LITERATURE SURVEY
GLOBAL SALES DATA ANALYTICS
TEAM ID: PNT2022TMID28374

1) BIG DATA ANALYTICS FRAMEWORK FOR PREDICTIVE ANALYTICS USING PUBLIC DATA WITH PRIVACY PRESERVING

AUTHORS: Duy H. Ho, Yugyung Lee

ABSTRACT:

There are increasingly leveraging public data with cities increasingly interested in driving both responsiveness to citizen demands and cost savings through data analytics. As public managers seek to augment existing data sources, such as 311 complaints, with existing secondary data, such as US Census products, severe challenges exist. This paper considers the problems inherent in data being collected at divergent geographic levels over different time horizons. An inductive analytical methodology is developed to create units of analysis that are both useful and analytically appropriate for public managers and policy leaders in urban areas. A big data analytics framework for public data, called BDAP, was presented predictive analytics for community need considering data the spatial and temporal location while addressing the data issues such as missing values, privacy-preserving, and predictive modeling. The findings illustrate the power of inductive data curation and privacy-preserving leading to benefits to the big data community. An application for the Open Data Platform was developed using KCMO's 311 data, crime data and census data.

RELATED WORK

A) Open data platforms

Danneels et al. defined open data platforms in terms of three different platform types: cognitivist as "standardized management of information," connectionist as "management of standardized information through communities," and autopoietic as "management of data through individual people." Neves et al. analyzed open data impact factors in smart city dimensions and found the big gap between the impacts and sustainable development of smart cities. The open data were defined as follows: the data are available online, in a machine-readable format that can be freely used, re-used, and redistributed by anyone . A recent study presented user behaviors as citizen-generated data through a 311 system tool . To overcome the limited knowledge of a 311 system and reduce the gap between consumer-oriented and user-oriented service of the 311 system, they analyzed the causal relationship among technology acceptance, residents' satisfaction, and frequency of public service use. The 311 service is a non-emergency service for residents to report problems and ask city-related questions in many cities, such as Baltimore, Buffalo, Chicago, New York City, San Francisco, Seattle, and Washington, D.C., etc. Some studies showed that an improvement in a city government's performance was achieved using 311.

The 311 calls were also used to enable citizen engagement in local government, fostering public services targeted to serve citizen needs. In these studies, the 311 customer data were combined with demographic data at the census tract level to develop insight as to patterns of 311 use by community members. Chatfield and Reddick discussed the effective use of big data analytics with a case study of Houston 311 data for data-driven government. Minkoff studied the distribution of physical conditions of government services using the New York City 311 data. In their study, through the tractlevel analysis of Census and 311 data, the problems were revealed in specific geographic spaces, such as older housing, vehicle traffic, and high population growth. O'Brien's study with Boston's 311 data was found more residents in the neighborhood are more active in reporting their close neighborhood issues; the residents were three times more likely to report the problems within two blocks of their home. Dawes et al. pointed out the importance of shared open platform for community studies. Stimulated by the works mentioned above, our goal in this paper is to enable communities to establish open data platforms for the reliable and sustained commitment of the platform actors and their connections.

B) Big Data Analytics for Public Data

There are ever increasing volume of data generated by the activity of government or individuals. This is an incredible opportunity for big data community to make more relevant and more responsive to situations, e.g., predict crime events, responds to emergent or non-emergent calls such as 911 or 311 calls. It is critical to integrate heterogeneous big datasets, which could be used to build spatio-temporal models in real-world settings. Deep predictive models could capture the complex spatial and temporal dependencies embedded in dynamic spatial-social environments such as crimes. Big data approaches was use to build predictive models for spatially and temporally dependent events in social contexts. Traditional machine learning algorithms were used to build classifiers for crime level prediction (low, medium, and high) with the accuracy of 83.95% Bayesian spatiotemporal methods to analyze the changing local patterns of crime over time, the criminal activity forecast using random forest regression and Poisson regression with features from diverse social data such as 911 data, 311 data, weather, and building data. Similarly, the crime rate of a neighborhood was predicted using crime data, demographic information, points of interest (POIs) data, and taxi data , the bluecollar/white-collar crime classification using gradient boosted trees, support vector machines, and random forest models. Crime prediction accuracy has been improved through a kernel density estimation and latent semantic and topics from tweets. The Steering Committee of the World Congress in Computer Science, Computer. Deep learning technologies have been used to predict crime hotspots with the images from Google Maps' street view, weather, demographics, and other crime incident information. Deep neural networks (DNNs) was used to estimate the likelihood of crime and the likely type of crime at the individual level based on individual criminal charge history data. Urban crime distribution was predicted using deep convolutional neural networks (DCNNs) to provide spatial policing patrol strategies in a coarse scale and for crime risk at the finer spatiotemporal scale.

