DETECTION OF BEHAVIOR PATTERNS OF INTEREST USING BIG DATA WHICH HAVE SPATIAL AND TEMPORAL ATTRIBUTES

R. W. La Valley[1*], A.Usher[2], A. Cook[3]

[1]OGSystems, Inc., Data Scientist and Senior Statistician, 14291 Park Meadow Dr # 100, Chantilly, VA 20151, Richard.LaValley@ogsystems.com
[2] Digital Globe Intelligence Solutions, Chief Technical Officer, 4350 Fairfax Dr., Ste 950, Arlington, VA 22203 Abe.Usher@digitalglobe.com
[3] Digital Globe Intelligence Solutions, 4350 Fairfax Dr. Ste. 950, Arlington, VA 22203, Adam.Cook@digitalglobe.com

KEYWORDS: Geospatial, Temporal, Aggregation, Location, Z-Curve, Space-Time Boxes, Geo-Temporal Hashing, Big Data

ABSTRACT:

New innovative analytical techniques are emerging to extract patterns in Big Data which have temporal and geospatial attributes. These techniques are required to find patterns of interest in challenging circumstances when geospatial datasets have millions or billions of records and imprecision exists around the exact latitude and longitude of the data. Furthermore, the usual temporal vector approach of years, months, days, hours, minutes and seconds often are computationally expensive and in many cases do not allow the user control of precision necessary to find patterns of interest.

Geohashing is a single variable ASCII string representation of two-dimensional geometric coordinates. Time hashing is a similar ASCII representation which combines the temporal aspects of date and time of the data into a one dimensional set of data attributes. Both methods utilize Z-order curves which map multidimensional data into single dimensions while preserving locality of the data records. This paper explores the use of a combination of both geohashing and time hashing that is known as "geo-temporal" hashing or "space-time" boxes. This technique provides a foundation for reducing the data into bins that can yield new methods for pattern discovery and detection in Big Data.

1. THE PROBLEM

Recent developments in geospatial analysis allow for the discovery of patterns of behaviors of entities of interest (Phithakkitnukoon, Husna, & Dantu, 2008) and (Ponienan, Salles, & Sarraute, 2013). Several issues present challenges for these approaches such as the Modifiable Areal Unit Problem (MAUP) which shows that the outcome of assessments can differ when data is aggregated to differing levels (provincial level, municipal level aggregated to provincial, neighborhood levels aggregated to municipal to provincial) (Openshaw, 1983). Tools used in geospatial analyses often have limited capacity to support rigorous quantitative assessments of geo-coordinates, and apply visualizations to provide exploratory data analysis and subjective discovery of patterns in the data (Lambiott, et al., 2008).

Precision and accuracy limitations with contemporary GPS devices such as smartphones, tablets, cameras and other handheld personal devices create noise which causes challenges in determining meaningful patterns in a combination of data from different devices (Eagle, Pentland. & Lazer, 2009). Perhaps the biggest challenge in analysis of spatio-temporal data is the sequence of observations and the validity of assessments to discovering patterns. Valid approaches to pattern detection and analysis must handle the temporal aspects of data to determine patterns such as co-location, loitering, meetings, et cetera. And finally, given the scale of big data datasets, analytical methods must be computationally efficient to facilitate rapid analysis.

2. INDEXING OF SPATIOTEMPORAL DATA

2.1 Spatial Data Indexing Using Geohashing

Multiple methods for the handling and storage of spatial data have been developed and emerged. These include R-tree (Guttman, 1984) and (Beckmann, Kriegel, Schneider, & Seeger 1990), Hilbert R-tree (Kamel, & Faloutos 1993), Quadtree (Bentley, 1975), and Geohashing (Niemeyer, 2008) using Z-order curve. Geohashing is a data encoding technique to produce spatial bounding boxes used for binning of nearby points based on Z-order curves (Morton, 1966). Unlike many of the other spatial binning techniques, the geohash algorithm allows for the user to define the precision determination of the level of precision for the geospatial data hashes to use for the discovery of an event of interest. Similarly the time hash provides a method for summarizing intervals of time at a level of precision defined by the user. The method described in this paper, geo-temporal hashing illustrates a technique which allows for the selection of a range of precision for both geo and temporal hashes. Geohashes represented as an ASCII string composed of 32 alphanumeric characters providing a visual representation of the granularity and the reduction of spatial coordinates into a single attribute that is easily used in the discovery of patterns. It is important to note that multiple precisions are obtained without processing the data more than once. Geohashes at a fixed level of precision (e.g. geohash size 14, which represents a very small rectangle within a Cartesian coordinate system) are produced from a single step allowing for the user to select the precision desired (geohash size 14, or size 13, or any of the hierarchically related geohash boxes). This technique has many advantages including

