

Re: What is md5sum?

Source: <http://linux.derkeiler.com/Newsgroups/comp.os.linux.setup/2004-06/0985.html>

From: Thomas Richter (*thor_at_cleopatra.math.tu-berlin.de*)

Date: 06/30/04

Date: 30 Jun 2004 11:22:34 GMT

> *Micha? Kosmulski <M.Kosmulski@nospam.elka.pw.edu.pl> wrote:*

>> *There is no need to be 2^{128} different files. 2 is enough to get a collision
>> if it's the right two files.*

> *Provide me them, or a construction that provides them (and a computer
> program that lists all of them and calculates the md5sums won't work –
> it would die before getting anywhere at all).*

There's a pretty simple proof that there must be two files with the same md5sum. You can either call this "The Counting Theorem" or "The Pigeon Hole Principle", or, as a mathematician would put it, "a mapping from a set A to a set B cannot be one-to-one if the cardinality of B is smaller than that of A". You know, one can prove things without giving examples. Luckily, we're beyond that stage for quite a while.

And, of course, the above algorithm will work. It's just not pretty efficient for today's machines. By analysing the hashing done in md5, one can of course do better. (I guess it's just a polynomial over the ring \mathbb{Z}_2 , so maybe some theory about the polynomial ring can be applied).

>> *Strangely enough, mathematics seems to be used for quite a few things
>> these days. While it is not a strict representation of our world (which
>> it doesn't try to be), it gives quite good results. I don't know how far*

> *That's OK – but remember that it's an approximation.*

Mathematics is not an approximation of anything. It is a set of rules agreed upon for reasoning. It might get an approximation as soon as you apply it to physical objects and formulate rules about physical objects in mathematics – these rules are as "precise" as they could be, but they only allow predictions of a finite precision. (And, as a very important rule, a physical theory worth its name should give you also the limits of its applications).

> *An "abstraction",*

- > if you prefer. A "theory". The theory has rules of formal logic and
- > that's all fine. The problem comes when you try and move from the
- > theoretical realm across into making predictions using the theory. Then
- > you have several problems to consider. At least:

- > 1) are the formal logical rules of reasoning a proper abstraction of
- > the way the universe is?
- > 2) are the logical hypotheses (tenets, beliefs) proper abstractions
- > of the situation in the real universe?

- > Nobody has bothered to consider that there aren't and can't be 2^{128}
- > (say 2^{100} , to please you!) files examined by any computer anywhere on
- > earth, therefore no computer can ever carry out the constructive
- > procedure you propose.

Now, is this a proof? No, that's in fact a non-argument. For example, it doesn't prove that there can't be any other, more efficient algorithm to create these files. Neither, it doesn't disprove that there can't be two such files. In fact, it is easy to prove that there can.

- >> engineering would be today if someone rejected using integrals since
- >> "they are a sum of infinitely many infinitely small components, and that

- > Riemannian integration has a perfectly well-defined theory that does
^^^^^^^^^^^^^^

Oh please. The name of the guy is "Riemann". (-;

- > not require infinitesimals (although you can use nonstandard logics
- > which do use infinitesimals in order to construct the Riehamannian
- > theory).

Standard analysis integration at least requires taking limits, and this again requires assumption of the existence of objects that are only constructable by a non-stopping algorithm. The problem starts already at the construction of the real numbers. "There is no such thing in the universe since there are only a finite number of particles, thus a finite precision and no limit". Now, taking your argument from above, one shouldn't compute with "reals" since they don't exist, and there can't be a computer application that uses them. In terms of this "limited computability", even natural numbers don't exist, only the subset of the interval $[0..10^{80}]$ (10^{80} = rough approximation of the number of particles known in the universe).

- > It's constructed on notions of limits – that is, you can get
- > as close as you like to the integral by making calculations as fine as
- > you like.

No, taking your argument, not "as fine as you like". Not finer than the number of digits representable by possible states of the universe. Now what?

>> *A dramatic misunderstanding ! One doesn't have to show them to prove they exist.*

> *Oh, yes one does.*

Not at all. For a mathematician, it is enough to give a prove for existence by showing that non-existence yields to a contradiction. Or at least, for "main stream mathematicians".

> *What makes you think you don't? (you are supposed to think about how you may convince me).*

Look, you can take your own view on mathematics as you like, and use your own set of axioms (for example, by dropping the Axiom of Choice), but you'll have a harder time, and you can prove less.

> *Sure, you can tell me about the properties of the number I chose. So? Can you show me that there is a decimal sequence that is impossible to describe?*

No, all sequences are possible to describe. But some sequences are harder to describe than others, and some of them are so hard to describe that you need **at least** the number of digits of the sequence to do that. Actually, the least number of digits required to describe a sequence is called its Kolmogorov-complexity. Interestingly, you can't compute the Kolmogorov-complexity of an arbitrary sequence, so you can't check whether a given sequence is "complete" or "complex" in the above sense. Thus, this is a non-computable property.