Governments in general, and urban governments more specifically, are awash in data yet have difficulty in fully developing insights from across divergent data sources. The comprehensive use of big data platforms has led to an explosion in data, yet the Kansas City example illustrates the difficulty of integrating data analysis into a useable workflow for public management decision making. Here we address the omnipresent issue of needing to mix data from divergent sources in order to aid in real-world decisions about how to allocate limited resources. An inductive method that relies on big data analytics techniques allows for the clustering of data across time and geographic space to dictate areas of a city's landscape where needs are emergent, all without the constraints of jurisdictional boundaries that can serve to segregate data, which in turn impedes an analyst's ability to create a holistic explanation for current and future trends of citizen need. The data revolution is exciting and daunting for city administrators. The ability to develop a method to aggregate and ultimately create inference from disparate data sources can serve to aid in administrators' quest to provide finite city services to communities in need. The continued evolution of the smart city requires such methods so that administrators can continue to innovate in service delivery.

REFERENCES

[1] S. E. Bibri and J. Krogstie, "On the social shaping dimensions of smart sustainable cities: A study in science, technology, and society," Sustainable Cities and Society, vol. 29, pp. 219–246, 2017.

[2] T. Yigitcanlar, M. Kamruzzaman, M. Foth, J. Sabatini-Marques, E. da Costa, and G. Ioppolo.

[3] C. Duvier, P. B. Anand, and C. Oltean-Dumbrava, "Data quality and governance in a uk social housing initiative: Implications for smart sustainable cities," Sustainable cities and society, vol. 39, pp. 358–365, 2018.

[4] I. N. Gregory and P. S. Ell, "Breaking the boundaries: geographical approaches to integrating 200 years of the census," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 168, no. 2, pp. 419–437, 2005.

[5] P. Pittaluga, "Pioneering urban practices in transition spaces," City, Territory and Architecture, vol. 7, no. 1, pp. 1–10, 2020.

2) A SYSTEMATIC REVIEW TOWARDS BIG DATA ANALYTICS IN SOCIAL MEDIA

AUTHORS: Md. Saifur Rahman and Hassan Reza

ABSTRACT:

The recent advancement in internet 2.0 creates a scope to connect people worldwide using society 2.0 and web 2.0 technologies. This new era allows the consumer to directly connect with other individuals, business corporations, and the government. People are open to sharing opinions, views, and ideas on any topic in different formats out loud. This creates the opportunity to make the "Big Social Data" handy by implementing machine learning approaches and social data analytics. This study offers an overview of recent works in social media, data science, and machine learning to gain a wide perspective on social media big data analytics. We explain why social media data are significant elements of the improved data-driven decision-making process. We propose and build the “Sunflower Model of Big Data” to define big data and bring it up to date with technology by combining 5 V’s and 10 Bigs. We discover the top ten social data analytics to work in the domain of social media platforms. A comprehensive list of relevant statistical/machine learning methods to implement each of these big data analytics is discussed in this work. “Text Analytics” is the most used analytics in social data analysis to date. We create a taxonomy on social media analytics to meet the need and provide a clear understanding. Tools, techniques, and supporting data type are also discussed in this research work. As a result, researchers will have an easier time deciding which social data analytics would best suit their needs.

