Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding

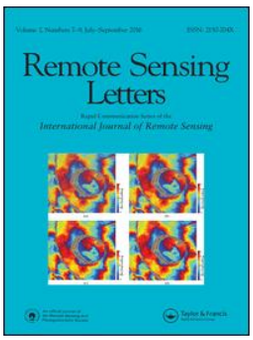
Article *in* Remote Sensing Letters · February 2019

DOI: 10.1080/2150704X.2018.1530480

3 authors, including:

Keying Huang
Chinese Academy of Sciences
3 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Guoqing Li
Chinese Academy of Sciences
120 PUBLICATIONS **625** CITATIONS
SEE PROFILE



Remote Sensing Letters

ISSN: 2150-704X (Print) 2150-7058 (Online) Journal homepage: http://www.tandfonline.com/loi/trsl20

Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding

Keying Huang, Guoqing Li & Jian Wang

To cite this article: Keying Huang, Guoqing Li & Jian Wang (2018) Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding, Remote Sensing Letters, 9:11, 1070-1078, DOI: 10.1080/2150704X.2018.1508907

To link to this article: https://doi.org/10.1080/2150704X.2018.1508907

Published online: 11 Sep 2018.

REMOTE SENSING LETTERS
2018, VOL. 9, NO. 11, 1070–1078
https://doi.org/10.1080/2150704X.2018.1508907

Taylor & Francis
Taylor & Francis Group

Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding

Keying Huang[a,b], Guoqing Li[a] and Jian Wang[a]

[a]Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China; [b]University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

With the rapid development of Earth observation technology, satellite data centres have accumulated large amounts of remote sensing data from different spaceborne and airborne sensors. The efficient management and quick retrieval of multisource, massive and heterogeneous remote sensing data in the Big Data age have become increasingly important. In this paper, a spatio-temporal organization model based on GeoHash coding is proposed. First, based on the ISO standard, the heterogeneous remote sensing metadata can be converted into a unified format, and the differences in the multisource remote sensing metadata are screened. Then, the GeoHash algorithm is used to encode and convert the latitude and longitude coordinates of the remote sensing metadata to reduce the remote sensing metadata dimensions under space retrieval conditions. Finally, by building an HBase key value model based on GeoHash, a primary key is used to realize the rapid retrieval of massive remote sensing metadata through the simulation of 1500 million remote sensing metadata retrieval experiments; by comparing with the traditional multi-conditional filtering retrievals, the results show that a spatio-temporal organization strategy for remote sensing metadata based on GeoHash coding can effectively improve the efficiency of remote sensing data retrievals.

ARTICLE HISTORY
Received 9 April 2018
Accepted 29 July 2018

1. Introduction

With the development of earth observation technology, a series of sensors with high spectral resolutions, high spatial resolutions, high temporal resolutions, multipolarization and multi-angle capabilities have been launched, resulting in the decrease in remote sensing data acquisition periods and an increase in timeliness, which has led to an explosive growth in data volume (He et al. 2017; Yan et al. 2017). Currently, China has satellite data centres for varying fields, including meteorology, terrestrial ecosystems, marine biology, resource management, and environmental sciences, and the storage of remote sensing databases at each satellite data centre has reached the PB level. Such a large number of multisource remote sensing data has caused great difficulties in the management of data queries, which causes the inability of current data organization storage modes and data management systems to meet the application requirements (Ye et al. 2017; Zhou et al. 2016).

CONTACT Guoqing Li ligq@radi.ac.cn Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China

GeoHash, which is a spatial lattice coding algorithm, can transform two-dimensional longitude and latitude coordinates into a simple, sortable and comparable string, which effectively reduces the spatial attribute parameter dimensions and facilitates the HBase primary key index. Therefore, this paper intends to use the GeoHash algorithm to encode and convert latitude and longitude coordinates from remote sensing metadata to construct an HBase primary key model that contains as many search conditions as possible to achieve and realize the rapid retrieval of multisource remote sensing metadata (Fan et al. 2016; Zhou et al. 2017).

