Topological Specifications of Information

A. Haouas

Faculty of Sciences, Department of Computer Science, Oran University for Science and Technology, 31000 – Oran,

Algeria.

haouasab@yahoo.fr

Abstract

In this paper, the notion of information, focusing on its representation was dealt with. A representation that takes into account of its specificities, especially the non-quantifiable ones like semantics was designed. It seems that the most suitable and abstract model that can encompass this approach is the concept of topology. The reason was explained and the fundamental links that rely on information and topology were provided. Through this representation, some fundamental operations that treat information and situate the Hausdorff property and its role in this context were discussed. Among these operations, for instance, were sortability and track-ability. Finally, a criterion for extractability of information through the compactness property of the associated topological space was given.

Key words: *Information, semantics, qualitative aspects, topology, heuristics*.

INTRODUCTION

The mathematical theory of information is usually interested in measuring quantities related to the concept of information. It also studies its representation to handle its transmission, storage and coding (Shannon, 1948). The fields of its applications are extremely varied, due to the increasing research developed around this topic, creating an incredible synergy. The quantifiable aspect of information becomes insufficient in certain contexts like semantics, allusion, ambiguity, etc. It is then natural to try and build a theory that unifies all of the various aspects of information theory, with a view to making it less complex. One can cite telecommunications, electronics, computing, cybernetics, linguistic, psychology and biology (Rajaraman and Ullman, 2003). In each of these domains, information is viewed under a specific aspect.

For instance, in communication, we have the landmark work of C. E. Shannon who produced a theory of information in his famous paper A Mathematical Theory of Communication. This work describes Information by measuring its entropy. This theory manipulates the concept of information as a measurable content, in terms of the probability of messages. It does not pretend to deal with its qualitative or semantic aspects. Shannon showed that the information contained in a message is measurable and expressible by probabilistic tools and proved that the information 'I' of a message is computable by the formula:

$$I = log_2 1/p$$

Where p is the probability of a message to be chosen among others.

This approach nevertheless ingenious, presents a certain dependence on the "support" - space, if we want to be more abstract - where the information is specified (that is a *two*-state discrete space). Besides this, the processing space is not refined in the sense that all types of perturbations like noise and redundancy are taken into account, so that one can treat more than the pure concept of information (Shannon, 1948). Since the advent of the Internet, the notion of information has radically changed. The support now, is a virtual object, and the invariant mean to consult document is the hypertext link. Information seeking has become very challenging, due to the exponential proliferation of documents, especially in the deep web. Algorithms used in the searching engines are based on the syntax and occurrence of items. The problems encountered raise essentially from semantics. Moreover, a lot of research works have been led without a formal model of the continuous aspect of semantics, the proximity of meanings, and so on.

These considerations suggest that information as a whole and pure concept with not only its quantitative aspect but also its qualitative aspects, such as the semantic or the intuitive aspects of fuzziness and concision, should be taken into account.

Here, a form-based representation in order to formalize this view was suggested. The aim of this study is not to deal with the notion of measure. It can be clearly noticed that information is not uniformly distributed in time and space. There may exist some spots where it is true, credible or

pertinent, etc. Outside this point, information undergoes a certain loss. This study then treats the aspect of the spreading of the information and the places where it is more significant. The two main features on which the approach of this study was built were considered expressly. It seems that the most suitable and abstract model that can encompass this approach is the concept of topology.

Topology is quite a deep topic in Mathematics and presenting it here is out of the scope of this article. However, it was given in the following section some notions that will permit the reader to have a general idea.

THE BEHAVIOURAL PATTERN OF INFORMATION AND ITS LINK WITH TOPOLOGY

Some introductory notions on topology

Now, it will be appropriate to try to briefly recall what topology does involve. It consists in studying the mathematical properties that are invariant under geometric distortion, or under continuous transformation of objects.

When space is curved, stretched, twisted or generally distorted, some properties stay unchanged. Topology is interested especially in these properties. While geometry takes account of the notions that change with the form of the considered space, topology studies notions like the objects' configurations, or their general forms. It is based on the fundamental notions of continuity and limit. The main problems that topology treats are the continuous behavior of phenomena and corollary of the discontinuous one (also called catastrophe) (Thom, 1997).

The catastrophic behaviors are interesting as they enable us to classify forms such that this classification is not categorical and express the intuitive aspect of human view. This point will be expanded below.

The Information behavior and its topological Analysis

For instance, if we buckle a sheet of paper with a circle drawn on it, this circle formally changes. But how far does one consider it as just a deformed circle? This question leads to the consideration of the only properties that remain unchanged for how far. This will be referred to by the fact of non-categorical classification. It seems that there is a conservation of the notion of proximity until the catastrophic deformation.

It can be noted that there is a natural homology between the qualitative aspects of information and general topological spaces. This apparent homology is effectively a mapping between these two entities. The point where information is the most significant (the case of the perfect circle) was not only retrieved, but also its spreading aspect (the cases of deformed circles).

The set constituted by this point (the information accounting for the perfect circle) and what surrounds it (every information accounting for a deformed circle), was referred to as an open set.

The interest here, namely and firstly, in forms, is that topological space is not necessarily required to be supplied with a metric. One can invoke this property when one needs information quantification. Through this representation, some fundamental operations that treat information were discussed. Among these were, for instance, sortability and track-ability. Finally, a criterion to extractability of information through the compactness property of the space was given.

The basic links

Information is carried by some structured space and secondarily this structure is not categorical. Intuitively, upon further considerations, it can be said that the information contained in a geometric shape is conserved until a certain threshold. One can assert the fact that the notion of interiority rests invariant as well as the notion of neighborhood and that they are the only aspects that remain unchanged.

