Measuring Information

Conference or Workshop Item

Measuring Information

Abstract

It is easy to look at developments arising from digital technology such as social media, smart phones and tablets, digital photography, and go along with the general perception that there is vastly more information around than ever before. Claiming that there is more information around, however, assumes that it is possible to quantify information, and that in turn seems to assume that we know what information is.

Engineering disciplines have measures for what they call information. Claude Shannon used one in his celebrated work on a mathematical theory of communication which led to the discipline of information theory, and a variation known as algorithmic information theory (AIT) was developed by Andrey Kolmogorov, Ray Solomonoff and Gregory Chaitin.

It has long been debated whether these information theories have relevance to semantic information, with some authors dismissing them on the grounds that they are not about the content of information. However, using the example of the information contained in a school report, this paper shows how information theory can quantify semantic information. Based on the modelling developed by Shannon, it is shown that school reports generated with the assistance of dedicated report-writing computer programmes contain a lot less information than it might appear at first sight, and may contain a lot less information than a hand-written report from the days before digital technologies.

Keywords

Information Theory, Algorithmic Information Theory, Kolmogorov Complexity, School Reports

Introduction

The idea that there is more information around than ever before is widely taken as a given fact of the age. Alongside the technological developments leading to a genuinely exponential increase in the size of computer memory, the rise of social media has recruited mass-participation to the task of filling-up all the ever-expanding storage capacity. With cameras on mobile phones and wireless broadband access, photographs and videos are taken and shared at will, while text, images and streamed or downloaded audio and video is available from multiple servers to inform or entertain 24/7.

From time to time attempts are made to quantify the amount of information being generated, and, for example, in *The Ethics of Information*, Luciano Floridi[1] talks about the 'zettabyte era', quoting a study which had shown that:

> In 2011 the amount of information created and replicated will surpass 1.8 zettabytes (1.8 trillion gigabytes) – growing by a factor of 9 in just five years"

This figure was arrived at essentially by counting data bits. It is not quite as crude as directly counting every bit stored in every computer and phone memory, but it is nevertheless counting bits. However it is easy to identify situations in which there are bits that should not be identified as information, or at least need careful consideration before awarding them the status of information. The bits in the computer representing this sentence might reasonably be accounted as information. But what if it contains nonsense text at random? Do the bits that represent the following characters deserve the status of 'information'? sahduos8wbb. Furthermore there are multiple copies of this file on different computers, so the bits are duplicated in several different places. Do the duplicate copies constitute additional information? These are just two examples of the reasons why we should be wary of counting bits in order to measure information.

This paper therefore starts by addressing the question of what information is. It then looks at two engineering measures of information: that based on the work Claude Shannon ("Shannon Information") and that based on the work of Andrey Kolmogorov, Ray Solomonoff and Gregory Chaitin (algorithmic information). It illustrates how these measure might be exploited using the example of a school report and closes with some conclusions about information.

What is information?

Descriptions that have been used for information include 'data with meaning' or 'data in context'[2] and Gregory Bateson's 'difference that makes a difference'[3], but at best these need some unpicking. A useful categorisation which is sometime ascribed to Norbert Wiener distinguishes between matter, energy and information (arguably this distinction is, somewhat self-referentially, a difference that makes a difference).

Consider, for example, a digital device such as a smart phone. Matter (the metal, plastic, glass, silicon used to build the phone) and energy (the charged battery) are uncontroversial, but what is important about the phone is the information: the multimedia (text, audio and video) that it is used for, but also the design of the phone. We know that the multimedia exits as binary data, patterns of 1s and 0s, but beyond that, what is it? This three way categorisation is not restricted to manufactured artefacts, since the categories of matter, energy and information are equally relevant to the understanding natural objects, including biological entities such as plants and animals.

Although matter and energy are uncontroversial, we still can not really say what they are. There are few people who have understood physics as well as Richard Feynman, yet he said of energy[4]:

> "[I]n physics today, we have no knowledge of what energy is. […] However, there are formulas for calculating some numerical quantity, and when we add it all together we get "28" – always the same number."

Feynman is saying that while we can not say what energy is, we can measure it. It may be the same with information but that takes us back to where we started because the reason for wanting to know what information was in order to be able to measure it. We have come full circle, and perhaps understanding information is iterative.

A way in to the loop is through the existing measures of information from the field of computing and communications engineering.