The structure of this article is as follows: section 2 describes background knowledge and related works; section 3 describes the GeoHash spatio-temporal organization model used for remote sensing metadata encoding; section 4 describes the experimental results of the massive remote sensing metadata retrieval based on the GeoHash code; and the conclusion is in section 5.

2. Background knowledge

2.1. *HBase introduction*

HBase was created in 2007 and was initially part of Hadoop. HBase stores data as tables. The tables consist of rows and columns, each of which belongs to a specified column family. A data unit consisting of rows and determined columns in a table becomes an element (i.e., cell), where each element holds multiple versions of the same data, which are represented by timestamps [Ma et al. (2015); Song, Zhu, and Li (2014)].

A row key is a unique identifier that determines the table record and primary key for retrieving records. The line key can be any string (with a maximum length of 64 KB). Within HBase, row keys are sorted in the form of a byte stream array and in byte order. There are three ways to access the rows in an HBase table: (1) through a single row key; (2) through a range of row keys; and (3) a full table scan (Ma et al. 2016).

A column family is a collection of logical relational columns in an HBase table. Each column in the HBase table belongs to a column family. The column family must be defined before the table is used, which is part of the table scheme, and the columns are dynamically generated during use. The data in HBase are sorted using a dictionary based on the order of row keys. When the line keys are the same, the data are sorted using a dictionary based on the order of column names (Ji et al. 2014).

2.2. *GeoHash introduction*

The GeoHash algorithm is a geographic data coding technique based on a grid partitioning proposed by Gustavo Niemeyer. By converting the target latitude and longitude coordinates into an encoded string similar to that of a URL, the algorithm can reduce the redundancy generated by the traditional method to guarantee a precise target position. Therefore, it is widely used in terrain perimeter queries and the geo-fencing technologies because it provides more accurate peripheral user recommendations.

The main idea behind the GeoHash algorithm is that the latitude and longitude coordinates (0, 0) are the original points, and the Y and X axes, which represent latitude and longitude, respectively, treat the ranges for longitude and latitude as a two-dimensional

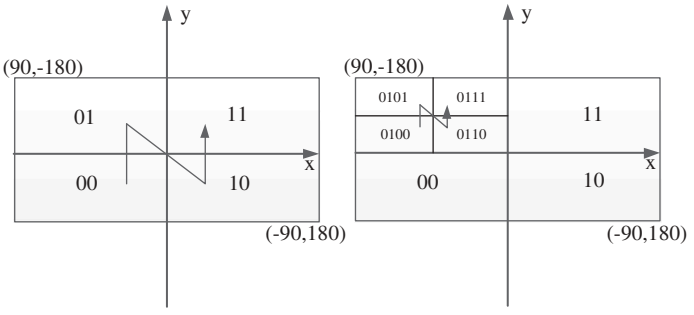


Figure 1. GeoHash grid division diagram.

rectangle in the coordinate system. The rectangle is divided into two points along the longitude and latitude axes, subsequently. The smaller latitude or longitude intervals are encoded as 0, while those that are bigger are encoded as 1, as shown in Figure 1(Liu et al. 2014; Jeansoulin 2016).

Assuming that x represents latitudinal values, the interval [−90, 90] is divided into two categories, [−90, 0] and [0, 90], which are called the left and right intervals, respectively. The interval [−90, 0] is smaller, and the interval [0, 90] is bigger. Suppose that x belongs to the right interval, [0, 90], recursively in the above procedure; regardless of the first iteration, x always belongs to interval [a, b]. With each iteration, interval [a, b] continuously narrows. It is known from the limit that anytime [a, b] converges to x (i.e., any given ε), there is always an N, such that $\delta = | x\text{-}a/2\ N | < \varepsilon$, where x is any given latitude. The above analysis process guarantees the convergence of the algorithm, and the length of the sequence is related to the given convergence multiplied by N. Based on the recursion, after a binary sequence of N bits is combined with 0 and 1, each of the 5 bit binary numbers are compiled into a group, and the group is converted into a number between 0 and 9 or lowercase letters b to z (excluding a, l, l, and o) in English for a 32-bit character to obtain 32 bit GeoHash grid encoding, as shown in Figure 2.

