Information Distance: New Developments

Paul M.B. Vitányi
CWI, Amsterdam, The Netherlands
(*Invited Lecture*)

Abstract

In pattern recognition, learning, and data mining one obtains information from information-carrying objects. This involves an objective definition of the information in a single object, the information to go from one object to another object in a pair of objects, the information to go from one object to any other object in a multiple of objects, and the shared information between objects. This is called "information distance." We survey a selection of new developments in information distance.

I. The Case $n = 2$

The clustering we use is hierarchical clustering in dendrograms based on a new fast heuristic for the quartet method [5]. If we consider n objects, then we find n^2 pairwise distances. These distances are between natural data. We let the data decide for themselves, and construct a hierarchical clustering of the n objects concerned. For details see the cited reference. The method takes the $n \times n$ distance matrix as input, and yields a dendrogram with the n objects as leaves (so the dendrogram contains n external nodes or leaves and $n-2$ internal nodes. We assume $n \geq 4$. The method is available as an open-source software tool, [2].

Our aim is to capture, in a single similarity metric, *every effective distance*: effective versions of Hamming distance, Euclidean distance, edit distances, alignment distance, Lempel-Ziv distance, and so on. This metric should be so general that it works in every domain: music, text, literature, programs, genomes, executables, natural language determination, equally and simultaneously. It would be able to simultaneously detect *all* similarities between pieces that other effective distances can detect seperately.

Such a "universal" metric was co-developed by us as a normalized version of the "information metric" of [1], [9]. There it was shown that the information metric minorizes up to a constant all effective distances satisfying a mild density requirement (excluding for example distances that are 1 for every pair x,y such that $x \neq y$). This justifies the notion that the information distance is universal.

We may be interested what happens in terms of properties or features of the pair of objects analyzed, say x and y. It can be shown that the information distance captures every property of which the Kolmogorov complexity is logarithmic in the length of $\min\{|x|,|y|\}$. If those lengths go to infinity, then logarithm of those lengths go to infinity too. In this case the information distance captures every property.

This information distance (actually a metric up to minor additive terms) is normalized so that the resulting distances are in $[0,1]$ and can be shown to retain the metric property, [8]. The result is the "normalized information distance" (actually a metric up to neglidgible terms). All this is in terms of Kolmogorov complexity [9].

It articulates the intuition that two objects are deemed close if we can significantly "compress" one given the information in the other, that is, two pieces are more similar if we can more succinctly describe one given the other. The normalized information distance discovers all effective similarities in the sense that if two objects are close according to some effective similarity, then they are also close according to the normalized information distance.

Put differently, the normalized information distance represents similarity according to the dominating shared feature between the two objects being compared. In comparisons of more than two objects, different pairs may have different dominating features. For every two objects,

Affiliation: National Research Center for Mathematics and Computer Science in the Netherlands (CWI). Address: CWI, Science Park 123, 1098XG Amsterdam, The Netherlands. Email: Paul.Vitanyi@cwi.nl

this normalized information metric distance zooms in on the dominant similarity between those two objects out of a wide class of admissible similarity features. Since the normalized information distance also satisfies the metric (in)equalities, and takes values in $[0,1]$, it may be called *"the" similarity metric*.

Unfortunately, the universality of the normalized information distance comes at the price of noncomputability. Recently we have shown that the normalized information distance is not even semicomputable (this is weaker than computable) and there is no semicomputable function at a computable distance of it [13].

Since the Kolmogorov complexity of a string or file is the length of the ultimate compressed version of that file, we can use real data compression programs to approximate the Kolmogorov complexity. Therefore, to apply this ideal precise mathematical theory in real life, we have to replace the use of the noncomputable Kolmogorov complexity by an approximation using a standard real-world compressor. Starting from the normalized information distance, if Z is a compressor and we use $Z(x)$ to denote the length of the compressed version of a string x, then we arrive at the *Normalized Compression Distance*:

$$NCD(x,y) = \frac{Z(xy) - \min(Z(x), Z(y))}{\max(Z(x), Z(y))}, \qquad (1)$$

where for convenience we have replaced the pair (x,y) in the formula by the concatenation xy, and we ignore logarithmic terms in the numerator and denominator, see [8], [3]. In [3] we propose axioms to capture the real-world setting, and show that (1) approximates optimality. Actually, the NCD is a family of compression functions parameterized by the given data compressor Z.

A. Web-based Similarity

To make computers more intelligent one would like to represent meaning in computer-digestable form. Long-term and labor-intensive efforts like the *Cyc* project [7] and the *WordNet* project [11] try to establish semantic relations between common objects, or, more precisely, *names* for those objects. The idea is to create a semantic web of such vast proportions that rudimentary intelligence and knowledge about the real world spontaneously emerges. This comes at the great cost of designing structures capable of manipulating knowledge, and entering high quality contents in these structures by knowledgeable human experts. While the efforts are long-running and large scale, the overall information entered is minute compared to what is available on the Internet.

