[DM-01-070] Problems of Large Spatial Databases

Abstract

Large spatial databases often labeled as geospatial big data exceed the capacity of commonly used computing systems as a result of data volume, variety, velocity, and veracity. Additional problems also labeled with V's are cited, but the four primary ones are the most problematic and focus of this chapter (Li et al., 2016, Panimalar et al., 2017). Sources include satellites, aircraft and drone platforms, vehicles, geosocial networking services, mobile devices, and cameras. The problems in processing these data to extract useful information include query, analysis, and visualization. Data mining techniques and machine learning algorithms, such as deep convolutional neural networks, often are used with geospatial big data. The obvious problem is handling the large data volumes, particularly for input and output operations, requiring parallel read and write of the data, as well as high speed computers, disk services, and network transfer speeds. Additional problems of large spatial databases include the variety and heterogeneity of data requiring advanced algorithms to handle different data types and characteristics, and integration with other data. The velocity at which the data are acquired is a challenge, especially using today's advanced sensors and the Internet of Things that includes millions of devices creating data on short temporal scales of micro seconds to minutes. Finally, the veracity, or truthfulness of large spatial databases is difficult to establish and validate, particularly for all data elements in the database.

Keywords: data mining, geospatial big data, spatial database

Author & citation

Usery, E. L. (2019). Problems of Large Spatial Databases. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2019 Edition), John P. Wilson (Ed.). DOI: 10.22224/gistbok/2019.2.13

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Problems of large spatial databases. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions

geospatial big data: datasets with locational identifiers that exceed the capacity of current computing systems to manage, process, or analyze the data with reasonable effort

2. Introduction

Large spatial databases including geospatial data for images, maps, and other data used in geographic information system and location-based applications most often are referred to currently (2019) as **geospatial big data**. By definition, geospatial big data are datasets with locational identifiers that exceed the capacity of current computing systems to manage, process, or analyze the data with reasonable effort. These excesses result from volume, variety, velocity, veracity, and other Vs of the spatial data (Firican, 2017; Panimalar et al., 2017). Traditional forms of large spatial databases are vector-formatted points and linework and raster images, including satellite images and aerial photographs. Additional forms of large databases appearing now with locational components include a plethora of new sensors, such as lidar and other electronic sensors, video systems, human sensors with cell phones and other data collections. Data collection by these sensors result in Volunteered Geographic Information (VGI) and a wealth of social media data with locations, including Twitter feeds and photographic archives. Although these big data sources often include location information, they exceed current capacities for processing, much less spatial analysis. Geospatial big data provide a major source for innovation, competition and productivity, but simultaneously are problematic for data handling, processing, storage, retrieval, and use. The challenges confronting the management and application of geospatial big data have necessitated the development of new software tools and techniques, as well as parallel computing hardware architectures to meet the data handling requirements (Wright & Wang, 2011; Wang & Goodchild, 2019).

3. Classification

Geospatial big data may be classified into groups with similar data characteristics. A basic classification, after Yao and Li (2018), using types of data for organization, follows.

3.1 Remote sensing

Traditionally, from the initial launch of satellite systems, such as Landsat 1, to the modern day high spatial and temporal resolutions to new sensor designs, remote sensing data form a class of geospatial big data. Current systems acquire multispectral and hyperspectral images, in multiple resolutions and multiple temporal acquisition dates from multiple sensor systems, resulting in big data volumes in addition to other obstacles to their management and use (Chi et al., 2016). Collected and archived in traditional raster formats, these customarily continuous-tone images usually exceed current computer processing capabilities requiring new solutions such as parallel computing and enhanced distributed cyberinfrastructure (Ma et al., 2015).

3.2 Surveying and mapping data

Geographical data, often referenced to the Global Navigation Satellite Systems (GNSS), includes industrial infrastructure graphics, thematic maps, digital map and image products,

such as Digital Line Graphs, Digital Elevation Models, Digital Raster Graphics and Digital Orthophotographs, data for most national mapping organizations, such as The National Map of the U.S. Geological Survey, land use, and other basic surveying and mapping data. In recent years, high resolution lidar data have become the dominant form of mapping data and the data volumes (terabytes to petabytes) result in geospatial big data (Sugarbaker et al., 2014). Mobile mapping with moving sensors and sensor fields have also impacted the big data problem in the mapping industry.

3.3 Location-based data for location services

Location-based data typically include geographic and human social information data with spatial and temporal locations. These data are primarily acquired with GNSS inputs generated with smart phones, field collected data, and human and traffic trajectories. Examples include restaurant locations and ratings, highway and street representations, personal route traces, photo locations, and others. These geospatial big data have become a critical resource in socially-based service industries, vehicle routing, and other activities that sense the activities of human groups (Liu et al., 2015).