When the length of the 32-bit GeoHash encoding is 8, the precision of the corresponding space is approximately 19 meters; when the length of the encoding is 9, the precision is approximately 2 meters. Therefore, during the actual spatial location coding process for remote sensing images, the specific GeoHash encoding length should be chosen according to its spatial resolution size.

3. Spatio-temporal organization model for remote sensing metadata based on GeoHash coding

Remote sensing metadata stores the descriptive information of the remote sensing image data; it provides information regarding the characteristics of remote sensing image data, such as identification, imaging time, imaging location, product grade, quality and spatial reference system (Henri et al. 2017). Currently, the standards for commonly used satellite metadata are different, such as the NASA-EOSDIS metadata format, the hierarchical data format (HDF)-EOS (Wei et al. 2007), and the custom XML metadata format from the China Centre for Resources Satellite Data and Application. Metadata of different forms used for

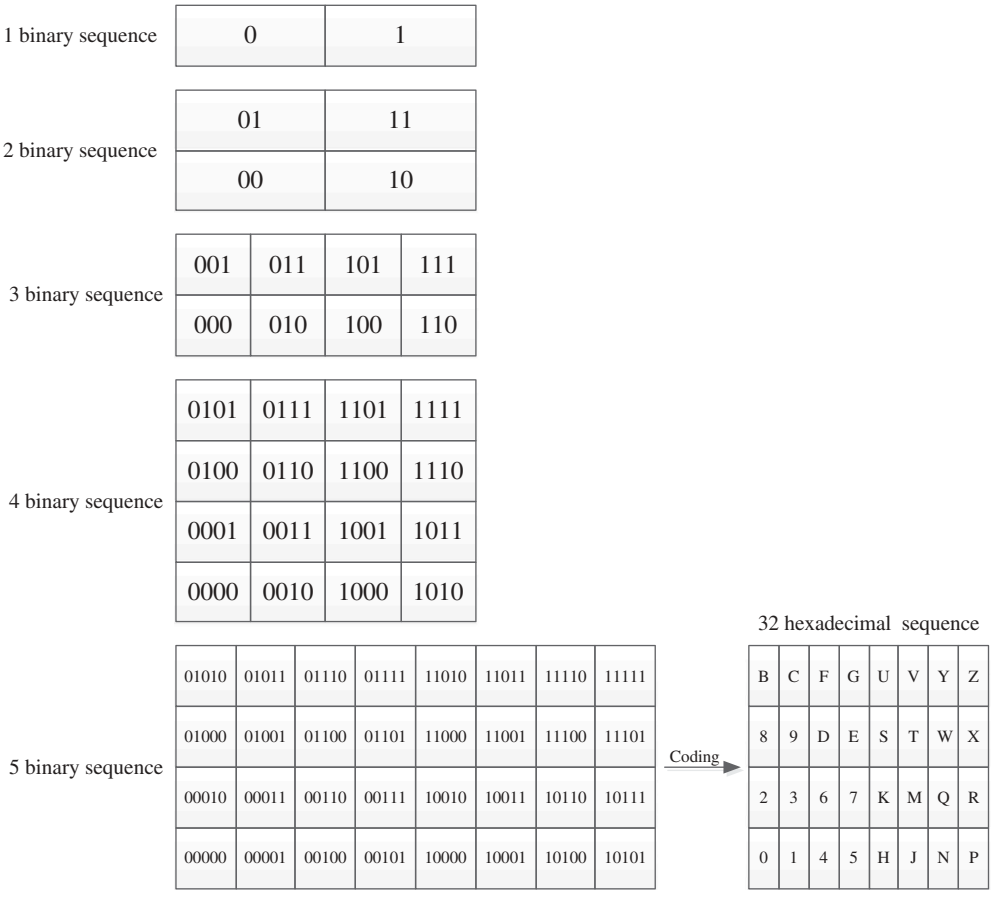


Figure 2. 32 bit GeoHash grid encoding.

remote sensing satellites pose great difficulties for the integrated management of multi-source remote sensing data; therefore, it is necessary to develop a widely accepted standard metadata format (Chen and Hu 2012; Devarakonda et al. 2010).

