An Entropy Maximizing Geohash for Distributed Spatiotemporal Database Indexing

Taylor B. Arnold
AT&T Labs Research
33 Thomas Street
New York, NY 10007
taylor@research.att.com

Abstract

We present a modification of the standard geohash algorithm for which the data volume, rather than spatial area, is constant for a given hash prefix length. This property is particularly useful for indexing large distributed databases, where load distribution of large range scans is an imporant aspect of query performance. Distributed spatiotemporal databases, which typically require interleaving spatial and temporal elements into a single key, reap additional benefits from a balanced geohash by creating a consistent balance between spatial and temporal precision even across areas of varying data density. We apply our algorithm to data generated proportional to population as given by census block population counts provided from the US Census Bureau. An efficent implementation for calculating an arbitrary balanced geohash is also provided.

1 Introduction

Queries over large distributed databases often take the form of a series of large range scans; balancin

2 Entropy Balanced Geohash

2.1 A Formulation of the Standard Geohash Encoding

A geohash is a scheme for mapping two-dimensional coordinates into a hierarchical, one-dimensional encoding. It is explained in several other sources, but we re-construct it here in a format which will be most conducive to generalizations.

The first step is to map latitude and longitude coordinates in a standard unit square; this is done by the following linear mapping:

$$x = \frac{\text{lon} + 180}{360} \tag{1}$$

$$y = \frac{\text{lat} + 90}{180} \tag{2}$$

This choice is by convention, and any other method for mapping coordinantes into the unit square is equally viable. The geohash formulation actually works on any continuous two-dimensional data, and could be used as-is to encode other two-dimensional datasets.

The x and y coordinates need to be expressed in as a binary decimals. Formally, we define the unique $x_i \in \{0, 1\}$ and $y_i \in \{0, 1\}$ such that

$$x = \sum_{i=1}^{\infty} \frac{x_i}{2^i}, \tag{3}$$

$$y = \sum_{i=1}^{\infty} \frac{y_i}{2^i}. \tag{4}$$

A geohash representation of (x, y) is constructed by interleaving these binary digits. The q-bit geohash $g_q(\cdot, \cdot)$ can symbolically be defined as

$$g_q(x, y) := \sum_{i=1}^{\lceil q/2 \rceil} \frac{x_i}{2^{2i-1}} + \sum_{i=1}^{\lfloor q/2 \rfloor} \frac{y_i}{2^{2i}}. \tag{5}$$

It is fairly easy to show that the geohash function is monotone increasing in q, with the growth strictly bounded by 2^{-q}, so that

$$0 \leq g_{q+m}(x, y) - g_q(x, y) < \frac{1}{2^q} \tag{6}$$

For all m greater than zero.

2.2 Entropy

A geohash is typically used as an index in the storing and querying of large spatial processes. A simple theoretical model for a stream of spatial data can be constructed by assuming that each observation is an independent identically distributed random variable $\mathfrak{F}$ from some distribution over space. Borrowing a concept from information theory, we can define the entropy of a geohash over a given spatial distribution by the equation

$$H(g_q) := -1 \sum_{v \in \mathcal{R}(g_q)} \mathbb{P}\left[g_q(\mathfrak{F}) = v\right] \cdot \log_2 \left\{\mathbb{P}\left[g_q(\mathfrak{F}) = v\right]\right\} \tag{7}$$

Where $\mathcal{R}(g_q)$ is the range of the q-bit geohash. It is a standard result that the entropy of a discrete distribution is maximized by the uniform distribution. Therefore we can use this as a proxy for how balanced a geohash is for a given distribution of spatial data.

2.3 The Generalized Geohash

As the q-bit geohash function is bounded and monotonic in q, we can define the infinite precision geohash, which we denote as simply $g(\cdot, \cdot)$, to be the limit

$$\lim_{q \to \infty} g_q(x, y) := g(x, y). \tag{8}$$

With this continuous format, one can see that if we compose g with an appropriate new function h, the composition $h \circ g(x, y)$ which can be thought of as a rescaled version of the traditional geohash. To be precise, we would like a function h to have the following properties:

$$h : [0, 1] \to [0, 1], \tag{9}$$

$$h(0) = 0, \tag{10}$$

$$h(1) = 1, \tag{11}$$

$$x < y \iff h(x) < h(y). \tag{12}$$

Note that Equation 12 implies that h is also continuous. From here, we can define the analogue to a q-bit geohash be truncating the binary representation of w, the value of $h(z)$,

$$h(z) = \sum_{i=1}^{\infty} \frac{w_i}{2^i} \tag{13}$$

To the its first q-bits

$$h_q(z) := \sum_{i=1}^{q} \frac{w_i}{2^i}. \tag{14}$$

In the remainder of this document, we refer the $h_q \circ g(x, y)$ as a generalized geohash.

