vaccinations and the mobility to workplace).

The rest of the paper is structured as follow: Section II discusses the technical details for CovidLens, covering data processing, visualization and recommendation. Section III presents multiple example usage scenarios of CovidLens.

II. The CovidLens System

A. Spatio-Temporal Data

Spatial-temporal data keeps the space and time information of each record in the data, such that each record is in the form of $\langle s_i, t_j, v_{ij} \rangle$, where s_i represents a space point or range,

2375-0324/22/$31.00 ©2022 IEEE
DOI 10.1109/MDM55031.2022.00066

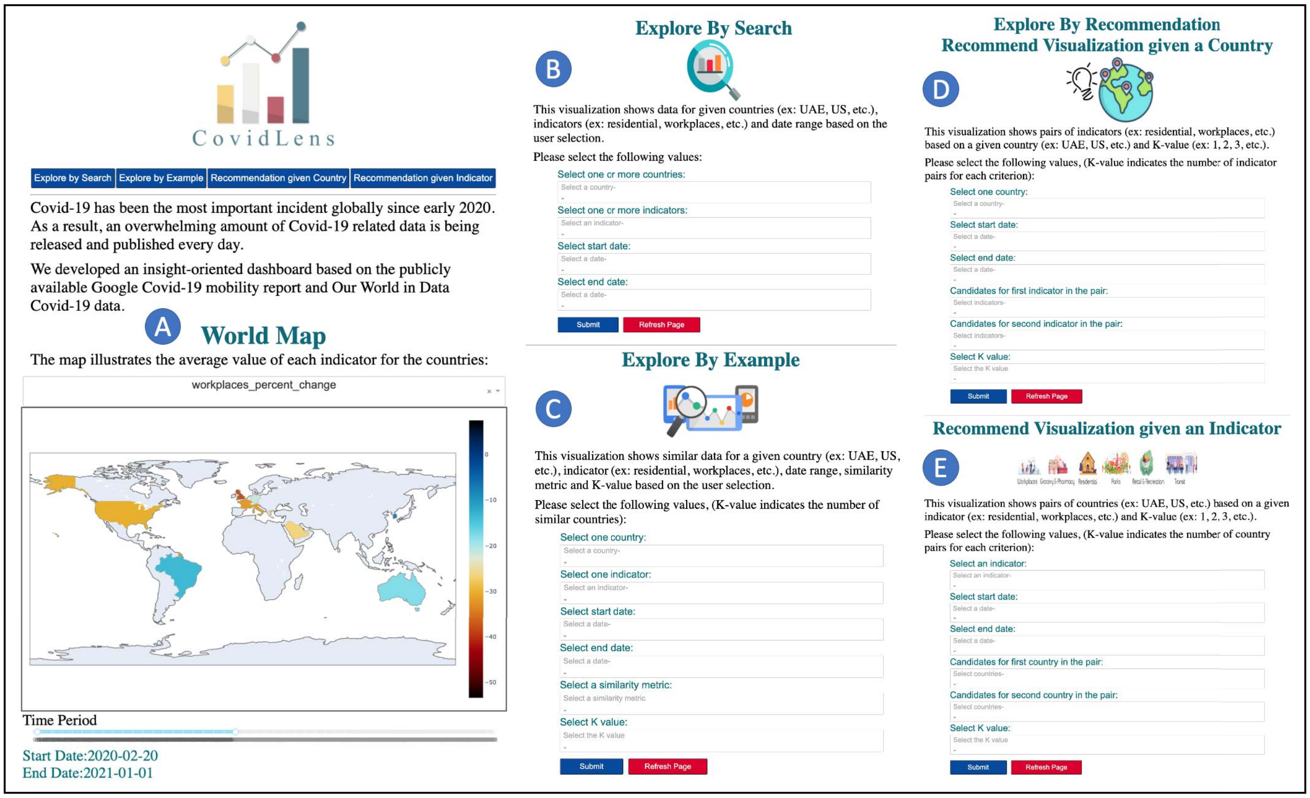

Fig. 1. CovidLens GUI overview (https://db.cs.pitt.edu/covidlens/)

t_j represents a time point or range, and v_{ij} represents the corresponding indicator value.