3.1. *Remote sensing metadata format based on the ISO standard*

The ISO 19115–2 is the second part of the ISO 19115 standard and is an extension of the image and the grid data, which provides information on the acquisition of spaceborne remote sensing data, the description of remote sensing data bands and other information. The ISO 19115–2:2009 has been integrated into the general metadata warehouse CMR (common metadata repository) (Yue, Gong, and Di 2010; Gilman and Shum 2016), and it became a standard for data exchange and the integration and retrieval between international geographic information organizations and geo-data centres.

This study is based on ISO 19115–2:2009 geo-information metadata standard and aims to determine the characteristics of remote sensing data and establish a unified standard format for remote sensing metadata (Table 1). Remote sensing metadata for each distributed satellite data centre need to be converted into a standard format before data integration.

Table 1. The ISO 19,115–2:2009-based uniform metadata format.

Classification	Metadata fields	Descriptive information
Metadata information	Creation	metadata creation time
	Lastrevision	last modified
Image information	MD_identifier	image name
	Timeperiod_beginposition	data start time
	Timeperiod_endposition	data end time
	MI_platform	satellite name
	MI_instrument	device name
	MI_sensor	sensor
	Datacenter	data center
	recStationid	satellite receiving station
	spatialResolution	spatial resolution
	Westboundlongtude	longitude of the west boundary
	Eastboundlongtudei	longitude of the east boundary
	Southboundlatitude	longitude of the south boundary
	Northboundlatitude	longitude of the north boundary
	Centerlongtude	center point longitude
	Centerlatitude	center point latitude
	Scenepath	path
	Scenerowi	row
	Formatspecificationcitationi	image format
	Referencesysitemidentifier	reference coordinate system
	Cloudcoverpercentagei	cloud cover volume
	Imagequalitycode	image quality
	Processinglevel	processing level

3.2. *HBase key value model of the remote sensing metadata based on GeoHash coding*

After the multisource remote sensing metadata format has been made uniform, the metadata can be imported into the HBase database for the management and retrieval of mass remote sensing metadata. HBase is a column family database, and rapid retrieval efficiency is primarily implemented by the row key; therefore, a well-designed key-value model is required. However, remote sensing metadata are typically spatial data, which means that the joint retrieval based on 'satellite + sensor + time + space' is unavoidable. Therefore, it is important to realize the rapid retrieval of HBase to reduce remote sensing data retrieval dimensions by unifying the coding of satellites, sensors, time and space, which would allow the metadata to be reduced to a single key retrieval value whenever possible.

(1) Coding of satellites and sensors

Regarding the characteristics of multisource remote sensing metadata, this research codes the names of satellites and sensors with binary codes to replace strings, and the codes are used as satellite and sensor attributes in the HBase key value. The codes of the satellites and sensors are shown in Table 2.

(2) Temporal coding

For the collection of time data, uniform encoding requires an 8-bit decimal value; this is the time attribute of the HBase key value code.

Table 2. Satellite and sensor codes.

Satellite	Sensor	Coding
Landsat 5	TM	00000000
Landsat 7	ETM	00010000
Landsat 8	OLI	00100000
HJ1A	HSI	00110001
HJ1B	CCD	01000000

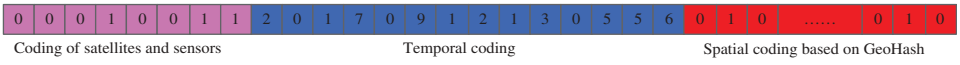


Figure 3. HBase key value model.

