URLs, hashes

Sevki

Tue Feb 3 19:09:28 CES 2009

Abstract

This article is briefly describes how to setup a sample application that works with OPC clients.

Building a url shortener because I thought it would be cool to have a QR image with Clint Eastwood that has the url of the current page encoded because fistful of bytes. Before we go any further bookmark the link below, or just go and read the amazing writeup by Russ Cox

Figure 1: qrs

Problem as it turns out the image on the left is you can see you can't make out blondie, the URL `http://fistfulofbytes.com/how-to-bypass-ssl-validation-for-exchange-webservic` is 91 characters long. On right hand-side `http://lea.cx/MQb2dg==` is encoded in QR because the URL is considerably shorter, you can sort of make out blondie's outline.

Straight forward way

Have a sql db, stick urls one by one, look up if a url has been been inserted, return the id if it has and so on. Problem with this approach is, every time you want to shorten a url you have to query a server. Can we do without the querying bit?

If something that needs to be less of that same something, we compress it right?

Plan B: Compression

Traditional compression methods such as `gzip bzip lzw` are not effective on short text. However there is a better solution by the great mind of antirez called `smaz` which does just that, make long urls shorter.

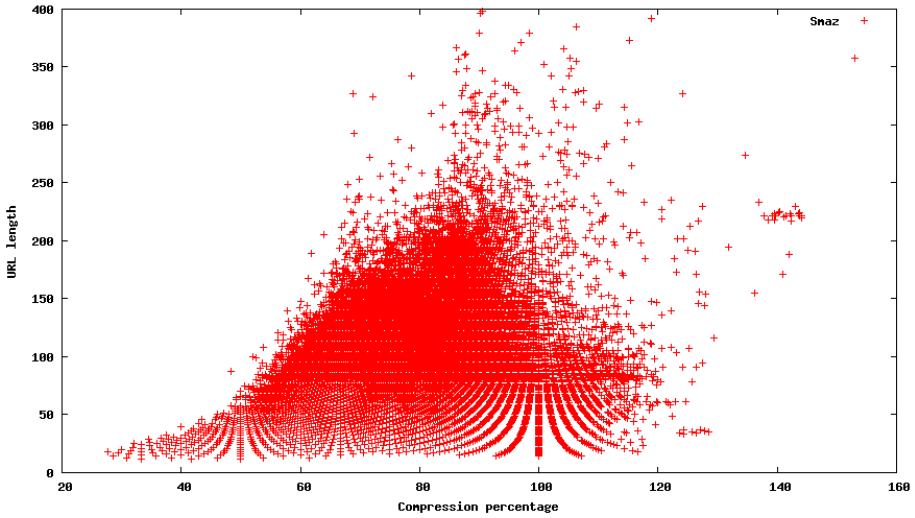


Figure 2: smaz

It works fairly well if you are storing gajillion urls. However if you are trying to make a url shortener it'll only compresses it to about 75%, which doesn't work 100% of the time, but even when it does, it assumes that you have about 100-250 characters to begin with so getting the size to 60% is an act in futility.

Plan A: Hashing

Ok, so compression was never going to work but hashes are probably going to work.

Why Hash?

Hashes are deterministic functions that always return `a` for any given `f(x)`. Most of them are also standardized by NIST and IEEE so they remain deterministic across devices and languages; therefore one might guess what the key of an object is going to be on another system, given that there are not collisions.

When building a URL shortener they come in handy; for instance given that you know that a `id:url` tuple exists, you don't have to query your service to serve a shortened version of the url, you can just serve the hash of the url. This saves you a roundtrip to your service.

Does it work?

YES. Here is comparison of 242 286 unique urls that are gathered from HN hashed with `crc16`, `crc32` and `crc64`.