SOCIAL MEDIA STATISTICS

The abundance of consumer data makes social media a powerful resource for data analysis and research. In an unstructured format, social media content delivers a vast amount of big data. Consumer-generated social big data ensure the data’s integrity and value. This makes social data more attractive to researchers, businesses, the government, and others. The majority of social media sites offer Application Program Interfaces (APIs) that allow easy and public access to large amounts of social data for research purposes. Facebook, Google, Instagram, LinkedIn, and a slew of other social media platforms provide individuals and organizations customdesigned API. According to a recent Brandwatch report, there are 3.499 billion active social media users among the 7.7 billion global population, accounting for 45.4 percent of the overall population and nearly 80 percent of the total of netizens. This report was published on June 13, 2019. This is due to the fact that the number of people using social media increased by 202 million from April 2018 to April 2019. Amongst 81 percent of youngsters believe that social media have a beneficial influence on personal life. Each of these users spent an average of 142 min each day on social media. This statistic shows that the retailers and businesses corporation invested 74 billion dollars on social media advertising in 2018. As a result, 91 percent of retail companies use social media sites, while 81 percent of small and midsize enterprises use social media platforms to stay competitive and maximize marketing efforts. Data generated from users of social media become so revolutionary in volume, which results in the most important source of large data. The number of engagement on various social media sites. This table of data is generated based on the statistics of Brand watch. Based on the number of active users, Facebook was the most popular social media platform, with 2.3 billion in a month. There were 1.9 billion active users on the video-sharing website YouTube. Similarly, WhatsApp had 1.6 billion users. The professional networking website LinkedIn had 610 million subscribers, while another major network, Twitter, had 330 million active users and so on. The percentage of active social media consumers on various platforms is presented in the following Fig. 3. Figure 3 illustrates that Facebook has the most consumers (23 percent) who are also actively participating in social communication, followed by 19 percent on YouTube, 16 percent of all are active on WhatsApp, consumers are equally active on Instagram and WeChat and so on.

BIG DATA ANALYTICS IN SOCIAL MEDIA

The systematic computing and interpretation of data using statistical methods is known as analytics. Analytics uses mathematics, statistics, and artificial intelligence to help with data analysis in difficult-to understand formats so that better decisions may be made. At the same time, big data analytics assists data analysis by revealing trends, patterns, and other insights from messy social data. In this study, the terms "big data analytics" and "social media analytics" are used interchangeably. Text mining, social graph theory, opinion mining, social influence analysis, sentiment analysis, statistical analysis, cyber risk analysis, and others are some of the diverse approaches of big data analytics in social media. Furthermore, by merging, modifying, and extending ways to handle massive social data, these analytics contribute to the development and assessment of systems and informatics tools. Different firms might use the results of big data analytics to improve their production or marketing strategies to stay competitive in the digital business world. For example, social media analytics may help businesses get user input on their products, which can be used to make changes and get more value out of their brand. Leading companies such as Apple, Microsoft, Google, Honda, Facebook, NVidia, Amazon, Samsung, and others employ social media analytics regularly to improve their corporate strategies and customer relations practices. Research, civil defense, healthcare, banking, telecommunication, public transport system, insurance, and a variety of other industries can gain benefits from social media analytics to prepare for the future and make better data-driven recommendations while remaining flexible and agile. Sensitive events like elections frequently use sentiment and opinion mining in local and national elections processes

CHALLENGES AND LIMITATIONS

Many disciplines and sectors have advanced as a result of the widespread use of social media data and big data analytics. There are numerous hurdles and limitations to working in this field.

- With the increasing abundance of social media data, files are now being distributed over multiple physical sites. Public access is becoming difficult and technical skill is needed to access these data.
- The maintenance of large social datasets is challenging and expensive.
- Consumes continuously sharing status updates, photos, videos, etc., are not always useful for analysis. Data cleaning and filtering are required to extract necessary data from this complex dataset that is costly and time-consuming.

CONCLUSION

Big data, along with advances in computing tools, have evolved as a significant data analytics for understanding human behavior by analyzing data from social media. All types of organizations, from industry to government, can be benefitted using social data, data science, and social data analytics. This study fills in the research gap by identifying the ten most widely accepted and used big data analytics for analyzing social data and making decisions. Considering the overlap among the approaches of social media analytics, we design a taxonomy of big data analytics in the social media domain. We create three main categories and then assign these ten analytics to each one depending on the purpose, nature of usage, and working area. Data analysis in social media is aided by machine learning techniques. Each of these social data analytics has a long list of machine learning or statistical methodologies associated with it. We present social data analytics along with the methodologies. Until now, the most widely utilized analytics for social data analysis has been "Text Analytics". In addition, researchers are advancing their ability to extract useful information from an image, audio, and visual material in social data. The social media platforms provide data continuously to evaluate because it allows people to offer their perspectives on the most current event, product, tools, talents, and other topics. We must take advantage of the benefits of this massive amount of data. This research looks at big data analytics in social media in a broad, generic way. A specific field of interest, for example, business analytics in social media, geospatial/location based analytics, social media data analysis for political science research, etc. can be explored to serve the same purpose. We will continue our investigation by focusing on a small number of social media platforms, such as Facebook, Twitter, and Snapchat. Any of the ten big data analytics described above can be explored further. We do not get enough time and resources to investigate deep on the list of machine learning algorithms on each big data analytics. We aim to keep working on this project to find a shortlist of acceptable algorithms for each of these ten big data analytics categories. We also want to figure out and decide a few common attributes/characteristics of big data analytics by which we can tune up one analytics and perform comparative performance analysis.