(3) Spatial coding based on GeoHash

Using a spatial grid coding algorithm, GeoHash can transform a two-dimensional latitude and longitude coordinate into a simple string that can be sorted and compared, which effectively reduces the dimensions of the spatial attribute parameters and facilitates the primary key index of HBase. Therefore, for remote sensing images with four longitude and latitude coordinates, in order to effectively reduce the dimension of the spatial parameter index, this study uses remote sensing images centred about a longitude and latitude to carry out the GeoHash spatial encoding, which is used as the spatial attribute in the coding of the HBase key value.

The main idea for using GeoHash for remote sensing image encoding is a) the level of GeoHash space encoding (i.e., the length of the GeoHash code), which is determined according to the width of the satellite sensor; b) the longitude and latitude coordinates of the central point, which are calculated according to the latitude and longitude coordinates from the four corners of the remote sensing image; c) the GeoHash coding is calculated according to the GeoHash spatial coding level and the latitude and longitude coordinates of the centre point.

The final HBase key value model is shown in Figure 3.

4. Experiment and results

4.1. *Test data*

The experiment is performed to test the performance of remote sensing metadata management based on GeoHash coding; the tested dataset results in approximately 15 million remote sensing metadata generated by simulations of the Landsat 8 OLI_TIRS, Landsat 7 ETM+, Landsat 5 TM, Landsat 1–5 MSS, Aster L1T, CEBERS-1/2 CCD, HJ-1A/B CCD, HJ-1A HSI, and FY-3A/B VIRR metadata. These remote sensing metadata are all based on GeoHash coding to generate HBase key values and are imported into the HBase database system.

4.2. *Remote sensing metadata retrieval experiment*

The test experiment for the metadata retrieval is divided into six types according to the type and volume of the simulated metadata. The volume of each type of experimental metadata is 1 million, 2 million, 5 million, 8 million, 10 million, and 15 million, respectively, and all of the simulated metadata contained the same volume as that of the Landsat 5 TM metadata. For each type of data retrieval test experiment, the satellite sensor is set to Landsat 5 TM, and the retrieval time range is set to 1 day, 1 month, and 1 year. The spatial range is 115.41°E – 117.50°E and 39.43°N – 41.05° N, and the primary key is retrieved according to HBase key value. To exclude the influence of accidental factors, such as network delay, each test experiment is executed 20 times by taking the average value as the return time of the query result (Figure 4).

In addition, in order to illustrate the advantages of using the HBase value model based on GeoHash encoding for each type of experiment, traditional satellites, sensors, time and space conditions were added and combined with multiple retrieval modes and multi-column filters based on HBase for the metadata retrieval experiment. For simplicity, the HBase value model based on GeoHash encoding method is simply referred to as HBaseGeoHash, and the longitude and latitude retrieval method is referred to as HBaseLatLon

The final results are shown in Figure 4.

As can be seen from Figure 4, as a whole, an increase in the volume of remote sensing metadata indicates a linear increase in the retrieval time. Specifically, (1) when the retrieval time remains unchanged with an increase in the amount of remote sensing metadata, the increase in HBaseGeoHash retrieval time is much smaller than that of the HBaseLatLon retrieval time, especially when the data amount is greater than 8 million and the retrieval efficiency of HBaseGeoHash is superior to that of HBaseLatLon; and (2) when the total amount of metadata is constant with an increase in retrieval time, the retrieval time for HBaseGeoHash did not obviously increase, and the retrieval time of HBaseLatLon significantly increased. This may be because the multi-column filtering query mechanism of HBase needs to scan the entire table. On the other hand, the