CovidLens integrates multiple spatio-temporal datasets together with the Google mobility report data [2]. The main measurement in Google's data is the *mobility change ratio*, and it is recorded for six types of different place categories (i.e., residential area, workplace, etc.). Particularly, the change ratio for a category is measured by the change in number of visitors to that category compared to a *baseline* number. That baseline number is the median value from a 5-week period between Jan 3, 2020 and Feb 6, 2020. To be more precise, there are actually seven baseline numbers; one for each day in the week. Hence, the change ratio for a specific week-day (e.g., Saturday) is calculated based on the corresponding baseline number for that day.

The *Mobility* table in Fig. 2 shows some example records from the mobility dataset. For instance, the first record indicates that in the UAE on 2020-04-01, the mobility ratio to workplace was 46% lower than its baseline, whereas the stay duration in residential areas was 25% higher than its baseline.

In addition to the mobility data, CovidLens integrates Covid-19 indicator data such as new cases, new deaths, vaccination count, etc. [3]. Examples of such indicators are shown in Table *Covid* in Fig. 2. For instance, the first record shows that for the UAE on 2020-04-01, the number of new cases per million and new deaths per million are 15.0 and 0.2.

Mobility

country	date	workplaces	residential
UAE	2020-04-01	-46	25
UAE	2020-04-02	-47	28
UAE	2020-04-03	-44	22

Covid

country	date	new cases	new deaths
UAE	2020-04-01	15.0	0.2
UAE	2020-04-03	24.0	0.1

Joined

country	date	workplaces	residential	new cases	new deaths
UAE	2020-04-01	-46	25	15.0	0.2
UAE	2020-04-02	-47	28	null	null
UAE	2020-04-03	-44	22	24.0	0.1

Fig. 2. Example join of the two datasets

B. Data Preprocessing

In CovidLens, all the integrated indicators are combined into a single table. Particularly, that combined table is generated through a left outer join between the mobility data, and the datasets for each of the other integrated indicators. Fig. 2 shows a sample of the integrated data, in which the complete mobility data is set as the left side of the join.

Clearly, the primary key of the integrated table is the composite key of country name and date. Currently, CovidLens integrates all the indicators from the Google mobility dataset, the six indicators are mobility change percentage over baseline for six place categories, namely `retail and recreation`, `grocery and pharmacy`,

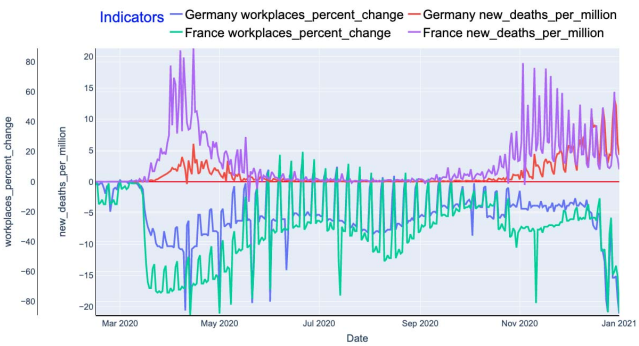

Fig. 3. Example Explore-by-Search result

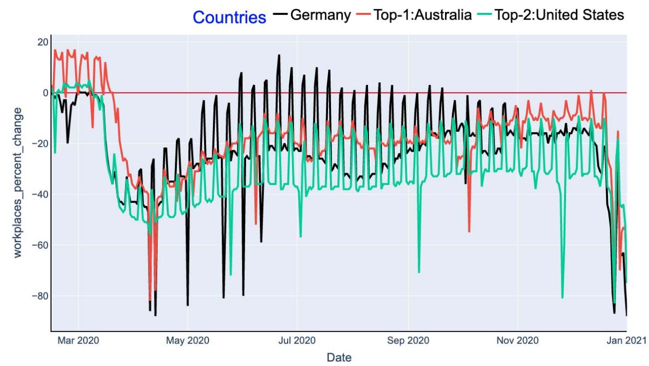

Fig. 4. Example Explore-by-Example result

`parks`, `transit`, `workplaces`, and `residential`. Additionally, we integrate publicly available data for the following indicators: `new cases`, `hospitalized patients`, `new deaths`, `vaccinated people`, `fully vaccinated people`, and `poverty ratio`.

C. Data Visualization

An overview of the CovidLens interface is shown in Fig. 1. CovidLens provides five interactive panels: two for data visualization, one for similarity search, and two for visualization recommendation. Next, we describe the operations supported by each panel.