REFERENCES

[1] V. Dhawan and N. Zanini, Big data and social media analytics, Res. Matters A Cambridge Assess. Publ., no. 18, pp. 36–41, 2014.

[2] K. Smith, 126 amazing social media statistics and facts, https://www.brandwatch.com/blog/amazing-socialmedia-statistics-and-facts/, 2019.

[3] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.

[4] N. A. Ghani, S. Hamid, I. A. Targio Hashem, and E. Ahmed, Social media big data analytics: A survey, Comput. Human Behav., vol. 101, pp. 417–428, 2019.

[5] P. V. Paul, K. Monica, and M. Trishanka, A survey on big data analytics using social media data, in Proc. 2017 Innov. Power Adv. Comput. Technol. (i-PACT), Vellore, India, 2017, pp. 1–4

3) ADVANCED DATA ANALYTICS PLATFORM FOR MANUFACTURING COMPANIES

AUTHORS: Tim Voigt, Nico Migenda, Marvin Schone, David Pelkmann, Matthias Fricke

ABSTRACT:

Data analytics is a key factor to make fully informed and data-driven decisions in modern manufacturing companies. With increasing international competition, classical approaches to data analytics have to be enhanced toward advanced data analytics to create smarter products, production processes and customer

services. This leverages the full potential of the data to quickly gain insights and make decisions. Application areas include predictive analytics, augmented analytics and real-time analytics. In this paper an advanced data analytics platform (ADAP) for manufacturing companies is presented and verified on a real-world production line. ADAP combines process data obtained in a physical environment with domain knowledge from experts to create data-driven models in a digital environment. In addition, ADAP is well scalable and provides a comprehensible data structure throughout data workflow. ADAP is verified on two different advanced data analytics applications on an electrocoating plant. In order to relate the applications to real production requirements, the needs of local manufactures from the region Ostwestfalen Lippe are taken into account.

ADVANCED DATA ANALYTICS IN MANUFACTURING

Upgrading manufacturing lines toward Industry 4.0 is of great interest for all companies around the globe. To provide a complete overview of data analytics platforms, it is necessary to show conceptional solutions, general ideas, examinations of existing approaches, and real-world applications. Because a complete overview is out of scope for this work, this review focuses on solutions that are successfully applied in the real world, with a special emphasis on OWL. A more extended international review is given.

A) International State of the Art

A modeling approach to forecast the energy consumption in factories for event venues is investigated. A platform is presented that starts with a data set and describes all necessary steps to build a prediction model. While promising results are shown, the data acquisition and storage steps are neglected. A more universal approach is presented, by the example of a plastic injection moulding process, providing a detailed workflow from data acquisition toward analysis for unplanned machine downtimes. A big data service architecture for monitoring cyber-physical production systems (CPPS) is presented. First, a conceptual platform and a 5-layer architecture are presented which are validated on data from a real production plant. In addition, different database types are compared. Enhancements in the production environment of a shipyard 4.0 are investigated. Industrial Internet of Things related challenges and opportunities for this specific case are given and a dashboard for monitoring workshop performance per ship is presented. While the previous work has focused on ships, the focus of is on building infrastructure for agricultural education. A framework for the entire data workflow from machine sensors toward learning apps and dashboards is presented. The framework using IoT-protocols is fully implemented. In addition, the work of focuses on an energy-efficient architecture. The architecture is divided into three layers: (1) A sensing layer containing the data sources; (2) a gateway layer for communication purposes; (3) a control layer containing control nodes. The performance is analyzed in an experimental study and achieves proper results with less resource utilization than classical computation infrastructure.