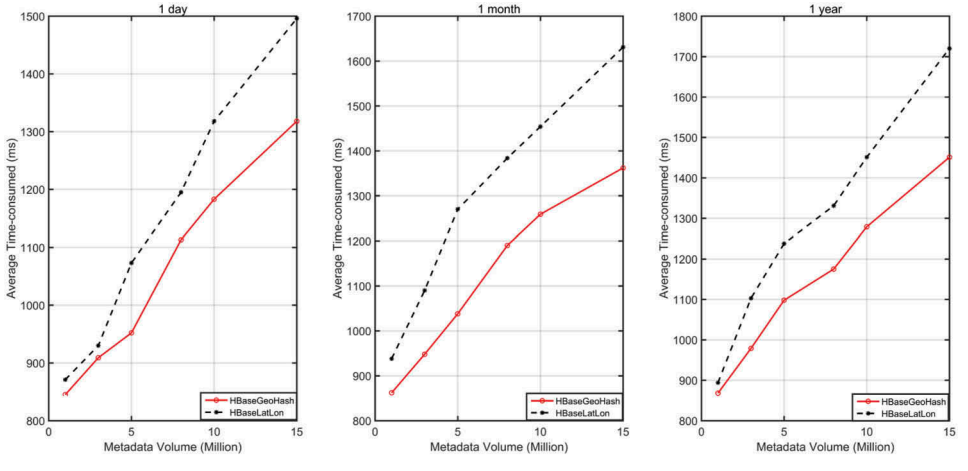


Figure 4. Experimental results of the remote sensing metadata retrieval.

HBase key retrieval based on GeoHash encoding did not need to scan and filter the whole table based on this condition; it only needed to query based on the primary key, which greatly improved the retrieval efficiency. This strongly demonstrates the superiority of the HBase key-value retrieval method based on GeoHash encoding.

5. Conclusions

In view of the current problem regarding the slow retrieval efficiency of massive remote sensing metadata, this paper proposes a fast retrieval strategy for massive remote sensing metadata based on GeoHash coding. First, to solve the problem caused by the differences in multisource remote sensing metadata formats, the unified format conversion of the remote sensing metadata is realized based on the ISO standard. Second, in order to reduce the dimension of the multi-condition joint retrieval of the remote sensing metadata, a remote sensing metadata key value model based on GeoHash encoding is constructed, which is derived from multiple elements, such as satellites, sensors, time and space, and reduced to an HBase single bond retrieval value. Finally, by comparing the 15 million remote sensing metadata retrieval experiments with the traditional multi-conditional joint retrieval experiments, this study finds that the spatio-temporal organization strategy for the remote sensing metadata based on GeoHash encoding can effectively improve the efficiency of massive remote sensing data retrieval.

However, the GeoHash coding retrieval used in this paper is directly aimed at the entire remote sensing image. In the process of retrieval, the purpose of the retrieval process is to handle remote sensing images as a single point. Because the original remote sensing image has a wide width, GeoHash encoding of the centre point is used to identify the whole image. Although this improves the efficiency of data retrieval, it also reduces the accuracy of the data query to a certain extent. Therefore, it is the focus of future work to establish a dissection face based on the segmentation of the remote sensing image using DataCube and smaller tile width to retrieve the data.

Funding

This work was supported by the National Key Research and Development Program of China [2016YFB0501503].

References

Chen, N. C., and C. L. Hu. 2012. "A Sharable and Interoperable Meta-Model for Atmospheric Satellite Sensors and Observations." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (5): 1519–1530. doi:10.1109/JSTARS.2012.2198616.

Devarakonda, R., G. Palanisamy, B. E. Wilson, and J. M. Green. 2010. "Mercury: Reusable Metadata Management, Data Discovery and Access System." *Earth SCI Inform* 3 (1–2): 87–94. doi:10.1007/s12145-010-0050-7.

Fan, Z., W. Zhang, D. Zhang, and L. Meng. 2016. "An Automatic Accurate High-Resolution Satellite Image Retrieval Method." *Remote Sensing* 9 (11): 1092–2011. doi:10.3390/rs9111092.

Gilman, J. A., and D. Shum. 2016. "Making Metadata Better with CMR and MMT." *Teaching material presented at Summer ESIP*, Durham, NC, July 19–22.

