Big Data in Social and Economic Analyses

Abstract

The passage highlights the challenges posed by the ever-increasing volume and variety of data generated from day-to-day business operations, along with the integration of social data, which traditional statistical approaches struggle to handle due to their development before the internet era. To address this, academics and researchers have been developing advanced and complex analytics techniques that can provide valuable insights for the commercial sector. The chapter focuses on the fundamental characteristics that form the foundation of social big data analytics (SBD) and presents a framework for predictive analytics within this context.

Predictive analytics is particularly emphasized as a crucial component of SBD, as it enables businesses to make informed decisions and predictions based on data patterns and trends. The framework for SBD predictive analytics likely includes various methodologies, tools, and algorithms that aid in making accurate predictions and extracting meaningful insights from the vast amount of social data.

Additionally, the chapter likely discusses various predictive analytical algorithms, their implementation in essential applications, and the use of top-tier tools and APIs to facilitate the analysis of social data. These algorithms and tools help businesses extract valuable information from social big data, leading to better decision-making and strategic planning.

To reinforce the significance and practicality of predictive analytics in the context of SBD, the chapter might present experiments and case studies. These experiments showcase real-world applications and demonstrate how predictive analytics can be leveraged effectively to gain insights from social data and drive meaningful business outcomes.

Overall, the chapter aims to shed light on the importance of predictive analytics in handling social big data, showcasing its potential benefits and how it can be practically applied to solve real-world problems and improve business operations.

Keywords: ML, Predictive Learning, Integrity

Authors:

Mahadevi Somnath Namose

Research Scholar, Department of Computer Science Sardar Patel University, Bhopal, MP, India

Dr. Tryambak Hiwarkar

Professor, Department of Computer Science2 Sardar Patel University, Bhopal, MP, India

I. INTRODUCTION

The socio-economic system is a paradigm which can be defined as complex because behavior change of free agents can result in other individuals and organizations experiencing chaotic dynamics, non-linear interactions and other cascading effects. To embrace sustainability in such a paradigm requires the adoption of emerging technologies that can lead to the next quantum leap including the big data platform. Big data provides us with a stream of new and digitized data exploring the interactions between individuals, companies and other organizations. However, to understand the underlying behavior of social and economic agents, organizations and researchers must manage large quantities of unstructured and heterogeneous data. To succeed in this undertaking requires careful planning and organization of the entire process of data analysis, taking into account the particularities of social and economic analyses such as the wide variety of heterogeneous sources of information and the existence of strict governance policy.

In recent years, many tools for both qualitative and quantitative models have been developed to describe and better understand complex systems. These tools include stochastic and dynamic systems, multivariate statistics, network models, social network analysis, inference and stochastic processes, fuzzy theory, relational calculus, partial order theory, multi-criteria decision methods and other tools which have been widely used to address problems in socio-economic systems. Traditional quantitative methods for acquiring socioeconomic data are limited in their ability to examine the complexities of socio-economic systems. Therefore, big data collected from satellites, mobile phones, and social media, among other data sources, allow researchers to build on and sometimes replace traditional methods providing greater frequency and timeliness, accuracy and objectiveness as well as defining sustainable models.

II. Socio-Economic Indicators

The importance of understanding and predicting the volatility of socio-economic indicators, particularly in developing nations. You highlight the negative impact of volatility on economic health, such as the drain on national coffers and the exacerbation of poverty. You also mention that commodity exports play a significant role in the revenue of certain countries, and fluctuations in commodity prices can have devastating effects. Additionally, you note that currency exchange rate instability can affect the cost of commodities.

To better comprehend and anticipate the volatility of socio-economic indicators, economists and computational scholars have employed various approaches. Economists often rely on established economic models to analyze and predict these variables. On the other hand, computational scholars have

turned to computational modeling tools, particularly when examining structured time series data.One significant advancement in recent years has been the explosion of unstructured data streams on the internet, which provides a wealth of information for understanding economic and social shifts. Additionally, the development of cutting-edge computational linguistics algorithms has further enhanced the ability to analyze this unstructured data. These algorithms can automatically infer events, create knowledge graphs, and forecast outcomes based on unstructured news streams.