Shannon Information

For many people the age of information began with the publication in 1948 paper of *A Mathematical Theory of Communication*, by Claude Shannon[5]. David Mackay said that this paper "both created the field of information theory and solved most of its fundamental problems"[6] and it has certainly been influential, having received more 10,000 citations as of

December 2014 according to the Web of Science Database. Though it addressed the specific task of modelling electrical communication systems, the breadth of fields among those 10,000 citations – including such unexpected topics as Pediatrics, Fisheries, Public Administration, Women's Studies, Art and Religion – reveals that it has found application a long way beyond its origins.

Shannon was working on electrical communication systems and modelled that system with a diagram that itself has become an iconic on the world of information theory. This model is shown in Figure 1, but with a slightly changed layout to emphasize the communication of a message from the source to the destination.

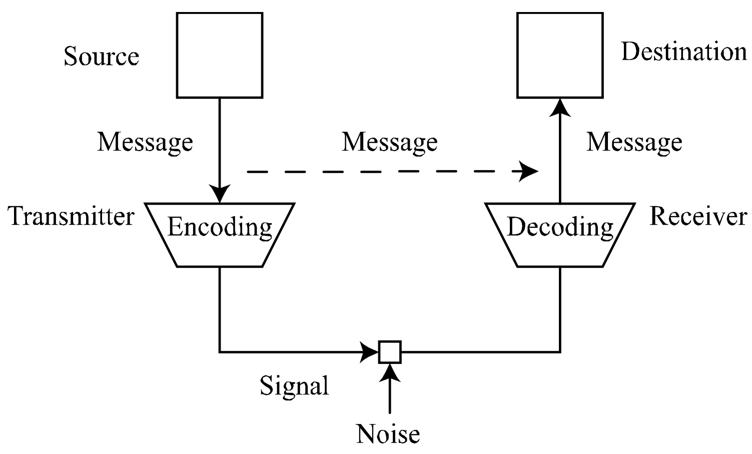

Figure 1. Shannon's model of a communications channel.

Shannon supposes that a source has messages to send to the destination. The source sends a continuous stream of messages, one at a time, each selected from a finite set of possibilities. Each message is encoded by the transmitter and sent as a signal through the channel. The signal might be corrupted by noise in the channel, then is decoded by the receiver, ideally recovering the transmitted message which is delivered to the destination.

Shannon wanted to know the rate at which messages could be sent down the channel, and he derived his famous equation:

$$C = B\log_2\left(1 + S/N\right)$$

B, S and *N* are physical, measurable features of the channel: the bandwidth, *B*; the power of the signal, *S*; and the power of the noise, *N*. *C* is called the channel capacity and is measured in bits per second, but it is what Shannon – following the work of others such as Hartley – called information.

One way of interpreting *C* is that it is the rate at which binary choices can be sent down the channel. This can be measured in information bits, where one bit represents the selection between two equally-probably alternatives. These information bits, as opposed to regular data bits, are sort of idealized bits that are doing the maximum work any bit can do. Every information bit supplies one binary decision about something which would otherwise be equally likely to be either of the two possibilities.

The world does not come made out of binary decisions ready to be conveyed by information bits, but Shannon's paper showed how to calculate the equivalent information content of real messages. Simplified, the calculation is based in the idea that if you are communicating a message and the probability of sending that message is *p*, then the information content of the message measured in bits is $-\log_2 p$ (or, equivalently, $\log_2\left(1/p\right)$). Notice that this is consistent with the idea that a single decision about something with two equally-likely outcomes delivers one bit of information, because in that case the probability, p, would be 0.5 and $\log_2(1/0.5) = \log_2 2 = 1$.

The idea of Shannon information becomes more intuitive with more examples. Consider parents communicating information about their new baby. First, they send a message about the gender. Given that boys and girls are equally likely, this is one bit of information. Suppose it is a boy, and they then send a message with the name. One approach to understanding the information content of the name is to consider what names are possible. Suppose that the parents are English and we can assume they use an English name. A list of English boys names identifies 610 possibilities. If they are all equally likely, the probability of any of them is 1/610 and the information conveyed by one of them is $-\log_2(1/610) = \log_2 610 = 9.25$ bits. Names are not all equally likely, however. Of the 358,438 boys born during 2013 in England and Wales, 6,949 of them were called 'Oliver'[7], so, with no further information, the probability that a boy was called Oliver in 2013 was 6,949/358,438 = 0.0194 and the information content of a message saying a boy is called Oliver that year was $-\log_2 0.0914 = 5.69$ bits. Six boys that year were called Theon[8], so the probability that a boy was called Theon was 6/358,438 = 1.67 x 10^{-5} and the information content of a message announcing this name was $-\log_2\left(1.67\times10^{-5}\right) = 15.87$ bits. In general being told something improbable provides more information than being told something you expected to be the case anyway (there is more information in being told you have won the lottery than in being told you have not won it).