* Corresponding Author

computational efficiency which enables the processing of billions of data.

Figure 1 illustrates how applying geohash can be thought of as a hierarchical decomposition of boxes of various sizes that can be tuned to a particular level of geospatial precision ranging from hundreds of kilometers to less than a meter.

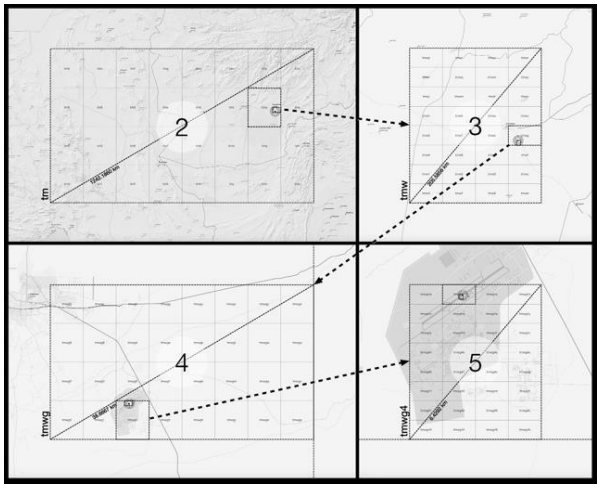

Figure 1. GeoHash – tmwg4 and each of the four images illustrate the five levels of precision

2.2 Temporal Data Indexing Using Time Hashing

Time hashing is a lossey precision technique for temporal encoding (Usher, 2010). This technique represents intervals of time and produces units of time into a single ASCII string for easy query and discovery of patterns. Like geohashing, this technique is hierarchical decomposition method which produces multiple levels of amalgamation of time in the ASCII string which is computationally efficient. It can be implemented by selecting a fixed period of 128 years such as 1970-01-01 00:00:00 GMT to 2097-12-31 23:59:59 GMT to take advantage of the base-2 encoding that allows for efficient subdivision of intervals into smaller time periods. The calculation of the time hashes of 14 levels of precision can produce the temporal bins as large as 64 years to nanoseconds. Each ASCII in the string represents bits of time (e.g., +/- 16 years, +/- 8 years, 91.2 days, +/11.4 days, etc.).

Figure 2 illustrates the various levels of precision which can be achieved in a single pass of data. This temporal index can

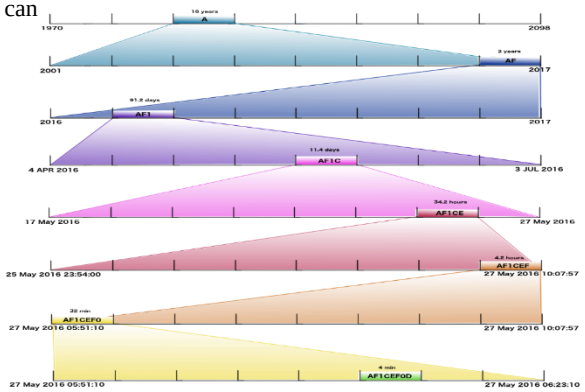

Figure 2. Time hashes with eight levels of precision

be used by a user to determine the level of precision.

2.3 Space-Time Boxes

The discovery of patterns in the data requires the creation a primitive for processing with both spatial and temporal components. The concatenation of geohashes and time hashes yields a usable primitive once the user determines the desired level of precision for both space and time. Unlike other methods, this methodology initially processes the data into s high level of precision for both space and time hashes. These hashes are concatenated into a single value to represent co-occurring three-dimensional bins of latitude, longitude and time (La Valley, Usher, & Halman 2015).