The rise of the Internet has enticed millions of users to type in trillions of characters to create billions of web pages of on average low quality contents. The sheer mass of the information available about almost every conceivable topic makes it likely that extremes will cancel and the majority or average is meaningful in a low-quality approximate sense. Below, we give a general method to tap the amorphous low-grade knowledge available for free on the Internet, typed in by local users aiming at personal gratification of diverse objectives, and yet globally achieving what is effectively the largest semantic electronic database in the world. Moreover, this database is available for all by using any search engine that can return aggregate page-count estimates like Google for a large range of search-queries.

While the previous NCD method that compares the objects themselves using (1) is particularly suited to obtain knowledge about the similarity of objects themselves, irrespective of common beliefs about such similarities, we now develop a method that uses only the name of an object and obtains knowledge about the similarity of objects by tapping available information generated by multitudes of web users. The new method is useful to extract knowledge from a given corpus of knowledge, in this case the pages on the Internet accessed by a search engine returning aggregate page counts, but not to obtain true facts that are not common knowledge in that database. For example, common viewpoints on the creation myths in different religions may be extracted by the web-based method, but contentious questions of fact concerning the phylogeny of species can be better approached by using the genomes of these species, rather than by opinion. This approach was proposed by [4]. We skip the theory.

In contrast to strings x where the complexity $Z(x)$ represents the length of the compressed version of x using compressor Z, for a search term x (just the name for an object rather than the object itself), the code of length $G(x)$ represents the shortest expected prefix-code word length of the event $\mathbf{x}$ (the number of pages of the Internet returned by a given search engine). The associated *normalized web distance* (NWD) is defined just as (1) with the search engine in the role of compressor yielding code lengths $G(x), G(y)$ for the singleton search terms x, y being compaired and a code length $G(x,y)$ for the doubleton pair (x,y), by

$$NWD(x,y) = \frac{G(x,y) - \min(G(x), G(y))}{\max(G(x), G(y))}. \qquad (2)$$

This NWD uses the background knowledge on the web as viewed by the search engine as conditional information.

The same formula as (2) can be written in terms of frequencies of the number of pages returned on a search query as

$$NWD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)\}}{\log N - \min\{\log f(x), \log f(y)\}}, \quad (3)$$

and if $f(x), f(y) > 0$ and $f(x,y) = 0$ then $NWD(x,y) = \infty$. It is easy to see that

1) $NWD(x,y)$ is undefined for $f(x) = f(y) = 0$;
2) $NWD(x,y) = \infty$ for $f(x,y) = 0$ and either or both $f(x) > 0$ and $f(y) > 0$; and
3) $NWD(x,y) \geq 0$ otherwise.

The number N is related to the number of pages M indexed by the search engine we use. Our experimental results suggest that every reasonable (greater than any $f(x)$) value can be used for the normalizing factor N, and our results seem in general insensitive to this choice. In our software, this parameter N can be adjusted as appropriate, and we often use M for N. In the [4] we analyze the mathematical properties of NWD, and prove the universality of the search engine distribution. We show that the NWD is not a metric, in contrast to the NCD. The generic example showing the nonmetricity of semantics (and therefore the NWD) is that a man is close to a centaur, and a centaur is close to a horse, but a man is very different from a horse.

B. Question-Answer System

A typical procedure for finding an answer on the Internet consists in entering some terms regarding the question into a Web search engine and then browsing the search results in search for the answer. This is particularly inconvenient when one uses a mobile device with a slow internet connection and small display. Question-answer (QA) systems attempt to solve this problem. They allow the user to enter a question in natural language and generate an answer by searching the Web autonomously. the QA system QUANTA [15] that uses variants of the NCD and the NWD to identify the correct answer to a question out of several candidates for answers. QUANTA is remarkable in that it uses neither NCD nor NWD introduced so far, but a variation that is nevertheless based on the same theoretical principles. This variation is tuned to the particular needs of a QA system. Without going in too much detail it uses the maximal overlap of program p going from file x to file y, and program q going from file y to file x. The system QUANTA is 1.5 times better (according to generally used measures) than its competition.

II. $n > 2$

In many applications we are interested in shared information between *many* objects instead of just a pair of objects. For example, in customer reviews of gadgets, in blogs about public happenings, in newspaper articles about the same occurrence, we are interested in the most comprehensive one or the most specialized one. Thus, we want to extend the information distance measure from pairs to multiples. This approach was introduced in [10] while most of the theory is developed in [14].

Let X denote a finite list of m finite binary strings defined by $X = (x_1, \ldots, x_m)$, the constituting strings ordered length-increasing lexicographic. We use lists and not sets, since if X is a set we cannot express simply the distance from a string to itself or between strings that are all equal. Let U be the reference universal Turing machine. Given the string x_i we define the information distance to any string in X by $E_{\max}(X) = \min\{|p| : U(x_i, p, j) = x_j \text{ for all } x_i, x_j \in X\}$. It is shown in [10], Theorem 2, that

$$E_{\max}(X) = \max_{x:x\in X} K(X|x), \tag{4}$$

up to a logarithmic additive term. Define $E_{\min}(X) = \min_{x:x\in X} K(X|x)$. Theorem 3 in [10] states that for every list $X = (x_1, \ldots, x_m)$ we have

$$E_{\min}(X) \leq E_{\max}(X) \leq \min_{i:1\leq i\leq m} \sum_{x_i, x_k \in X \,\&\, k\neq i} E_{\max}(x_i, x_k), \tag{5}$$

up to a logarithmic additive term. This is not a corollary of (4) as stated in [10], but both inequalities follow from the definitions. The lefthand side is interpreted as the program length of the "most comprehensive object that contains the most information about all the others [all elements of X]," and the righthand side is interpreted as the program length of the "most specialized object that is similar to all the others."