Returning to the question of what information is, Shannon's work has come in for criticism on the grounds that it is not about semantics, so how can it be anything to do with information?

Shannon himself said it his 1948 paper:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem."

But to say this makes it irrelevant to information is to miss the point. As Donald Mackay pointed out[9]

> "The trouble here appears to be largely due to a confusion of the concept of information with that of *information-content* – the confusion of the *thing* with the *measure* of the thing." (Emphasis in original)

As explained earlier, knowing how to measure something – be it energy of information – may not be the same as knowing what the thing is, but it is one aspect of understanding it.

Algorithmic Information Theory

Shannon information is one strand of information theory, and it is fair to say is the mainstream information theory, but another strand is Algorithmic Information Theory, AIT, which was devised independently by Andrey Kolmogorov and Ray Solomonoff, and has been further developed by Gregory Chaitin. It is also frequently referred to as Kolmogorov Complexity or Kolmogorov Information.

In a standard textbook on the theory[10], Ming Li and Paul Vitanyi have an introductory example along the lines of the following. Suppose you toss a coin 50 times and write down a 1 if you get a head and a 0 if you get a tail. You do it once and get the following result.

0100101110111101001101011000010001000111110

You then do it again and get a head every time, giving the next sequence.

0000000000000000000000000000000000000000000

According to simple probability theory, these two are equally likely outcomes. Every possible sequence of 50 bits is equally likely, so the particular mix of heads and tails represented by the first sequence just as likely as getting 50 tails, as represented by the second sequence. (They are in fact both extremely unlikely, with a probability of $1/2^{50}$ which is about one in 10^{15}, one in a thousand million-million.)

Intuitively there does appear to be a qualitative difference between these sequences, and in AIT there is a difference on the basis that it is possible to describe one of them much more concisely than the other. It is possible to describe the second sequence by saying “repeat 0 fifty times”, whereas there is no such simple description of the first.

Sequences do not necessarily have to be obviously simple in order to have a short description. Consider the following 50 digits.

314159265358979323846264338327950288419716

This might just look random, or else you might spot that it is the first 50 digits of the number pi. It can be described concisely by saying “the first 50 digits of π”, but that presupposes that the reader knows the digits of the number π. However, it is possible to write an algorithm (a computer programme) that generates the digits of pi. Such an algorithm is unlikely to be shorter than 50 digits, so it would not be a concise way of describing these digits, but the principle could be used for communicating, for example, a sequence of 10,000 digits in which case the algorithm might well be much more concise than sending the 10,000 digits explicitly. In this way, AIT defines the ‘Kolmogorov complexity’ of a sequence x as K(x), where K(x) is the shortest algorithm that can generate the sequence x.

While AIT can be seen as ‘merely’ about data compression, it has also been interpreted as having more profound significance. It is claimed to be about finding the absolute minimum number of bits needed to describe a sequence, and therefore to be a measure of the absolute information content of the sequence.

Li and Vitanyi argued that AIT is superior to Shannon as a measure of information[11].

> “Shannon ignores the object itself but considers only the characteristics of the random source of which the object is one of the possible outcomes, while Kolmogorov considers only the object itself to determine the number of bits in the ultimate compressed version irrespective of the manner in which the object arose.”

The problem, however, is that according to AIT the sequence with the highest information content is one consisting of an entirely random sequence of digits. Arguably, meaning – semantics – is entirely absent from AIT, whereas a calculation of the amount of information based on Shannon's modelling, requires an exploration of the message content.

An example based on school reports is instructive.

The information content of a school report

Forty years ago, in the days before personal computers and word-processing, a school report was typically a slim document with a few handwritten notes from the teachers. These days, school reports are much bigger. Often a report is a booklet with lots of detail about what the pupil has been doing and has achieved in each subject. The report appears to contain a lot of information, apparently a lot more than in the days of small hand-written (or typed) reports.

Today's reports, however, are generated with a lot of help from computer programmes. An example of report-writing software for use in schools following the National Curriculum for England and Wales worked along the following lines[12].