This three-dimensional variable may be thought of as a space-time box. This technique enables the possibility of analytical methods of discovery and exploration of space-time patterns and the discovery of previously unknown relationships through the use of simple visualizations such as heat maps. Table 1 illustrates the granularity and flexibility which is at the disposal of the user through the techniques described.

Figure 3. provides a 3D visual of this concatenation creating geo-temporal hashing.

Precision	Geohash (at equator)	Timehash	Precision	SpaceTimehash (at equator)
1	≈ 5009*km* x 4993*km*	16years	2	≈ 5009*km* x 4993*km* x 16years
2	≈ 1252*km* x 624*km*	2years	4	≈ 1252*km* x 624*km* x 2years
3	≈ 157*km* x 156*km*	91.2days	6	≈ 157*km* x 156*km* x 91.2days
4	≈ 39.1*km* x 19.5*km*	11.4days	8	≈ 39.1*km* x 19.5*km* x 11.4days
5	≈ 4.89*km* x 4.89*km*	34.2hours	10	≈ 4.89*km* x 4.89*km* x 34.2hours
6	≈ 1.22*km* x 0.61*km*	4.2hours	12	≈ 1.22*km* x 0.61*km* x 4.2hours
7	≈ 152.9*m* x 152.4*m*	32mins	14	≈ 152.9*m* x 152.4*m* x 32mins
8	≈ 38.2*m* x 19.1*m*	4mins	16	≈ 38.2*m* x 19.1*m* x 4mins
9	≈ 4.77*m* x 4.77*m*	30secs	18	≈ 4.77*m* x 4.77*m* x 30secs
10	≈ 1.19*m* x 0.59*m*	3.6secs	20	≈ 1.19*m* x 0.59*m* x 3.6secs
11	≈ 14.9*cm* x 14.9*cm*	–		
12	≈ 3.72*cm* x 1.86*cm*	–		

Table 1. Precision achieved using Geo-temporal hashes (Space Time Boxes)

2.4 Data Locality of Geohashes and Temporal Hashes

One of the key features of space-filling curves is that they preserve data locality by ensuring that if two spatial or temporal curves are close, their corresponding hashes will be

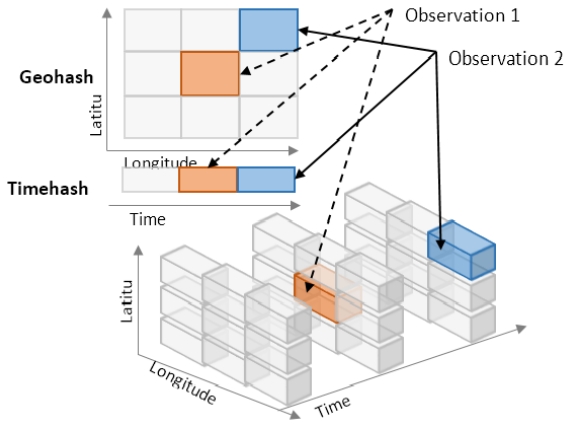

Figure 3. Concatenation of Geohashing and Temporal Hashing of two observations

numerically close. The closeness of the locality can be exploited when looking for certain patterns of interest which require either a spatial or temporal closeness to be of interest to the user.

2.5 Multiple precision levels without additional processing

A key advantage of Z-order curves to over space filling curves is the computation of a high precision as described in sections 2.2 and 2.3 with a single set of algorithmic calculations. Using simple truncation of the least significant bits of the ASCII representation to achieve lower precision levels. The ease of having multiple levels of precisions easily available without recalculating for each precision provides an ideal environment for the discovery of patterns of interest.

3. SIMPLE OPERATIONALLY SIGNIFICANT PATTERNS OF INTEREST

3.1 Spatiotemporal Patterns in Maritime AIS Data

The utility of these primitives (geohash, timehash & geo-temporal hashes) outlined is illustrated by providing several patterns of operational significance and interest to users using maritime data from maritime AIS (Automatic Identification System) data. Table 2 shows the variables which are broadcast in AIS every fifteen minutes by each system.