The first panel displays a world map to help users get a quick and overall picture of the different indicators in different countries (Fig. 1 panel A). Particularly, this panel allows users to select an indicator from the top dropdown menu and a time range from the bottom slider, then it calculates and displays the average values for the selected indicator using a heat map. For example, the world map in Fig. 1 displays average workplace mobility change for different countries in year 2020.

To support analysts who want to further explore interesting patterns revealed by the world map, CovidLens provides an *Explore by Search* panel (Fig. 1 panel B), in which users can pinpoint specific indicators for further investigation. Particularly, that panel allows users to select multiple countries and indicators and visualizes the corresponding data in a multi-trace line chart over the user-specified time range. For example, Fig. 3 visualizes the workplace mobility change and new Covid deaths for Germany and France in year 2020.

D. Similarity Search

In CovidLens, similarity search is supported by the *Explore by Example* panel (Fig. 1 panel C). Particularly, given a country, an indicator and a similarity metric, CovidLens will search for the *top-k* countries with similar trends. For example, Fig. 4 shows that in 2020, Australia and the USA are the top-2 countries similar to Germany in terms of workplace mobility reductions.

CovidLens supports three similarity metrics: *Euclidean distance*, *Dynamic time warping distance* (DTW) [7], and *Longest common subsequence* (LCSS) [8]. The Euclidean distance, for example, is calculated using Equ. 1,

$$d(V_1, V_2) = \sqrt{\sum_{j=1}^{n} (v_{1j} - v_{2j})^2} \tag{1}$$

where V_1 and V_2 are the two indicator series; v_{1j} and v_{2j} are the corresponding values in the two indicator series for time stamp j; n is the total number of time stamps.

While Euclidean distance calculation is based on continuous one-to-one matching between the two series, DTW and LCSS bring some matching flexibility through allowing one-to-many/many-to-one matching and time stamp skipping respectively. The focuses of the similarity metrics are also different. While Euclidean distance and DTW focus on the overall distance between the two series, LCSS focuses on the common subsequence of the two series (i.e., the total number of time stamps at which the two series are close enough to each other).

E. Visualization Recommendation

CovidLens provides two panels for automatically recommending insightful visulizations from the integrated datasets. The first such panel (Fig. 1 panel D) recommends visualizations based on the indicator measures, whereas the visualizations recommended by the second panel (Fig. 1 panel E) are based on location (i.e., country).

In the first panel, after the user selects a country and a time range, CovidLens will automatically recommend the top-k visualizations, each of which displays *a pair of indicator series* showing an interesting insight. For example, for a user selecting Germany during 2020, the top recommended visualization is shown in Fig. 5, which reveals a high anti-correlation between the new death cases vs. change in workplace mobility. To achieve such recommendation, we expand on our previous work (e.g., [9]–[12]), where the main idea is to automatically generate all possible visualizations (i.e., all pairs of indicator series). Then the interestingness of each visualization is measured using some utility score, then the top-k visualizations based on that score are recommended. In

Fig. 5. Example indicator pair recommendation

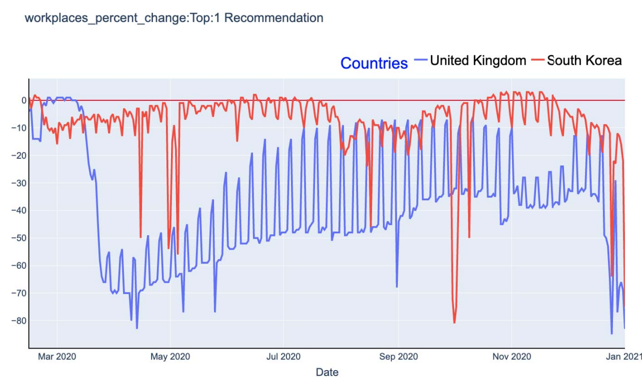

Fig. 6. Example country pair recommendation

CovidLens, we adopt two different utility measures, namely deviation and correlation (e.g., [5], [6]).