2.4 The Empirical Entropic Geohash

We have introduced the concept of a generalized geohash in order to construct a spatial encoding scheme which better optimizes the entropy as defined in Equation 7. Assume $\{z_i\}_{i=0}^N$ is a set of independent sample from realizations of the random variable $\mathfrak{F}$. The empirical cumulative distribution function G of the standard geohash function $g(\cdot,\cdot)$ is given by

$$G(t) := \frac{1}{N} \cdot \sum_{i=0}^{N} 1_{g(z_i)\leq t},\ t \in [0,1]. \tag{15}$$

From Equation 15 we can define the balanced entropy maximizing geohash function b (balanced), assuming that every point z_i has a geohash $g(z_i)$ strictly between 0 and 1, to be

$$b^q(t) := \begin{cases} 0 & \text{if } t = 0 \\ 1 & \text{if } t = 1 \\ \frac{N}{N+2} \cdot G^{-1}(t) & \text{if } \exists\, i \in \mathbb{Z} \text{ s.t. } t = i/2^q \\ \text{linear interpolation of the above points} & \text{else} \end{cases} \tag{16}$$

The balanced geohash is essentialy the inverse of the empirical function G, with some minor variations to satisfy Equations 9-12. If the points $\{z_i\}$ are unique, and N is sufficently large, the q-bit analogue b_q of Equation 16 will have an entropy $H(b_q)$ of approximately equal to q.

More formally, we can prove the following bound on the entropy of the balanced geohash:

Theorem 1. *Let b_q^q be the entropy balanced geohash estimated from a sample of N unique data points. Then, with probability at least $1 - 2e^{-0.49 \cdot 2^{-2q} N \cdot [1-A]^2}$ the entropy $H(b_q^q)$ is bounded by the following simultaneously for all values $A \in [0,1]$:*

$$H(b_q^q) \geq q \cdot \frac{N}{N+2} \cdot A \tag{17}$$

Proof. Let $F(\cdot)$ be the true cumulative distribution function of the variable $\mathfrak{F}$, and $F_N(\cdot)$ be the empirical cumulative distribution function from a sample of N independent observations. Setting $\epsilon = [1-A] \cdot \frac{N}{N+2} \cdot 2^{-(q+1)}$, the Dvoretzky-Kiefer-Wolfowitz inequality states that the follow holds for all values $A \in [0,1]$ with probability $1 - e^{-2N\epsilon}$:

$$|F(x) - F_n(x)| \leq [1-A] \cdot \frac{N}{N+2} \cdot 2^{-(q+1)} \tag{18}$$

Therefore, the empirical entropy should be bounded as follows:

$$H(b_q^q) \geq -1 \cdot \sum_{i=0}^{2^q-1} \left[F(i/2^q) - F((i+1)/2^q)\right] \times \log_2 \left[F(i/2^q) - F((i+1)/2^q)\right] \tag{19}$$

$$\geq -1 \cdot \sum_{i=0}^{2^q-1} \left[F_n(i/2^q) - F_n((i+1)/2^q) - 2\epsilon\right] \times \log_2 \left[F_n(i/2^q) - F_n((i+1)/2^q) - 2\epsilon\right] \tag{20}$$

$$= -2^q \cdot \left[\frac{N}{N+2} \cdot \frac{1}{2^q} - 2\epsilon\right] \times \log_2 \left[\frac{N}{N+2} \cdot \frac{1}{2^q} - 2\epsilon\right] \tag{21}$$

$$= -1 \cdot \frac{N}{N+2} A \times \log_2 \left[\frac{N}{N+2} \cdot \frac{1}{2^q} \cdot A\right] \tag{22}$$

$$\geq q \cdot \frac{N}{N+2} A \tag{23}$$

Which, plugging in the appropriate ϵ into the probablity bound, yields the desired result. □

Plugging in $A = 2/3$, the bound in Theorem 1 holds with probability greater than 0.99 for a 5-bit balanced geohash when N is at least $1e5$ and for a 10-bit geohash when N is at least $1e8$. We will see in the our empirical examples that the rate of convergence of the entropy to q is typically much faster.