Information distance for multiples, that is, finite lists, appears both practically and theoretically promising. The results below appear in [14]. In all cases the results imply the corresponding ones for the pairwise information distance defined as follows. The information distance in [1] between strings x_1 and x_2 is $E_{\max}(x_1, x_2) = \max\{K(x_1|x_2), K(x_2|x_1)\}$. In the [14] $E_{\max}(X) = \max_{x:x\in X} K(X|x)$. These two definitions coincide for $|X| = 2$ since $K(x,y|x) = K(y|x)$ up to an additive constant term. The reference investigate the maximal overlap of information which for $|X| = 2$ specializes to Theorem 3.4 in [1]. A corollary in [14] shows (4) and another corollary shows that the lefthand side of (5) can indeed be taken to correspond to a single program embodying the "most comprehensive object that contains the most information about all the others" as stated but not argued or proved in [10]. The reference proves metricity and universality which for $|XY| = 2$ (for metricity) and $|X| = 2$ (for universality) specialize to Theorem 4.2 in [1]; additivity; minimum overlap of information which for $|X| = 2$ specializes to Theorem 8.3.7 in [12]; and the nonmetricity of normalized information distance for lists of more than two elements and the failure of certain proposals of a normalizing factor (to achieve a normalized version). In contrast, for lists of two elements we can normalize the information distance as in Lemma V.4 and Theorem V.7 of

[8]. The definitions are of necessity new as are the proof ideas. Remarkably, the new notation and proofs for the general case are simpler than the mentioned existing proofs for the particular case of pairwise information distance.

III. Conclusion

By now applications abound. See the many references to the papers [8], [3], [4] in Google Scholar.

The methods turns out to be more-or-less robust under change of the underlying compressor-types: statistical (PPMZ), Lempel-Ziv based dictionary (gzip), block based (bzip2), or special purpose (Gencompress). Obviously the window size matters, as well as how good the compressor is. For example, PPMZ gives for mtDNA of the investigated species diagonal elements ($NCD(x,x)$) between 0.002 and 0.006. The compressor bzip2 does considerably worse, and gzip gives something in between 0.5 and 1 on the diagonal elements. Nonetheless, for texts like books gzip does fine in our experiments; the window size is sufficient and we do not use the diagonal elements. But for genomics gzip is no good.

References

[1] C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, W. Zurek, Information Distance, *IEEE Trans. Information Theory*, 44:4(1998), 1407–1423.

[2] R.L. Cilibrasi, The CompLearn Toolkit, 2003–, www.complearn.org

[3] R.L. Cilibrasi, P.M.B. Vitányi, Clustering by compression, *IEEE Trans. Information Theory*, 51:4(2005), 1523–1545.

[4] R.L. Cilibrasi, P.M.B. Vitányi, *IEEE Trans. Knowledge and Data Engineering*, 19:3(2007), 370–383.

[5] R.L. Cilibrasi, P.M.B. Vitányi, A fast quartet tree heuristic for hierarchical clustering, *Pattern Recognition*, 44 (2011) 662-677

[6] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission*, 1:1(1965), 1–7.

[7] D.B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, *Comm. ACM*, 38:11(1995),33–38.

[8] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitányi. The similarity metric, *IEEE Trans. Information Theory*, 50:12(2004), 3250- 3264.

[9] M. Li, P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*, 3nd Ed., Springer-Verlag, New York, 2008.

[10] M. Li, C. Long, B. Ma, X. Zhu, Information shared by many objects, Proc. 17th ACM Conf. Inform. Knowl. Management, 2008, 1213–1220.

[11] G.A. Miller et.al, WordNet, A Lexical Database for the English Language, Cognitive Science Lab, Princeton University.

[12] An.A. Muchnik, Conditional complexity and codes, *Theor. Comput. Sci.*, 271(2002), 97–109.

[13] S.A. Terwijn, L. Torenvliet, P.M.B. Vitányi, Nonapproximability of the Normalized Information Distance, *J. Comput. System Sciences*, 77:4(2011), 738–742.

[14] P.M.B. Vitányi, Information distance in multiples, *IEEE Trans. Inform. Theory*, 57:4(2011), 2451-2456.

[15] X. Zhang, Y. Hao, X.-Y. Zhu, M. Li, New Information Distance Measure and Its Application in Question Answering System, *J. Comput. Sci. Techn.*, 23:4(2008), 557–572.