Overall, the integration of big data analytics algorithms with unstructured data sources holds promise for improving our understanding of socio-economic volatility and facilitating more accurate predictions in the field. In our chapter, our aim to address the challenge of dealing with unorganized text data on the web by mining relevant information and providing a concise and accurate depiction of the collected data. By extracting data from these texts, you propose a new approach that focuses on events as a lower dimensional representation of the information reported in web papers. Viewing events as discrete data points allows for a more precise representation of the occurrences taking place globally.

In our chapter also emphasizes the importance of understanding the connections between these events. By analyzing the context of time and space, hidden connections among events become apparent. To uncover these connections, you introduce a method for constructing knowledge graphs, which visually illustrate the relationships between various events.

Furthermore, in conjunction with observed time-series data of socio-economic indicators, you utilize the lower-dimensional representation of web text (events) and the latent relationships between them (knowledge graphs) to gain insights into the factors driving these indicators. By understanding the types of events that influence these phenomena, you aim to develop models that can analyze, describe, and forecast shifts in socio-economic indicators. Overall, in our chapter focuses on leveraging the structured representation of events and knowledge graphs derived from unstructured web text to enhance the understanding and prediction of socio-economic phenomena.

In our chapter, you aim to address the challenge of dealing with unorganized text data on the web and provide a concise and accurate representation of this vast amount of information. By mining relevant data from these texts, you propose new approaches to depict events, which serve as a more precise and lower-dimensional representation of the information found in web papers reporting on events and occurrences worldwide. Each of these events can be seen as discrete data points.

By analyzing events in the context of time and space, you highlight that many hidden connections between events become evident. To uncover these connections, you introduce a method for constructing

knowledge graphs, which visually illustrate the relationships between various events.In conjunction with observed time-series data of socio-economic indicators, such as fluctuations in indicators, you utilize the lower-dimensional and precise representation of events derived from web text and the latent relationships captured in knowledge graphs. The goal is to understand the types of events driving these phenomena and leverage them for prediction purposes.

In our Chapter compares two forms of information at a high level: structured data of various socio-economic metrics and unstructured news feeds. Analyzing unstructured data presents challenges due to their sheer volume, noise, and lack of consistent patterns. However, your techniques have been developed to handle large datasets effectively and robustly, taking into account these challenges. The information extracted from unstructured news is condensed using events, visualization frameworks, and knowledge graphs, which can be understood by subject-matter specialists. This allows for the integration of third-party technologies with your data and enables the creation of forecasting and analysis tools for macroeconomic indicators.

The merging of structured and unstructured data presents additional difficulties, as these two forms of data have distinct structures and features. In our chapter addresses the challenges associated with efficiently and successfully combining them. Figure 1.5 provides an overview of the different condensed forms of online news and illustrates the overall flow of the process through these components and their interactions.

What exactly is “big data” in the context of economic applications? It can be defined as datasets that require advanced computing hardware and/or software tools to conduct the analysis. One such tool is distributed computing that shares the processing of a task across several machines, instead of a single machine as typically done by economists. Examples of large datasets used in economic analysis are administrative data (e.g. tax records for the whole population of a country), commercial datasets (e.g. consumer panels), and textual data (e.g. such as Twitter or news data) just to mention a few. In some cases, the datasets are structured and ready for analysis, while in other cases (e.g. text), the data is unstructured and requires a preliminary step to extract and organize the relevant information. As discussed in Einav and Levin (2014), economists are still in the early stages of analysing big data and are learning from developments in other disciplines. In particular, there is renewed interest in machine learning (ML) algorithms after the early applications of the 1990s (Kuan & White, 1994). Varian (2014) discusses techniques that can be used to analyse large datasets.