3.4 Social media platforms

These data are Internet and Web-based and include geospatial location. These locations may be specified on Web page data, or in social media platforms, such as Twitter, Facebook, Google Plus, and other social platforms. Location specification may be as specific as a set of global positioning coordinates or as vague as a simple disambiguated place name. Geospatial big data from social media carry a host of problems for data handling, such as location dereferencing, failure to meet basic statistical assumptions of independent, unbiased samples, and others (Goodchild, 2013; Kitchin, 2013; Tsou, 2015).

3.5 Internet of Things (IoT)

The IoT includes sensors that monitor and collect data including environmental and atmospheric measurements, water quality, volumes, and flow, intelligent devices in the household, in field collection for science and management, wearable devices, and a host of sensors that contribute real-time data in microseconds to minutes to big data servers. Most of these data acquired by new sensors are data streams of arbitrary high density; include many different dimensional measurements, such as optical, acoustical, and mechanical; and have different positional accuracies and precision. Compared to traditional data on the Internet, IoT data are generally of much greater variety and much greater frequency leading to a true geospatial big data problem (Alelaiwi, 2017).

4. Big Data Challenges

These include architectures for processing the large data volumes as well as inherent problems in the data such as heterogeneity, vagueness of geographic feature definitions and boundaries, and uncertainty of position or attribution. Among the identified processing problems are quality assessment, data modeling and structuring, functional programming for geospatial big data streams, geospatial big data analytics, data mining and knowledge discovery, and geospatial big data visualization and visual analytics (Li et al., 2016).

4.1 Inherent geospatial big data problems

All of the types of geospatial big data listed above suffer from problems common to all geospatial data which include vagueness and indeterminate boundaries for geographic features (Usery, 1995). For example, using lidar data for terrain feature extraction immediately requires definition and conceptualization of the terrain features themselves. That is, where are the boundaries of the natural geographic feature that can be used to extract the entity from the data. Couple this indeterminacy with the big data problem and the natural vagueness of the features become more problematic with the large data volumes and heterogenous data. Another inherent problem is uncertainty of geospatial big data and the lack of methods and specifications for measuring and quantifying such uncertainties. Many of these datasets lack normal scientific standards of replicability and rigorous sampling (Goodchild, 2013).

4.2 Quality Assessment

Geospatial big data are often continuous measurements, such as pixel values in satellite images or lidar returns, and, as abstractions of real time variables, uncertain. The sheer volume of data magnifies the uncertainty and requires quality assessment to assure appropriate abstraction, processing, feature extraction, analysis and visualization processing. Standards and rigorous procedures for monitoring quality have been developed for many traditional structured geospatial datasets. For example, the international standard ISO 19157 specifically addresses the quality of geospatial data. A similar standard exists for lidar data collections: the U.S. Geological Survey Lidar Base Specification 1.3 (Heidemann, 2018). These and other similar standards frameworks define quantitative measures of data quality, including spatial, temporal and thematic accuracy, spatial, temporal, and thematic resolution, consistency, and completeness (Li et al., 2016). Thus, data quality assessment is defined for the collection processes and creation of metadata reflecting the quality of those processes. Similar standards do not exist for the majority of geospatial big data. For example, assessment of the quality of social media data is extremely difficult since many of these data violate basic statistical assumptions and are often obtained through biased and limited sampling, for example, those who own cell phones and those who participate in the data collection process for a particular type of big data. Quality of locations in geospatial big data is also problematic, and often the locational component, as with Twitter feeds, is determined from context rather than through rigorous surveying methods.

4.3 Geospatial Big Data Streams

Twitter, Facebook, and other social media platforms provide continuous streams of geospatial big data. These streams require continual processing and analysis. The geospatial components of these streams are often hidden, ambiguous, and only determined by the context of the message or the scenes of a photograph. Thus, the spatial component is inherently uncertain and must be resolved in any analytics or visualizations related to these types of data. Similar big data streams are collected from sensor fields and objects on the Internet of Things. With these sensor fields, high-volume data streams are more problematic than the spatial location since most of most of these devices have fixed locations for the sensors and that can be used in analytics and visualization.

5. Geospatial Big Data Management and Processing

The volume and characteristics of spatial data and the velocity with which some data are collected present impediments to their management and processing. Among the solutions for geospatial big data processing are data organization methods such as indexing based on spatial location, feature identifiers, or thematic or temporal attribution.

5.1 Data modeling and structuring

A common approach to handling and processing geospatial big data is dependent upon data model and structure, particularly in traditional vector and raster formats. For example, vector data consists of points, lines, and polygons, and sometimes volumetric figures. Standard arc-node based models have been developed with high degrees of indexing that allows direct access to the geometric, topologic, thematic, and temporal attributes and relationships of these geospatial features. Raster data can be supported with a variety of indexing methods such as encoding using space-driven structures such as linear trees, quadtrees and KD-trees, or data-drive structures such as the R-tree (Samet, 2006). The use of ontologies and semantics to structure and index geospatial big data now offers the potential to support modeling, analytics, and visualization (Zhang et al., 2017).