Name of Variable	Description
MMSI	Maritime Mobile Service Identity. Unique identifier for the ship
Name	The Ship Name
Latitude	Decimal Degree
Longitude	Decimal Degree
Speed	KPH – Kilometers per hour
Heading	Compass direction of the ship is traveling at time of transmission
Transmission Time	Time of AIS broadcast
Destination	Description or port name of the destination of the ship (optional, manually entered by ship crew
ETA	Estimated time of arrival (Optional, manually entered by ship crew

Table 2. Variables in maritime Automatic Identification System (AIS) transmissions

The following patterns illustrate the utility of geo-temporal hashing (space time boxes) for auto- matically discovering patterns of interest or events of interest in the maritime data in AIS. These patterns will be described in each section and provide a brief description of their utility to address analytical problems of interest.

3.1 Co-Location of Multiple Entities

When two or more entities occupy the same geo-temporal hash or space time box, there is a potential for it being an event of interest to users. The user will need to set the precision required for the event to be significant to be an event of interest as illustrated in Figure 4.

However, the co-location in a single space time box does not always mean that it is an event of interest. For example, if the geohashes occur in known commercial shipping lanes, then the event could be a simple passing of two or more vessels during the time hash. To refine the automatic discovery of events of interest, the user should look for loitering of two or more vessels over several space time boxes or possibly adjust the precision of the time hashes to refine the discovery of the event.

Another example of a co-location in a single space time box is when the multiple vessels share the same geohash in a known port. This discovery may or may not be of interest to the user.

Automatic discovery of certain events of interest such as illegal off-loading of cargo may require filtering out of events which occur in geohashes which are near to known commercial shipping lanes or near known shipping ports.

3.2 Loitering of Vessels

Another event which can be discovered using geo-temporal hashing is loitering. Loitering can be defined when the user has the desired precision for both geo and temporal hashes, and a vessel is found to occupy multiple consecutive or near

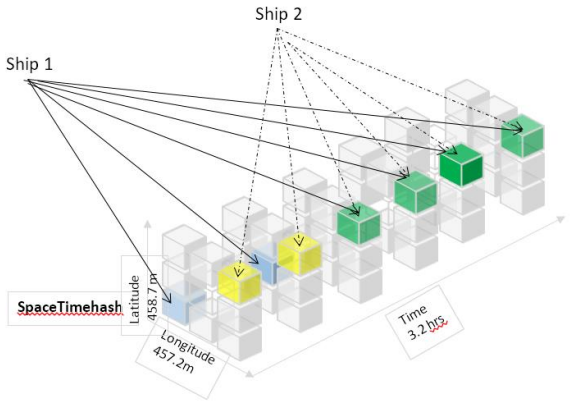

Figure 4. Illustration of the Co-Location of more than one vessel loitering for more than 3 or more time hashes and same geo-hash

consecutive time hashes. This loitering event may or may not of interest to the user and can occur when a vessel or ship is at a known port offloading its cargo. If the vessel is discovered to be loitering and isn't near a known port facility, it may be an event of interest to the user. For example, if the vessel or ship is in the open sea, then it could be broken down.

It could also be stopped and waiting for another vessel. Either may be an event worth investigating. Coastal areas and known shipping lanes around the world be used and mapped to a geohash. The user uses geohashes to identify a vessel or ship which is demonstrating a loitering activity as defined above and is near a coastline geo-hash with no known port facility. This pattern could be an event of interest as it could be a ship is utilizing an area which is not normally designated a shipping port for illicit purposes such as possible offloading of cargo for smuggling. If the loitering is continuous for consecutive or near consecutive time hashes, then the event is likely to be either a broken down vessel or another event of possible interest to the user.

3.3 Co-Location of Multiple Entities Loitering in Unusual Places

There are several potential events of interest when there is co-location of multiple vessels at a geo-hash that is not near a known shipping lane or not near a known port. If multiple vessels are discovered to be in the same time hash or several

consecutive time hashes, then that event is possible an event of interest to the user.

This pattern is possibly the off-loading of cargo or loading of an illicit cargo of interest between vessels. If more than one vessels are found to be at the same geohash but at different times, then this could be an event of interest to the user. This event could be a "drop" of illicit cargo which is picked up by the second or subsequent vessels. If two or more vessels occupy the same space time does not include a known port, then the loitering event is possibly an event of interest to the user.