B) **Advanced Data Analytics in OWL**

The economy in OWL is characterized by SMEs in the fields of mechanical engineering, food industry, IT industry, automation technology, and furniture industry. More than 200 companies, research institutions, and organizations work and do research together in the technology network intelligent technical systems OWL ("it's OWL") on Industry 4.0 and artificial intelligence for industry. The current strategy of "it's OWL" focuses on autonomy, dynamic networking, sociotechnical interaction, and interlinking products and services. Autonomy implies that systems autonomously solve complex tasks within a given application domain. For example, a system autonomously detects an anomaly, evaluates a produced component for quality and failures, performs a (partially) safety assessment of manufacturing machinery, or determines redundancies of information sources. To achieve a desired overall system behavior, several subsystems must be able to interact. This requires custom interfaces for information exchange. One challenge is to ensure that this dynamic networking works consistently across value chains and manufacturers. The technological transformation toward Industry 4.0 will result in an intensified interaction with intelligent technical systems. For this sociotechnical interaction, it is necessary that a system can communicate in an intuitive way, for example, via speech or text. An additional aspect that must be ensured in this interaction, is security. For example, if process control takes place via a mobile device, secure authentication must be provided. In addition to the technical challenges of Industry 4.0, SMEs must also consider economic and social factors. A sociotechnical performance assessment for SMEs in the context of Industry 4.0 is presented. It allows to assess the current sociotechnical situation and helps to identify company specific actions for improvement. However, this transformation also affects the market, giving rise to new digital data-driven business models. The resulting business models, which are based on the interlinking of products and services, require new concepts for implementation. In a concept for identification, development and implementation of these business models is described. For example, a blockchain-based pay-per-use business model for industrial equipment is proposed

CONCLUSIONS & FUTURE WORK

In order to remain competitive, companies need a fast adoption of Industry 4.0 technologies. A key method to enable the development of these technologies is advanced data analytics, which requires in-depth domain knowledge on the one hand and experience in big data technologies, data mining and machine learning on the other hand. To face these challenges, ADAP was developed to integrate, store, process and analyze process data for Industry 4.0 applications. ADAP provides an ecosystem for building data-driven models using data mining and machine learning. The platform was designed with respect to the needs of manufacturing companies in OWL, which are mostly small or medium-sized enterprises that are specialized in a specific domain and operate highly automated production lines. Due to the needs of these companies, ADAP is created for existing automated manufacturing processes and for data that should be kept secure either by trusted third party providers or by the company itself to provide data sovereignty. However, ADAP is not limited to these needs. The structure of ADAP is composed of a big-data Hadoop framework, implemented in a computing cluster, called Data Analytics Cluster, and an additional machine server for each production environment. In this way, several Industry 4.0 applications can be realized simultaneously, independent of whether all assets belong to the same company or production site. Furthermore, the structure provides scalability and customizability. Encapsulating the Data Analytics Cluster and each machine server into different security zones, combined with security features for authentication, authorization and accounting provided by Kerberos, LDAP and encrypted communication, ensures high levels of IT security. Moreover, different processing steps in ADAP ensure that domain knowledge is incorporated into the Industrie 4.0 application to improve data and model quality, e.g. by converting the OPC UA namespace of a data source into well structured and readable data objects. ADAP was finally tested on a real-world production process with a focus on predictive maintenance and offered great potential for building data-based models. For further work, the processing steps of ADAP could be extended by various question-answering technologies to increase the usability and incorporate more domain knowledge. A key development in Industry 4.0 is the asset administration shell (AAS) that is used to describe physical assets. To enable a more flexible use of ADAP, an integration of the ASS would be suitable. At this state, ADAP is only capable to handle continuous processes. To expand the range of applications, ADAP should be extended to handle discrete processes, e.g. by using data-driven state machine models or hybrid models.

REFERENCES

[1] H. C¸ ebi Bal and C¸ isil Erkan, “Industry 4.0 and Competitiveness,” Procedia Computer Science, vol. 158, pp. 625–631, 2019.

[2] C. Schroder, “The challenges of industry 4.0 for small and medium-sized ¨ enterprises,” Friedrich-Ebert-Stiftung: Bonn, Germany, 2016.

[3] H. Hirsch-Kreinsen, U. Kubach, G. von Wichert, S. Hornung, L. Hubrecht, J. Sedlmeir, and S. Steglich, “Key themes of Industrie 4.0,” acatech – National Academy of Science and Engineering, Tech. Rep., 2019.