3 Census Data Example

3.1 Population based

	V2	V3	V4	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
alabama	0.00	0.88	6.74	0.00	0.85	6.30	13.13	1.00	5.00	10.00	15.64	15.84	1.00	4.95
alaska	0.41	1.94	5.41	0.43	1.94	4.91	10.36	1.00	5.00	9.91	12.26	12.45	1.00	4.89
arizona	0.00	0.95	5.08	0.00	0.84	4.53	12.11	1.00	5.00	10.00	15.39	15.74	1.00	4.92
arkansas	0.09	0.51	6.72	0.11	0.53	6.25	12.64	1.00	5.00	9.99	15.11	15.35	1.00	4.96
california	0.00	0.78	6.59	0.00	0.78	6.23	13.88	1.00	5.00	10.00	16.91	17.65	0.99	4.95
colorado	0.00	0.86	5.19	0.00	0.83	4.66	11.82	1.00	5.00	10.00	15.27	15.60	1.00	4.91
connecticut	0.00	0.00	4.31	0.00	0.00	4.05	11.13	1.00	5.00	10.00	14.56	14.61	1.00	4.91
delaware	0.00	0.98	3.55	0.00	0.94	3.20	9.51	1.00	5.00	9.94	12.89	12.92	0.97	4.86
D.C.	0.00	0.00	1.22	0.00	0.00	1.23	7.24	1.00	5.00	9.76	11.33	11.33	0.99	4.94
florida	0.00	0.98	6.57	0.00	0.97	6.25	13.72	1.00	5.00	10.00	16.46	16.99	1.00	4.97
georgia	0.00	1.00	6.48	0.00	1.00	6.13	13.46	1.00	5.00	10.00	15.61	15.88	1.00	4.94
hawaii	0.00	0.81	3.35	0.00	0.80	3.32	9.34	1.00	5.00	9.91	12.24	12.30	0.99	4.88
idaho	0.74	1.41	5.44	0.77	1.40	5.19	11.30	1.00	5.00	9.99	14.35	14.57	1.00	4.93
illinois	0.37	0.83	5.75	0.35	0.77	5.07	12.60	1.00	5.00	10.00	16.61	17.19	0.98	4.87
indiana	0.00	0.75	6.52	0.00	0.73	6.14	12.83	1.00	5.00	10.00	15.99	16.39	1.00	4.94
iowa	0.00	0.00	6.79	0.00	0.00	6.38	12.17	1.00	5.00	10.00	15.25	15.83	1.00	4.98
kansas	0.00	0.38	6.01	0.00	0.33	5.62	11.76	1.00	5.00	10.00	15.00	15.59	1.00	4.95
kentucky	0.00	0.00	6.55	0.00	0.00	6.08	12.71	1.00	5.00	10.00	15.09	15.27	1.00	4.95
louisiana	0.32	0.32	6.08	0.28	0.28	5.49	12.35	1.00	5.00	10.00	15.29	15.54	1.00	4.93
maine	0.49	0.59	5.89	0.46	0.56	5.46	11.39	1.00	5.00	9.99	13.95	14.03	1.00	4.89
maryland	0.00	0.87	4.69	0.00	0.79	4.34	11.53	1.00	5.00	10.00	14.93	15.06	1.00	4.93
massachusetts	0.00	0.00	4.72	0.00	0.00	4.32	11.58	1.00	5.00	10.00	15.54	15.68	0.98	4.91
michigan	0.28	0.28	6.44	0.26	0.26	5.88	13.20	1.00	5.00	10.00	16.36	16.73	1.00	4.94
minnesota	1.00	1.00	6.11	0.98	0.98	5.35	12.11	1.00	5.00	10.00	15.59	16.00	0.98	4.89
mississippi	0.93	1.73	6.89	0.96	1.74	6.48	12.73	1.00	5.00	9.99	15.02	15.22	1.00	4.95