3.4 Co-Traveling

When more than one vessel occupies different geohashes across more than three or more time hashes, then the vessels could be considered to be "co-traveling" and is an event of potential interest to the user as illustrated in Figure 5.

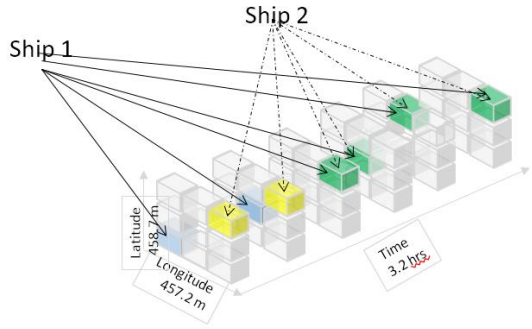

Figure 5. Co-traveling of multiple vessels for more than 3 or more time hashes and different geohashes

4. IMPLEMENTATION AND METRICS

Variations of this method have been developed for use as a forensic tool of exploring spatiotemporal data. The enrichment of spatiotemporal data with both geo and temporal hashes has been measured for the speed of computation with a Python programming implementation of the Niemeyer GeoHash and the Usher Temporal Hash routines found in GitHub. Random samples of AIS data of multiple sample sizes were extracted and the processing time was measured for each of 10 trials. Average Processing was used as a metric for comparison and a simple pattern of interest (co-location in same Space-Time Box) was used for comparison.

Desktop trials were performed on a MacBook Pro with 2.2GHz Core i7 processor and 16GB RAM using a Python implementation.

An Amazon Web Service (AWS) cloud implementation which used 10 SPARK EMR workers was performed for comparison.

4.1 Desktop Implementation and Metrics for Enrichment and Co-Location Pattern of Interest

Table 3 provides the results and metrics for multiple trials of various sample sizes which were used to explore the computational times for performing the full enrichment with precision of Geo13 and Time10 and finding the co-location pattern of interest for a specific level of precision Geo6 and Time7.

Implementation Results	Sample Size of AIS events						
Enrichment at Geo10Time12 Precision	10	100	1,000	10,000	100,000	1,000,000	10,000,000
Average time (seconds)	0.231	0.278	0.321	1.069	8.835	84.82	744.66
Average Unit time (milliseconds)	231	278	321	1069	8835	84820	744660
Enrichment at Geo6Time7 Precision and Finding Co-Location							
Average time (seconds)	0.314	0.316	0.373	0.968	6.854	67.77	658.82
Average Unit time (milliseconds)	314	316	373	968	6854	67770	658820

Table 3. Table of implementation results for desktop

The results on the desktop used in the creation of the geospatial and temporal hashes in Python are close to linear if it is plotted on a log scale. The table shows that for the co-location in a single space-time box with the desired precision in the Python implementation used was also a linear growth in a log scale.

4.2 Horizontal Scalable Cloud-Based Implementation

The same set of AIS sample sets used in the desktop implementation were reused for the AWS implantation using Spark, and the results were measured. Figure 6 illustrates the processing time per record comparing an eight core laptop with the AWS implementation.

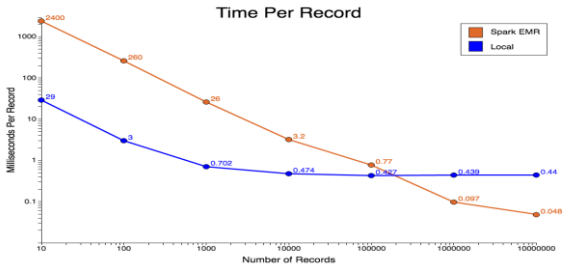

Figure 6. Processing results in AWS Implementation

As is illustrated, for relatively small data sets of less than 100,000 observations, the desktop outperforms the AWS implementation. This result was investigated and found to be due to the management overhead time of starting the Spark workers. But for datasets beyond 100,000 observations, the AWS implementation outperforms the desktop implementation.

The AWS implementation was tested on even larger datasets of 10 billion observations and the findings shown above were consistent.