How can big data contribute to a better understanding of the economy and to support policy? In the highly aggregate context of macroeconomic analysis, big data offer the opportunity to bring to light

the heterogeneity in consumers and firms that is typically neglected in official statistics. The high granularity of big datasets can be exploited to construct indicators that are better designed to explain certain phenomena, for example, along a geographic or demographic dimension. In addition, many economic models make assumptions about deep behavioural parameters that are difficult to estimate without detailed datasets. An example is represented by the work of Chetty et al. (2014b) where individual information about the school performance of a child is matched to his/her path of future earnings derived from tax data of the Internal Revenue Service (IRS). In other situations, big data allow to measure quantities that we could not measure until now. A field that is benefiting from these alternative sources of data is development economics. For instance, Storeygard (2016) uses night-light satellite data to estimate the income of sub-Saharan African cities.

Another important dimension in which big data can contribute to economic analysis is by offering information that is not only more granular but also more frequent in the time dimension. At times when economic conditions are rapidly changing, policy-makers need an accurate measure of the state of the economy to design the appropriate policy response. An example is provided by the early days of the Covid-19 pandemic in March 2020 when policy-makers felt the pressure to act in support of the economy despite the lack of official statistics to measure the extent of the slump, as discussed by Barbaglia et al. (2022). Many relevant economic indicators are observed infrequently, such as gross domestic product (GDP) at the quarterly frequency and the unemployment rate and the industrial production index at the monthly frequency. In addition, these variables are released with delays that range from a few days to several months. For these reasons, big data have the potential to produce indicators of business conditions that are more accurate and timely

III. CONCLUSION

The passage discusses the feasibility of using news articles as a data source for event extraction and the subsequent construction of knowledge graphs to represent event relationships in different analytic contexts. The dissertation introduces two types of event models and five strategies for building knowledge graphs, each tailored to varying levels of detail. These knowledge graphs can serve as valuable resources for both manual and automated analyses of news data.Event extraction involves identifying specific events or incidents from unstructured text, such as news articles. By extracting these events and representing their relationships through knowledge graphs, researchers and analysts can gain a deeper understanding of the interconnectedness between different events in the news domain.

The passage also highlights the various applications of event models and knowledge graphs. One significant application is the construction of prediction models for extraneous variables and indicators. By leveraging the information within knowledge graphs, researchers can build predictive models that have real-world implications. These models can help in forecasting future events, trends, or outcomes based on the relationships between events captured in the knowledge graphs.Overall, the dissertation likely explores the potential benefits of using news articles as a data source for event extraction and knowledge graph construction. It emphasizes the practicality of this approach and how it can be applied to a range of analytic contexts, enabling more informed decision-making, trend analysis, and predictive modeling in various domains. The utilization of knowledge graphs that incorporate event relationships from news data may prove to be a valuable tool in uncovering insights and predicting outcomes in real-world scenarios.

REFERENCES

1) Al-Dohuki, S., Wu, Y., Kamw, F., Yang, J., Li, X., Zhao, Y., Ye, X., Chen, W., Ma, C., Wang, F.: SemanticTraj: a new approach to interacting with massive taxi trajectories. IEEE Trans. Vis. Comput. Graph. 23(1), 11–20 (2017). http://doi.org/10.1109/TVCG.2016.2598416. http://doi.ieeecomputersociety.org/10.1109/TVCG.2016.2598416
2) Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM 26(11), 832–843 (1983). http://doi.org/10.1145/182.358434. http://doi.acm.org/10.1145/182.358434
3) Alvares, L.O., Bogorny, V., Kuijpers, B., de Macêdo, J.A.F., Moelans, B., Vaisman, A.A.: A model for enriching trajectories with semantic geographical information. In: GIS, p. 22 (2007)
4) Andrienko, G., Andrienko, N., Bak, P., Keim, D., Wrobel, S.: Visual Analytics of Movement. Springer Publishing Company, Incorporated, Berlin (2013)
5) Andrienko, G., Andrienko, N., Chen, W., Maciejewski, R., Zhao, Y.: Visual analytics of mobility and transportation: state of the art and further research directions. IEEE Trans. Intell. Transp. Syst. 18(8), 2232–2249 (2017). http://doi.org/10.1109/TITS.2017.2683539
6) Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I., Jern, M., Kraak, M.J., Schumann, H., Tominski, C.: Space, time and visual analytics. Int. J. Geogr. Inf. Sci. 24(10), 1577–1600 (2010). https://doi.org/10.1080/13658816.2010.508043
7) Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M., MacEachren, A., Wrobel, S.: Geovisual analytics for spatial decision support: setting the research agenda. Int. J. Geogr. Inf. Sci. 21(8), 839–857 (2007). https://doi.org/10.1080/1365881070134901
8) Andrienko, N., Andrienko, G.: Visual analytics of movement: an overview of methods, tools and procedures. Inf. Vis. 12(1), 3–24 (2013). https://doi.org/10.1177/1473871612457601