[4] S. Godt and W. Schenck, “Studie fur Aufbau und Betrieb eines An- ¨ wendungszentrums als zukunftsfahige Einrichtung f ¨ ur den Wissens- ¨ und Technologietransfer an Unternehmen,” CfADS, Bielefeld University of Applied Sciences, Tech. Rep.

[5] O. Luhr, J. Lambert, J. Struwe, J. Kreißig, J. Liesenfeld, and M. Bloser, ¨ “Green Economy Report North Rhine-Westphalia 2015 - Management Summary,” Ministry for Climate Protection, Environment, Agriculture, Conservation and Consumer Production of the State of North RhineWestphalia, Tech. Rep.

4) A VISUAL DATA SCIENCE SOLUTION FOR VISUALIZATION AND VISUAL ANALYTICS OF BIG SEQUENTIAL DATA

AUTHORS: Carson K. Leung, Yan Wen, Chenru Zhao, Hao Zheng[1], Fan Jiang.

ABSTRACT

In the current era of big data, huge volumes ofvaluable data have been generated and collected at a rapid velocityfrom a wide variety of rich data sources. In recent years, the initiates of open data also led to the willingness of many government, researchers, and organizations to share their data and make them publicly accessible. An example of open big datais healthcare, disease and epidemiological data such as privacy- preserving statistics on patients who suffered from epidemic diseases like the coronavirus disease 2019 (COVID-19). Analyzingthese open big data can be for social good. For instance, analyzing and mining the disease statistics helps people to get a better understanding of the disease, which may inspire them to take partin preventing, detecting, controlling and combating the disease. As“a picture is worth a thousand words”, having the pictorial representation further enhances people’s understanding of the data and the corresponding results for the analysis and mining. Hence, in this paper, we present a visual data science solution for the visualization and visual analytics of big sequential data. We illustrate the ideas through the visualization and visual analytics of sequences of real-life COVID-19 epidemiological data. Oursolution enables people to visualize COVID-19 epidemiological data and their temporal trends. It also allows people to visually analyze the data and discover relationships among popular features associated with the COVID-19 cases. Evaluation of these real-life sequential COVID-19 epidemiological data demonstrates the effectiveness of our visual data science solution in enhancing user experience in the visualization and visual analytics of big sequential data.

BACKGROUND AND RELATED WORKS

As for application to visualization and visual analytics of COVID-19 data, many visualizers and dashboards have been developed since its declaration as a pandemic. Notable ones include (a) World Health Organization (WHO) COVID-19 Dashboard, (b) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), and (c) COVID-19 dashboard by EuropeanCenter for Disease Prevention and Control (ECDC). They provide summary for global COVID-19 situations. Moreover, local governments (e.g., Government of Canada[4]) and media also provides visualizers and dash- boards for local COVID-19 situations. One commonality among these visualizers and dashboards is that they focus on the total numbers of new/confirmed cases and deaths, as well as their cumulative totals. They serve the purpose of fast disseminationof these crucial numbers related to COVID-19 cases. However,there is additional information and knowledge that are embedded in the data and yet to be discovered.

In response, we presented a big data visualization and visualanalytics tool for visualizing frequent patterns from the cumulative COVID-19 statistics. In terms of related works on visualizing frequent patterns, Jentner and Keim surveyed several visualization techniques for frequent patterns. These techniques can be broadly generalized into four categories:

- Lattice representation, which is the most intuitive representation of frequent patterns. With it, frequentpatterns are represented as nodes in a lattice. Immediate supersets and subsets of a frequent pattern are connected by edges.
- Pixel-based visualization, in which multiple frequent k-itemsets of the same length k are represented by a pixel.
- Linear visualization, in which frequent patterns are represented linearly. For example, FIsViz representsa frequent k-itemset in a polyline that connects k nodes in a 2-dimensional space. To avoid bending and crossing-over of polylines, FpVAT represents frequent patterns in a wiring-type diagram.
- Tree visualization, in which frequent patterns are represented according to a tree hierarchy. For example,PyramidViz shows frequent patterns with a side- view of the pyramid, in which short patterns are put on the bottom of the pyramid and longer related patterns are put on the top. Asanother example, FpMapViz shows frequentpatterns with a top-view, in which short patterns are putin the background and longer related patterns are overlayin the foreground.