He, C., Z. Zhang, D. Xiong, J. Du, and M. Liao. 2017. "Spatio-Temporal Series Remote Sensing Image Prediction Based on Multi-Dictionary Bayesian Fusion." *ISPRS International Journal of Geo-Information* 6 (12): 374. doi:10.3390/ijgi6110374.

Henri, L., D. Vincent, R. P. Hugo Antonio, M. Serge, and C. Jean-François. 2017. "Landsat-8 Cloud-Free Observations in Wet Tropical Areas: A Case Study in South East Asia." *Remote Sensing Letters* 8 (7): 537–546. doi:10.1080/2150704X.2017.1297543.

Jeansoulin, R. 2016. "Review of Forty Years of Technological Changes in Geomatics toward the Big Data Paradigm." *ISPRS International Journal of Geo-Information* 5 (9): 155. doi:10.3390/ijgi5090155.

Ji, Z., I. Ganchev, M. O'Droma, L. Zhao, and X. Zhang. 2014. "A Cloud-Based Car Parking Middleware for IoT-Based Smart Cities: Design and Implementation." *Sensors* 14 (12): 22372–22393. doi:10.3390/s141222372.

Liu, J., H. Li, Y. Gao, H. Yu, and D. Jiang. 2014. "A GeoHash-based Index for Spatial Data Management in Distributed Memory." *Paper presented at International Conference on Geoinformatics*, Taiwan, June 25–27. doi: 10.1109/GEOINFORMATICS.2014.6950819.

Ma, T., X. Xu, M. Tang, Y. Jin, and W. Shen. 2016. "MHBase: A Distributed Real-Time Query Scheme for Meteorological Data Based on HBase." *Future Internet* 8 (1): 6. doi:10.3390/fi8010006.

Ma, Y., H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie. 2015. "Remote Sensing Big Data Computing: Challenges and Opportunities." *Future Generation Computer Systems* 51: 47–60. doi:10.1016/j.future.2014.10.029.

Song, Y., Y. Zhu, and L. Li. 2014. "Large Scale Data Storage and Processing of Insulator Leakage Current Using HBase and Map-Reduce". *Paper presented at International Conference on Power System Technology*, Chengdu, China, October 20–22. doi:10.1109/POWERCON.2014.6993650.

Wei, Y. X., L. P. Di, B. H. Zhao, G. X. Liao, and A. J. Chen. 2007. "Transformation of HDF-EOS Metadata from the ECS Model to ISO 19115-Based Xml." *Computers & Geosciences* 33 (2): 238–247. doi:10.1016/j.cageo.2006.06.006.

Yan, J. N., Y. Ma, L. Wang, K. R. Choo, and W. Jie. 2017. "A Cloud-Based Remote Sensing Data Production System." *Future Generation Computer Systems* 2017. doi:10.1016/j.future.2017.02.044.

Ye, D., Y. Li, C. Tao, X. Xie, and X. Wang. 2017. "Multiple Feature Hashing Learning for Large-Scale Remote Sensing Image Retrieval." *International Journal of Geo-Information* 6 (11): 364. doi:10.3390/ijgi6110364.

Yue, P., J. Y. Gong, and L. P. Di. 2010. "Augmenting Geospatial Data Provenance through Metadata Tracking in Geospatial Service Chaining." *Computers & Geosciences* 36 (3): 270–281. doi:10.1016/j.cageo.2009.09.002.

Zhou, Y., S. De, W. Wang, K. Moessner, and M. S. Palaniswami. 2017. "Spatial Indexing for Data Searching in Mobile Sensing Environments." *Sensors* 17 (6): 1427. doi:10.3390/s17061427.

Zhou, Y., C. Liu, N. Li, and M. Z. Li. 2016. "A Novel Locality-Sensitive Hashing Algorithm for Similarity Searches on Large-Scale Hyperspectral Data." *Remote Sensing Letters* 7 (10): 965–974. doi:10.1080/2150704X.2016.1207255.