9) Baglioni, M., de Macêdo, J.A.F., Renso, C., Trasarti, R., Wachowicz, M.: Towards semantic interpretation of movement behavior. In: Advances in GIScience, pp. 271–288. Springer, Berlin (2009)
10) Bogorny, V., Renso, C., de Aquino, A.R., de Lucca Siqueira, F., Alvares, L.O.: Constant - a conceptual data model for semantic trajectories of moving objects. Trans. GIS 18(1), 66–88 (2014)
11) C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens. A time-based collective factorization for topic discovery and monitoring in news. In Proceedings of the 23rd international conference on World wide web, pages 527–538. International World Wide Web Conferences Steering Committee, 2014.
12) Chu, D., Sheets, D.A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., Chen, G.: Visualizing hidden themes of taxi movement with semantic transformation. In: Proceedings of the 2014 IEEE Pacific Visualization Symposium, PACIFICVIS '14, pp. 137–144. IEEE Computer Society, Washington, DC (2014). http://dx.doi.org/10.1109/PacificVis.2014.50
13) D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
14) F. Hogenboom, M. de Winter, F. Frasincar, and U. Kaymak. A news event-driven approach for thehistorical value at risk method. Expert Systems with Applications, 42(10):4667 – 4675, 2015
15) Fileto, R., May, C., Renso, C., Pelekis, N., Klein, D., Theodoridis, Y.: The baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. Data Knowl. Eng. 98, 104–122 (2015)
16) H. Tanev, J. Piskorski, and M. Atkinson. Real-time news event extraction for global crisis monitoring. NLDB '08, pages 207–218, 2008
17) Hamad, K., Quiroga, C.: Geovisualization of archived ITS data-case studies. IEEE Trans. Intell. Transp. Syst. 17(1), 104–112 (2016). https://doi.org/10.1109/TITS.2015.2460995
18) Hu, Y., Janowicz, K., Carral, D., Scheider, S., Kuhn, W., Berg-Cross, G., Hitzler, P., Dean, M., Kolas, D.: A geo-ontology design pattern for semantic trajectories. In: Tenbrink, T., Stell, J., Galton, A., Wood, Z. (eds.) Spatial Information Theory, pp. 438–456. Springer International Publishing, Cham (2013)
19) K. Radinsky and E. Horvitz. Mining the web to predict future events. WSDM '13, pages 255–264.ACM, 2013.
20) Kotis, K., Vouros, G.A.: Human-centered ontology engineering: the HCOME methodology. Knowl. Inf. Syst. 10(1), 109–131 (2006)