5. LIMITATIONS

Geohashes and Temporal hashes using Z space-filling curves have limitations. Niemeyer's require additional information to identify neighboring regions using the binary tree or Z-order curve technique. This will sometimes cause neighboring areas to have different lead strings without an apparent pattern of how to associate the ASCII strings. This problem will happen around the poles, equator and the prime meridian. Also, since Niemeyer's technique is lossy, and the accuracy of the data can decrease with multiple encoding and decoding.

User's temporal hash technique is a lossy technique, but it does not have the boundary mapping issue of geohashing. It does have temporal boundaries of 128 years with a starting point of 1/1/1970 and an ending point of 12/31/2098. This

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-4-W2-31-2017 | © Authors 2017. CC BY 4.0 License.

limits the discovery of patterns of interest to patterns between 1970 through 2098.

For datasets of billions of data, these geohash and temporal hashing techniques limitations need to be balanced with the computation efficiency achieved. For many patterns of interest to the user in Big DATA, the computational efficiency becomes a deciding factor.

6. CONCLUSION

There are many variations of the different patterns or events of interest provided that could be of interest to the user and the space time hash methods provide an efficient way to discover these patterns. The patterns illustrated in this paper were fairly simple and easy to illustrate. The foundation of the space-time box makes the discovery of more complex patterns easier and able to be automated because of the consistency of the boxes. This consistency and the enriching of spatiotemporal data with these methods lends itself to parallel computing and Cloud Based processing.

The authors' current research is the discovery of more complex spatiotemporal patterns than illustrated in this paper using this methodology and varying the precision of the geo and spatial hashes. It is believed that entire families of patterns of interest will be discovered. Also, it is anticipated that new and innovative methods for the automation of the discovery of those patterns will be required and programmed.

Other areas of research by the authors include exploring the possible use of these techniques for entity resolution and identifying possible spatiotemporal relationships in one or more space time boxes. There is an opportunity to leverage these techniques' discretizing continuous multi-dimensional spatiotemporal attributes which should facilitate many advanced entity association techniques based consecutive proximity in space, time or both which leverages one of the real strengths of these hashing techniques.

REFERENCES

Beckmann, N., Kriegel, H. P. Schneider, R., & Seeger, B. (1990). The R—tree: An efficient and robust access method for points and rectangles. ACM SIGMOD Record, 19 (2), p. 322-31.

Bentley, J. I. (1975). Multidimensional binary trees used for associative searching. Communications of the ACM, 18(9) p. 509-17.

Eagle, N., Pentland, A. S. & Lazer, D. (2009). Inferring friendship network structure using mobile phone data. Proceedings of the National Academy of Sciences. 106 p. 15274-8.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. (Vol. 14): ACM.

Kamel, I. & Faloutos, C. (1993). Hilbert R-tree: An improved R-tree using fractals: Proceedings of the Twentieth International Conference on Very Large Databases. Santiago de Chile, Chile.

Lambiotte, R., Blondel, V.D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and Its Applications, 387 (21), 5317-25.

La Valley, R., Usher, A., Halman, A., (2015). Detecting Dark Networks Using Geo-temporal and Pattern-Based Network Analysis Techniques. Illuminating Dark Networks: The Study of Clandestine Groups and Organizations. Cambridge University Press.

Morton, G. M. (1966) A computer oriented database and a new technique in file sequencing: Ottawa, Ontario: International Business Machines Company.

Niemeyer, G. (2008) Geohas.org: Short Links for referencing a position. Retrieved from http://forums.groundspeak.com/GC/index.php?showtopic=186412.

Openshaw, S. (1983). The modifiable areal unit problem. Norwich, U.K.: Geo Books.

Phithakkitnukoon, S., Husna, H., & Dantu, R. (2008) Behavioral entropy of a cellular phone user in H. Liu, J Salerno & M.J. Young (Editors) Social computing, behavioral modeling and predictions (pp160-7) New York: Springer.

Ponienan, N. B., Salles, A. & Sarraute, C. (2013). Human Mobility and Predictability enriched by social phenomena information; Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls, N.Y.

Usher, A. (2010). Temporal algorithms for processing and analyzing large datasets. Sterling Data LLC Report.