The teacher enters the name of the pupil and then chooses the grade of the pupil for each a subject, selecting from a pre-defined set such as: *h* for higher achiever, *m* for average to more able achiever, *l* for average to less able achiever and sen for students with special educational needs. The software then writes the whole report, drawing on the statements contained in the National Curriculum for England. A randomising element ensures that no two reports are identical, even if the grades are the same.

For example, in one version of this software, entering *m* for ICT for one pupil (John), generated the following

> John has extended his knowledge of a variety of computer programs and he can log into the network without support. He has explored a variety of features included in software for composing music and is aware that questions can be turned into search criteria when using data handling programs. He has found information relating to his topic work from given websites on the worldwide web and explains patterns that govern a computer simulation

This is 442 characters. A text file containing this paragraph has 442 bytes, which is 3,536 bits, but the amount of information will be much less.

Algorithmic information theory would measure the information content by finding the smallest number of bits required to generate the 442 bytes of the file. There is not an simple answer to this, but a first approximation which would give an upper limit is to use a standard compression algorithm. Zip, for example, compresses the file to 388 bytes, but this will be a long way from the minimum.

AIT considers the file itself, but to use ideas from Shannon information theory we need to consider the message that the file is conveying. In this case, given that we are interested in the information about the pupil being conveyed from the teacher (the message source, in Figure 1) to the parents (the destination), all of that information is contained in the selection of *m* (the message) by the teacher. The teacher judged John to be *m*, rather than *h*, *l* or sen. If the four options are equally probable the information content is $\log_2 \left(1/0.25\right) = 2$ bits. The amount of information conveyed about John from the teacher to the parents by the 3,536 bits of the file is 2 bits of information.

There are many reasons for the difference between the file size and the amount of information. The report contains information about the National Curriculum, for example, so some of the bits carry information about that. However, the important insight from this analysis is that the amount of information conveyed from the teacher to the parents about the pupil is 2 bits.

A very simple comparison with an old-fashioned hand-written report is possible by considering the choice of words available to the teacher when writing the report. The amount of information in the computer-generated report was so low because the teacher was selecting one from only four possible options. When writing a report, the options available to the teacher are in principle vast. Even a short sentence would come from a very large number of different sentences that might be written. The calculation also takes account of the probability of the different sentences, but even so the teacher is likely to be routinely choosing from a lot more than four options, so the information content will be correspondingly higher.

Conclusion

Information is just as real as energy and matter, and can be measured. A convenient unit of information is the information bit. A data bit can carry one information bit, but very often does not. Analysing data using ideas drawn from the information theory of Claude Shannon can help to identify the amount of information that is being carried by the data, sometimes leading to surprising results.

David Chapman

DTMD Group

Department of Computing and Communications

Faculty of Maths, Computing and Technology

The Open University

MK7 6AA

United Kingdom

Email: david.chapman@open.ac.uk

---

[1] Luciano Floridi "The Ethics of Information", (Oxford : OUP, 2013)

[2] See, for example, various chapters of Magnus Ramage and David Chapman (eds), "Perspectives on Information" (New York and Oxford: Routledge 2011)

[3] In several publications by Gregory Bateson, including: Gregory Bateson and Mary Catherine Bateson "Angels Fear" (London: Century Hutchinson Ltd, 1988) p. 17

[4] Richard P. Feynman, Robert B. Leighton and Matthew Sands, "The Feynman Lectures on Physics" Volume I, (Reading Mass.: Adison Wesley 1963), p. 4-2. Emphasis in the original.

[5] Claude E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27 (1948): pp 379–423 and 623–656

[6] David J. C. Mackay, *Information Theory, Inference, and Learning Algorithms*, (Cambridge, UK, Cambridge University Press 2003).

[7] Births data from the UK Office of National Statistics http://www.ons.gov.uk/ons/index.html accessed 15/12/2014

[8] Emily Dugan "Most popular baby names: The top 20 boys and girls names in England and Wales" *The Independent*, 15th August 2014 http://www.independent.co.uk/news/uk/home-news/most-popular-baby-names-the-top-20-boys-and-girls-names-in-england-and-wales-9671635.html accessed 15/12/2014

[9] Donald M. Mackay, "Operational aspects of some fundamental concepts of human consciousness", Synthese, (1953) Vol. 9 Nos 3/5, pp. 182-198.

[10] Ming Li and Paul M. B. Vitányi, *An introduction to Kolmogorov Complexity and its Applications* (New York, Springer 2009)

[11] Ming Li and Paul M. B. Vitányi. op. cit, p. 603

[12] This is based on 'The Report King' which was available for download from http://www.thereportking.co.uk as of 13/12/2014.