21) Kraak, M., Ormeling, F.: Cartography: Visualization of Spatial Data, 3 edn. Guilford Publications, New York (2010)
22) M. Liu, Y. Liu, L. Xiang, X. Chen, and Q. Yang. Extracting key entities and significant events from online daily news. IDEAL ’08, pages 201–209, 2008.
23) M. Okamoto and M. Kikuchi. Discovering volatile events in your neighborhood: Local-area topicextraction from blog entries. AIRS ’09, pages 181–192, 2009
24) Nogueira, T.P., Martin, H.: Querying semantic trajectory episodes. In: Proc. of MobiGIS, pp. 23–30 (2015)
25) Paiva Nogueira, T., Bezerra Braga, R., Martin, H.: An ontology-based approach to represent trajectory characteristics. In: Fifth International Conference on Computing for Geospatial Research and Application. Washington, DC (2014). https://hal.archives-ouvertes.fr/hal-01058269
26) Parent, C., Spaccapietra, S., Renso, C., Andrienko, G.L., Andrienko, N.V., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., de Macêdo, J.A.F., Pelekis, N., Theodoridis, Y., Yan, Z.: Semantic trajectories modeling and analysis. ACM Comput. Surv. 45(4), 42 (2013)
27) Peuquet, D.J.: It’s about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. Ann. Assoc. Am. Geogr. 84(3), 441–461 (1994)
28) Santipantakis, G., Doulkeridis, C., Vouros, G.A., Vlachou, A.: Masklink: Efficient link discovery for spatial relations via masking areas, arXiv:1803.01135v1 (2018)
29) Santipantakis, G., Vouros, G., Glenis, A., Doulkeridis, C., Vlachou, A.: The datAcron ontology for semantic trajectories. In: ESWC-Poster Session (2017)
30) Santipantakis, G.M., Glenis, A., Kalaitzian, N., Vlachou, A., Doulkeridis, C., Vouros, G.A.: FAIMUSS: flexible data transformation to RDF from multiple streaming sources. In: Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, 26–29 March 2018, pp. 662–665 (2018). https://doi.org/10.5441/002/edbt.2018.79
31) Santipantakis, G.M., Kotis, K.I., Vouros, G.A., Doulkeridis, C.: RDF-gen: Generating RDF from streaming and archival data. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, WIMS ’18, pp. 28:1–28:10. ACM, New York (2018). http://doi.org/10.1145/3227609.3227658. http://doi.acm.org/10.1145/3227609.3227658
32) Santipantakis, G.M., Vouros, G.A., Doulkeridis, C., Vlachou, A., Andrienko, G.L., Andrienko, N.V., Fuchs, G., Garcia, J.M.C., Martinez, M.G.: Specification of semantic trajectories supporting data transformations for analytics: The datacron ontology. In: Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, 11–14 Sept 2017, pp. 17–24 (2017). https://doi.org/10.1145/3132218.3132225

33) Soltan Mohammadi, M., Mougenot, I., Thérèse, L., Christophe, F.: A semantic modeling of moving objects data to detect the remarkable behavior. In: AGILE 2017. Wageningen University, Chair group GIS & Remote Sensing (WUR-GRS), Wageningen (2017). https://hal.archives-ouvertes.fr/hal-01577679

34) Spaccapietra, S., Parent, C., Damiani, M.L., de Macêdo, J.A.F., Porto, F., Vangenot, C.: A conceptual view on trajectories. Data Knowl. Eng. 65(1), 126–146 (2008)

35) T. Hofmann. Probabilistic latent semantic indexing. SIGIR '99, pages 50–57, 1999

36) Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. In: Survey Review XXII, pp. 88–93 (1975). https://doi.org/10.1179%2Fsre.1975.23.176.88. https://www.ngs.noaa.gov/PUBS_LIB/inverse.pdf

37) Vouros, G., Santipantakis, G., Doulkeridis, C., Vlachou, A., Andrienko, G., Andrienko, N., Fuchs, G., Martinez, M.G., Cordero, J.M.G.: The datacron ontology for the specification of semantic trajectories: specification of semantic trajectories for data transformations supporting visual analytics. J. Data Semant. 8 (2019). http://doi.org/10.1007/s13740-019-00108-0. http://link.springer.com/article/10.1007/s13740-019-00108-0

38) Wen, Y., Zhang, Y., Huang, L., Zhou, C., Xiao, C., Zhang, F., Peng, X., Zhan, W., Sui, Z.: Semantic modelling of ship behavior in harbor based on ontology and dynamic Bayesian network. ISPRS Int. J. Geo-Inf. 8(3) (2019). http://doi.org/10.3390/ijgi8030107. http://www.mdpi.com/2220-9964/8/3/107

39) Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. In Proceedings of the 18th ACM SIGKDD international conference on Knowledgediscovery and data mining, pages 123–131. ACM, 2012.

40) Yan, Z., Macedo, J., Parent, C., Spaccapietra, S.: Trajectory ontologies and queries. Trans. GIS 12(s1), 75–91 (2008). http://doi.org/10.1111/j.1467-9671.2008.01137.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2008.01137.x