missouri	0.23	0.61	6.52	0.24	0.60	6.08	12.89	1.00	5.00	10.00	15.84	16.26	1.00	4.92
montana	0.03	0.98	6.33	0.04	1.00	5.79	11.13	1.00	5.00	9.99	14.09	14.38	1.00	4.93
nebraska	0.00	0.32	5.60	0.00	0.32	5.21	11.16	1.00	5.00	10.00	14.78	15.28	1.00	4.93
nevada	0.00	0.72	3.35	0.00	0.71	2.91	10.55	1.00	5.00	9.98	13.78	13.94	0.99	4.92
new_hampshire	0.01	0.01	4.70	0.01	0.01	4.48	10.70	1.00	5.00	9.98	13.65	13.70	1.00	4.86
new_jersey	0.00	0.17	4.69	0.00	0.17	4.15	11.56	1.00	5.00	10.00	15.83	16.01	0.96	4.87
new_mexico	0.00	0.84	5.46	0.00	0.83	4.89	11.23	1.00	5.00	9.99	14.43	14.70	0.98	4.90
new_york	0.00	0.28	5.28	0.00	0.26	4.15	11.76	1.00	5.00	10.00	16.25	16.64	0.97	4.83
north_carolina	0.00	0.89	6.89	0.00	0.90	6.59	14.01	1.00	5.00	10.00	16.08	16.32	1.00	4.97
north_dakota	0.00	0.78	5.71	0.00	0.73	4.61	9.80	1.00	5.00	9.98	13.71	14.07	0.99	4.88
ohio	0.00	0.61	6.59	0.00	0.63	6.17	13.45	1.00	5.00	10.00	16.40	16.83	1.00	4.95
oklahoma	0.00	0.05	6.25	0.00	0.05	5.67	12.35	1.00	5.00	10.00	15.35	15.80	1.00	4.95
oregon	0.99	1.29	5.46	0.99	1.29	5.10	11.87	1.00	5.00	10.00	14.96	15.25	1.00	4.96
pennsylvania	0.00	0.84	6.39	0.00	0.86	5.96	13.18	1.00	5.00	10.00	16.65	17.17	1.00	4.96
rhode_island	0.00	0.00	2.55	0.00	0.00	2.35	9.34	1.00	5.00	9.97	13.29	13.31	0.99	4.88
south_carolina	0.00	0.96	6.21	0.00	0.98	5.85	13.07	1.00	5.00	10.00	15.31	15.49	1.00	4.96
south_dakota	0.54	1.33	6.08	0.51	1.32	5.53	10.56	1.00	5.00	9.97	13.81	14.11	1.00	4.94
tennessee	0.20	0.20	6.49	0.28	0.28	6.08	13.29	1.00	5.00	10.00	15.69	15.93	1.00	4.96
texas	0.00	0.94	7.27	0.00	0.85	6.74	14.07	1.00	5.00	10.00	16.71	17.56	1.00	4.95
utah	0.00	0.59	4.33	0.00	0.61	3.89	10.86	1.00	5.00	10.00	14.44	14.63	1.00	4.85
vermont	0.04	0.04	5.32	0.04	0.04	4.98	10.58	1.00	5.00	9.93	12.92	12.95	0.99	4.92
virginia	0.00	0.74	6.03	0.00	0.73	5.62	12.77	1.00	5.00	9.99	15.51	15.72	1.00	4.93
washington	0.00	0.09	5.76	0.00	0.08	5.35	12.48	1.00	5.00	10.00	15.47	15.76	1.00	4.95
west_virginia	0.00	1.29	6.20	0.00	1.26	5.78	12.13	1.00	5.00	9.99	14.61	14.79	1.00	4.95
wisconsin	1.08	1.08	6.59	0.93	0.93	5.95	12.49	1.00	5.00	10.00	15.88	16.25	1.00	4.94
wyoming	0.00	0.00	5.75	0.00	0.00	5.42	10.15	1.00	5.00	9.97	13.43	13.59	1.00	4.92

4 Distributed Application

5 Conclusions