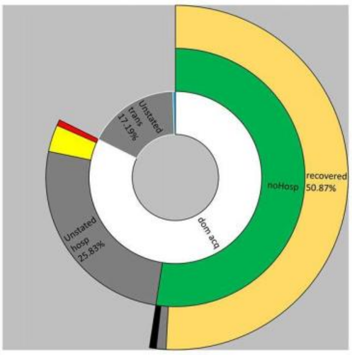

Fig: Visualization of clinical outcomes among those who domestically acquired COVID-19 via community exposures but did not require hospitalization.

OUR VISUAL DATA SCIENCE SOLUTION

To explore and visualize temporal changes, we design and develop a visual data science solution. It first collects and integrates data from a wide variety of rich data sources. It thenpreprocesses the data and builds a temporal hierarchy to generalize the temporal data. Depending on the application, collected data can be of a fine granularity. Having the temporalhierarchy enables us to pick an appreciate level of data granularity. For instance, although dynamic streaming data can be collected at a rapid rate, analyzingdata aggregated at a coarser level may lead to more meaningful and interesting results. As a concrete example, COVID-19 statistics is usually updated on a daily basis. We observed that, analyzing the data on a yearly basis may miss some details, but analyzing them on a daily basis maylead to a huge solution space and may be sensitive to unnecessary fluctuation. Consequently, analyzing them on a weekly basis appear to be appropriate.

After selecting an appropriate level of temporal hierarchy, the next key step is to mine frequent patterns at this level by aggregating frequency counts. The resulting frequent patterns help reveal frequently observed characteristics at a time instance. By repeating the mining procedure over all temporal points, we then compare and contrast similarities and differences among frequent patterns discovered at these temporal points. Asa concrete example, we mine sequences of COVID-19 data on aweekly basis by aggregating their daily counts of various features associated with the data to form the corresponding weekly counts. Then, we discover frequent patterns revealing characteristics (e.g., transmission method, hospital status, clinical outcome) of COVID-19 in a particular week.

In terms of visualization for sequential data, many related works focus on visualizing collections of individual sequences.In contrast, in this paper, we focus on visualizing sequences ofcollections of patterns. For example, instead of analyzing and visualizing the trend of each individual stock, our visual data science solution focuses on visualizing temporal changes in thecomposition of stocks. As another example, for sequences of COVID-19 data, our solution focuses on visualizing temporal changes in the composition of some features.

To visualize temporal changes over compositions of features, it is tempting to stacking all pie charts or outward sunburst diagrams. Given an outward sunburst diagram for a temporal point, one could repeat the mining and visualization process to generate multiple sunburstdiagrams (with one diagram for each temporal point). While thestack of sunburst diagrams capture all information for analysisof temporal changes, it may be challenges to view and comprehend the details for each diagram, let alone discoveringtheir temporal changes.

CONCLUSIONS

In this paper, we presented a visual data science solution forthe visualization and visual analytics of big sequential data. Weillustrate the ideas through its applications to real-life COVID-19 epidemiological data. Our solution represents compositionsof combinations of feature values in stacked columns, which enables easy comparison in the temporal dimension. Although we evaluated and showed its practicality by using COVID-19 data, it can be applicable to visualization and visual analytics ofother big sequential data. As ongoing and future work, we further enhance visibility, interpretability and explainability ofour visual data science solution in visualization and visual analytics of big sequential data.

REFERENCES

[1] C Ordonez, et al., “An intelligent visual big data analytics framework for supporting interactive exploration and visualization of big OLAP cubes,” IV 2020, pp. 421-427.

[2] A Perrot, et al., “HeatPipe: high throughput, low latency big data heatmap with Spark streaming,” IV 2017, pp. 66-71.

[3] I.M. Anderson-Grégoire, et al., “A big data science solution for analytics on moving objects,” AINA 2021, vol. 2, pp. 133-145.

[4] A.H. Diallo, et al., “Proportional visualization of genotypes and phenotypes with rainbow boxes: methods and application to sickle cell disease,” IV 2019, Part I, pp. 1-6.

[5] S. Hamdi, et al., “Intra and inter relationships between biomedical signals: a VAR model analysis,” IV 2019, Part I, pp. 411-416.