5.2 Geospatial big data analytics

Analytics often involve distributed and high-performance computing systems with algorithms adapted for parallel computation, processing, and visualization (Wang & Goodchild, 2019). Often approaches involve data mining and knowledge discovery. Methods of approach include parametric statistics, which require assumptions of a probability distribution function and often randomness and independence of samples; non-parametric statistics which simply assume local smoothness; and functional analysis including wavelets and spatial data generalization, spatial data clustering, and mining spatial association rules. Machine learning and the use of deep convolutional neural networks for geospatial big data mining and knowledge discovery are often implemented in high performance computing environments. These methods can be applied to the large variety of geospatial big data including the vast archives of satellite and other Earth observation imagery, sensor field data, social media, and the data feeds from the Internet of Things (Vatsavai et al., 2012).

5. Summary and Conclusions

Geospatial big data exceed the capacity of commonly used computing systems as a result of data volume, variety, velocity, and veracity. Many data types, including remotely-sensed images from satellite, aircraft, and hand held platforms; mapping data types including lidar, elevation, hydrography, land cover classifications, and others; location-based data for location services highly dependent on GNSS coordinates; social media feeds such as Twitter and images from Facebook; and sensor streams from sensor fields and the IoT, are examples of geospatial big data with extensive volumes and heterogeneity. Challenges and problems of geospatial big data include computer processing architectures, requirements for parallel and distributed computational systems, high performance computing, basic uncertainty of the spatial and non-spatial components of the data, quality assessment, analytical processing and statistical assumptions and violations, and analysis and visualization of large datasets with clusters of significant context. The future of geospatial big data will witness additional growth in the volume and variety of data and the harnessing of computational methods to turn geospatial big data into science results and applications

for society.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Adelaiwi, A. (2017). A Collaborative Resource Management Tool for Big IoT Data Processing in Cloud. Cluster-Computing: The Journal of Networks, Software, Tools and Applications, 20(2), 1791-1799

Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, I., & Zhu, Y. (2016). Big Data for Remote Sensing: Challenges and Opportunities. Proceedings of the Institute for Electrical and Electronics Engineers, 104(11), p. 2207-2219.

Firican, G. (2017). The 10 Vs of Big Data. Transforming Data with Intelligence (TDWI). Blog entry, retrieved September 11, 2018.

Goodchild, M. F. (2013). The Quality of Big (Geo)data. Dialogues in Human Geography, 3(3), 280-284.

Heidemann, H. K. (2018). Lidar Base Specification (ver. 1.3, February 2018): U.S. Geological Survey Techniques and Methods, book 11, chap. B4, 101.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. Dialogues in Human Geography, 3(3), 262-267.

Lee, J., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. Big Data Research 2(2): 74–81.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. ISPRS Journal of Photogrammetry and Remote Sensing 115, 119–33.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. Annals of the Association of American Geographers, 105(3), 512–530.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote Sensing Big Data Computing: Challenges and Opportunities, Future Generation Computer Systems, 51, 47-60.

Panimalar, A., Varnekha, S., & Veneshia, K. (2017). The 17 Vs of Big Data. International Research Journal of Engineering and Technology (IRJET), 4(9). Retrieved September 11, 2018.

Samet, H. (2006). Foundations of Multidimensional and Metric Data Structures. Morgan Kaufmann Publishers, 1024.

Sugarbaker, L. J., Constance, E. W., Heidemann, H. K., Jason, A. L., Lukas, V., Saghy, D. L., & Stoker, J. M. (2014). The 3D Elevation Program initiative—A call for action. U.S. Geological Survey Circular 1399, 35.

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. Cartography and Geographic Information Science, 42(1), 70-74.

Usery, E. L. (1996). A Conceptual Framework and Fuzzy Set Implementation for Geographic Features, In P. A. Burrough and A.U. Frank (Eds.), Geographic Objects with Indeterminate Boundaries. Taylor and Francis: London, pp. 71-85.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). Spatiotemporal Data Mining in the Era of Big Geospatial Data: Algorithms and Applications. BigSpatial '12 Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. Redondo Beach, CA, 1-10.

Wright, D. J. and Wang, S. (2011). The Emergence of Spatial Cyberinfrastructure. Proceedings of the National Academy of Sciences USA, 108(14): 5488–5491.

Yao, X., & Li, G. (2018). Big Spatial Vector Data Management: A Review. Big Earth Data, 2(1), 108-129.

Zhang, C., Zhao, T., & Li, W. (2017). Big Geospatial Data and Geospatial Semantic Web: Current State and Future Opportunities. In Y. Wu, F. Hu, G. Min, and A. Zomaya (Eds.), Big Data and Computational Intelligence in Networking. Taylor & Francis LLC, CRC Press. pp 43-64.