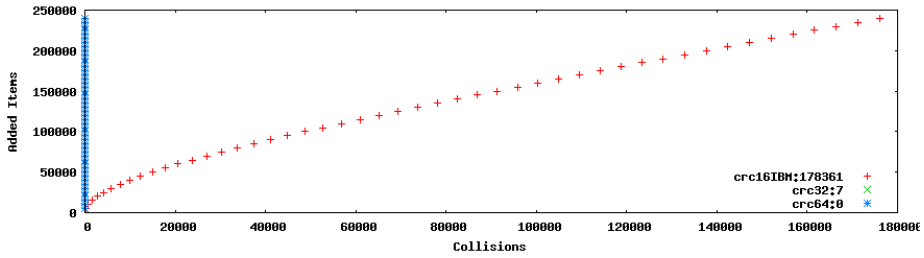


Figure 3: crc

As you can see the 4 byte and 8 byte hashes nearly and don't collide at all respectively. And 2byte hashes behave in an extremely predictable way when you are hashing 17mb of urls. Clear winner in this race is crc32 seeing that in 242286 we got 7 collisions out of only 4 bytes. Which is negligible, or sheer bad luck. When talking about hashes, bitwise length is raised as 2's power to calculate possible combinations as such:

bits

bytes

possible combinations

16

2

65,536

32

4

4,294,967,296

64

8

18,446,744,073,709,551,616

Seeing that there are a possible 4 294 967 296 combinations, 7 collisions seem like a happy conincidince.

Encoding, UTF-8, base64

Unfortunately there is a caveat, our 4 bytes as is, are not url safe nor are they type-able on a qwerty keyboard. Because type-able ascii characters are 7 bits, when you divide 32bits into 4 bytes and use as text each one thats greater than 127 should look like gibberish, but in reality it's even less than that because we want to use the stuff that doesn't need modifier keys to be typed. Enter http://en.wikipedia.org/wiki/Base64. Encoding something in base64 is the process of taking a bitmap and slicing it into 6 bit characters and mapping those bits to base64 chars. But a 4 byte array such as this `49,6,246,118` encoded in base64 looks like this `MQb2dg==`, and the length jumped to 8.

Can we go from 8 to 4?

Would halving the bytes encoded also half the encoded string length, it probably will. But we really don't need to go that far, for instance if the hash is just 2 bytes `49,6` it will be `MQY=` encoded in base64. Just like before there are `=` signs at the end which is padding. Because neither 32 nor 16 are exact multiples of 6 we need to pad them, but 24 is, which is 3 bytes.

49

6

246

118

00110001

00000110

11110110

01110110

M

Q

b

2

d

g

=

=

49

6

246

00110001

00000110

11110110

M

Q

b

2

49

6

00110001

00000110

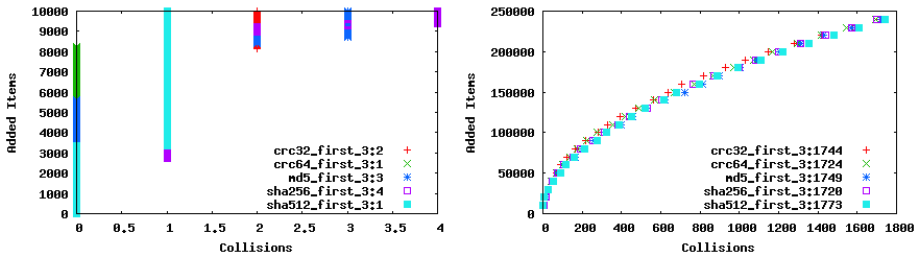
M

Q

Y

=

So how well does 24bit hashes work? Great, the thing about hashes are, if you have good enough entropy they work in an extremely predictable way.



Comparing first 3 bytes different hashes, they all perform very similarly when tested over a huge set, `md5` works very similarly to `crc32`.

In a scenario where there is sufficient randomy goodness in the original hash, if you take a smaller slice of the original hash it will preform very similar to other hash slices of the same size over a very large dataset. So toss a coin or choose the least computationally expensive hashing function, and stick with it because it's highly unlikely to get a collision until there are 4000~ urls, so this is a rather simpler solution to this particular problem.