A nice example: Go check for the "A Million Random Digits File" on Google. There's currently believed to be as much redundancy as 24 bits in it. Thus, while it is not "complex" in the above sense, it's pretty damn close. And you need possibly a million digits minus ten or twenty to describe it.

> *What's the matter? You "know" that there is such a nonempty set, but you find yourself unable to "choose" a representative element to show me?*

Yup.

> *Great. Now you know that "choosing" is not as axiomatic a thing as you may think. In fact it IS an axiom – formally unprovable, and formally impossible to contradict – that you may choose an element of any set.*

That is not a question. I can always do that. The Axiom of Choice (if that is what you're talking about) is about something different, and harder. (For an infinite sequence of sets, it is always possible to pick a sequence of elements, one from each set.)

comp.os.linux.setup: Re: What is md5sum?

- > *Fantastic, so even though you "know" that there "are" two files with*
- > *the same md5sum, you find yourself curiously unable to CHOOSE a pair to*
- > *show me!*

No, that is a *finite* problem, and not a matter of the Axiom of Choice. It is very well possible to build an algorithm that finds two. The best one is possibly just damn slow. Though, possibly that's not even an NP-complete problem.

>> *Just any 2 which have the same hash code.*

> *Show me them.*

The point is – you don't have to.

- > *It's not impossible that the moon will decide to change its orbit and*
- > *crash on your head tomorrow.*

A completely different story. This is physics, and has all its known limits of the applications of the theory. We're talking about mathematics. While it is possible that the moon crashes into the earth, just very unlikely, it is not possible that $2+2 = 5$, never ever. And for the same reason, two files exist with the same md5sum.

- > *But I am willing to bet you all the money*
- > *on earth (and what I own) that tomorrow nobody will find two files with*
- > *the same md5sum. That is truth.*

Oh great. Here's a business proposal for you ("The md5sum compressor") to test your ideas in the real world. After all, mathematics is one thing, successful business another, and if you're true, you can actually make some money out of it:

Given your state of understanding, an md5sum is unique (always, ever). Thus, I can compress every(!) single(!) file to its md5sum, and represent, say, a gigabyte harddisk just by its md5sum either. Thus, we've a super-duper compression algorithm. Expansion is a bit harder, but there's at least the trivial "try and error" algorithm (try all files of all length, compute the md5sum, check whether they're equal) but possibly there's a better one. Maybe you could just store your files in a big archive, and just make a table lookup thru the md5sums – I don't care. Given the fact that this is an ideal file compressor, you could make quite a lot of money since people won't have to have big harddisks (your archive might have, but though what – that'll pay), and all I – the customer – need is a single floppy disk to store the md5sums.

>> *People seem to have little understanding of how probability works.*

- > *No, you don't. And nobody knows how probability works, because it's not*
- > *a physical theory – try reading some books on the subject of how*
- > *probability theory may relate to reality, if at all.*

Re: What is md5sum?

Seems to be pretty usable for several purposes at all.

>> *The fact that probability of something is very low doesn't mean it's
>> never going to happen.*

> *You seem unfortunately confused between probability and statistics. The
> probability of something being low says nothing at all about whether
> the event associated with that probability will happen or not.*

And a mathematical prove doesn't say anything about probability. As a matter of fact, if you accept the proof, it is true, and that's it. Mathematics describes probabilities, but it doesn't work with it for its theorems.

>> *It probably won't ever happend, but it could.*

> *You seem confused. What are the experiments you are performing?*

Math is not about experiments.

> *Are you going to run the universe once, see if it happens, then stop the
> universe, restart it and run the experiment again?*

No need to consider the universe. It's an abstract language, I don't need no #!@! universe to do math. I can easily compute the volume of a cube in 123 dimensions, but there's nothing like that around here.

>> *So, we should not be afraid of someone finding a collision in MD5, but
>> strictly speaking, it is possible and might happen some day.*

> *No, it will not happen any day.*

I don't know whether it will or will not happen. With some skill, it might be even possible to compute the required files. It is, however, provable that such files exist, and one can really create them as real existing files, on a real existing harddisk if one really wants to. The counting argument just doesn't say a word how to get these; that's one of the abstractions one has to deal with when doing math.

> *Let me put it this way – if you think
> it is possible, then why not bet me on it? I am happy to bet you!*

Existence doesn't require examples, that's the point. I'm pretty sure there's a 10^{1000} prime number, but I cannot print it here. I can possibly tell you the approximate number of digits it might have – but that again requires you to believe in the Riemann Hypothesis, which is still unproven (though generally believed to be correct). All I can give you is a prove that there are infinitely many, and thus even a 10^{1000} 'th must exist. It's just damn large.

comp.os.linux.setup: Re: What is md5sum?

So long,
Thomas