The deviation score is calculated using Euclidean distance as shown in Equ. 1 and the correlation score is calculated using absolute value of Pearson correlation coefficient,

$$|r(V_1, V_2)| = \frac{|\sum_{j=1}^{n}(v_{1j} - \bar{v}_1)(v_{2j} - \bar{v}_2)|}{\sqrt{\sum_{j=1}^{n}(v_{1j} - \bar{v}_1)^2 \sum_{j=1}^{n}(v_{2j} - \bar{v}_2)^2}} \quad (2)$$

where $\bar{v}_1 = \sum_{j=1}^{n} v_{1j}$ and $\bar{v}_2 = \sum_{j=1}^{n} v_{2j}$; $|\cdot|$ is the absolute value operator; V_1, V_2, v_{1j}, v_{2j}, j and n have the same meaning as those in Equ. 1.

Differently from the panel described above, the second recommendation panel in CovidLens (Fig. 1 panel E) recommends visualizations, where each one shows *a pair of countries* exhibiting an interesting pattern. Particularly, for a given indicator and time range selected by the user, CovidLens will recommend the top-k pairs of countries that exhibit the highest utility scores. For example, Fig. 6 shows a recommendation according to the deviation-based utility metric for the workplace mobility reduction indicator. As the figure shows, in South Korea, the reduction in workplace mobility was much smaller than that in the UK.

III. Demonstration Scenarios

A. Demo Artifact

CovidLens was developed using the Plotly Dash framework, which is written on the top of Flask, Plotly.js and React.js.

B. Demo Plan

Equipment: The conference attendees will have the opportunity to interactively engage with the CovidLens web dashboard on a laptop, a tablet or a smartphone.

Datasets: We will pre-load the Google mobility dataset and the other Covid-19 indicators mentioned in the paper.

Scenarios: We already mentioned the *Location-aware exploration* scenario in Section II, in which participants can use the world map panel to discover interesting patterns between countries while having their spatial relations in mind.

A second scenario that participants can explore is *Inter-dataset analysis using multiple panels*. For example, after the user discovers the workplace mobility change difference between Germany and France in 2020 (as indicated by the color difference in Fig. 1 panel A), they could use the Explore-by-Search panel to view the precise day-by-day changes of the two countries. Then, to discover potential causes for the difference, they could continue with the indicator pair recommendation panel to discover the high correlation between workplace mobility change and new deaths in both countries (the Germany's case is shown in Fig. 5). After that, they could add the new deaths indicator to the Explore-by-Search panel (as shown in Fig. 3) and form a hypothesis that the mobility change difference in the two countries might be caused by the difference in new death cases in year 2020.

Acknowledgment. This work is partially supported by UAE University grant G00003352 and US NSF grant SES-2017614.

References

[1] "JHU CSSE COVID-19 Data," https://github.com/CSSEGISandData/... COVID-19, accessed: 2022-02-27.
[2] https://www.google.com/covid19/mobility, accessed: 2022-02-27.
[3] https://ourworldindata.org/coronavirus, accessed: 2022-02-27.
[4] https://gisanddata.maps.arcgis.com/apps/dashboards/... bda7594740fd40299423467b48e9ecf6, accessed: 2022-02-27.
[5] M. Vartak, S. Rahman, S. Madden, A. G. Parameswaran, and N. Polyzotis, "SEEDB: efficient data-driven visualization recommendations to support visual analytics," *VLDB*, vol. 8, no. 13, pp. 2182–2193, 2015.
[6] R. Ding, S. Han, Y. Xu, H. Zhang, and D. Zhang, "Quickinsights: Quick and automatic discovery of insights from multi-dimensional data," in *ACM SIGMOD*, 2019, pp. 317–332.
[7] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
[8] M. Vlachos, D. Gunopulos, and G. Kollios, "Discovering similar multidimensional trajectories," in *IEEE ICDE*, 2002, pp. 673–684.
[9] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, "Efficient recommendation of aggregate data visualizations," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 2, pp. 263–277, 2018.
[10] M. A. Sharaf and H. Ehsan, "Efficient query refinement for view recommendation in visual data exploration," *IEEE Access*, vol. 9, pp. 76 461–76 478, 2021.
[11] R. Mafrur, M. A. Sharaf, and G. Zuccon, "Quality matters: Understanding the impact of incomplete data on visualization recommendation," in *International Conference on Database and Expert Systems Applications*, 2020, pp. 122–138.
[12] X. Zhang, X. Ge, P. K. Chrysanthis, and M. A. Sharaf, "Viewseeker: An interactive view recommendation framework," *Big Data Res.*, vol. 25, p. 100238, 2021.