Hereafter, the essential steps encountered and that permit to link information and topology can be summarized:

i) Some information and some point in an abstract space

ii) The spreading of this information and a subset containing this point (considered in i)

iii) Information related to the first information, constitutes

a topic that can be linked to the interior of a subset

iv) The notion of proximity has to be linked to the notion

of neighborhood.

REPRESENTATION OF INFORMATION

The existence problem

Actually, the apprehension problem of the information concept in a pure and formal approach was reconsidered. It has thereby freed it from any reducing context. To realize that purpose, specifying it in an over structured environment has been avoided. By this, it is meant that the structure is neither of geometric nature nor of algebraic one (Thom, 1997). With the help of these considerations, the question of possibility for information to be represented seems therefore soluble.

In summary, answers to such questions written below can now be given:

Can information admit some representation that considers its specificities?

Which types of spaces are enough for this?

What type of structure can be conferred to these spaces?

The fit approach and the topological representation regarding the handlings

At this point, it can be said that there is now an understanding of this diffuse entity, called information.

This notion as previously seen, requires primarily a pure structural aspect reduced to a point and secondarily the notion of fuzziness, which surrounds this *structure*.

Now, another series of questions pose themselves:

What types of treatments can be supported by these structured spaces?

What properties must one design in order to allow effectiveness of such treatments?

Predicting the evolutions of stock markets using information is one of the hardest problems, for instance. It involves sets of temporal events, sets of spatial events, which are closely related, and often divergent. These facts are generally perceived intuitively and some of them are treated without measurable notion.

The diverse notions of measure in information treatments can be required in some precise situations, and they can be neglected in the contexts where the intuition takes precedence over the rest. A remarkable fact is the case where parameters evolve to give a stable state. One can cite for example, the case of the convergence towards a crash.

In the general case of convergence, it is interesting to know with what mode of convergence a system is evolving, alternatively expressed: Is convergence slow or rapid?

Other modes of convergence do exist in topology, but they will be referred to at the convenient moment.

SOME INSTANTIATIONS AND APPLICATIONS

Presented below are some examples showing how information is intuitively treated by a topological view:

1. Here is an instantiation in the actual field of computing:

Take for instance the deployment of information over Internet. The formalization can be given as follows: The element of information is in this case carried in an HTML (hypertext markup language) document that constitutes the initial element of information and the sets of hypertext links give a set of HTML pages that constitutes a neighborhood of this element.

2. The following illustration describes a more complex interpretation and can give possible applications in marketing analysis. Customers filling their shopping basket with items: Marketers lay their items physically in places according to the purchasing flow.

This example shows that a phenomenon, which occurs sequentially, has by some mental treatments of marketing analysts and finally by Marketers a consequence consisting in performing precise positioning of items on the shelves.

This behavior can be interpreted as homeomorphic mapping from the linear space of purchasing frequency of items to the show space (topologically viewed as a Space of Euclid). Let us recall that we can alternatively define a homemorphism as a linear correspondence which is bijective, continuous and whose reverse is also objective and continuous. Informally speaking, this type of mapping permits linear transformations or transports of objects without changing the notions of proximity between the elements of these objects.

3. The notion of Importance of Information topologically illustrated. One can cite the criterion of web-pages indexing used by the search engine Google called Page Rank. It is based on the recursive definition of importance, that is, a page is important if important pages link to it. Stochastic matrices are used to calculate Page Rank. Elements of these matrices are determined according to the above definition, as follows:

1. Each page i corresponds to row i and column i of the matrix.

2. If page j has no successors (links), then the ij^{th} entry is 1/n if page i is one of these n successors of page j, and 0 otherwise.

This definition emphasizes the occurrence frequency of the considered item. One can note that the statement of the definition of importance should be more closely related to topology than to the others fields of Mathematics because the notion of importance is generally used to describe something characterized by its qualitative aspect. Now, balls can be defined instead of matrices and importance and will be expressed through the concept of density. Finally, it was remarked that this approach describing information permits graphic depictions. If besides this, the representation space is supplied with a metric (or a norm), information can be concretely represented with the use of maps. This constitutes a considerable advantage especially if the current powerful visualization techniques and the priority given to graphical aids are taken into account (Fabrikant and Buttenfield, 2001).

Qualitative aspects of information

Let us recall that every topological space is characterized by its open sets. Indeed, the fineness of the topology increases with the number of the open sets defining (covering) the space.

The qualitative aspects of information can be expressed in this approach by the rank of fineness of the associated topology and what is meant by qualitative aspect is consequently the poorness or the richness of the informational space.

This space will be perceived as poor if the only open sets are the empty set and the total set itself, while its richness is expressible by the fact that there are a greater number of open sets.

The extreme case of richness is the one where every most significant point is itself an open set; i.e. the case where the space is supplied by a discrete topology. The opposite case is the one that has the rough topology.

The Hausdorff property and some of its consequences

A necessary condition for the problem of locating information

Information cannot be located if it cannot be distinguished from the rest. This problem is hence reduced to the question of distinguishability in information and the matter of finding a distinguishability criterion.

This property is easily expressible by the topological property, known as the *separation* property or the Hausdorff property (Schwartz, 1970). This forces the associated space to be a Hausdorff space. The consequences of this property are numerous, among which sortability of information expressible initially by its capability to be selected, and this is not possible without the property of Hausdorff. One can use the complete set of neighborhoods to perform inferences but it is sufficient to work with a base of neighborhoods. If this base is convergent and one has uniqueness in the convergence, it ensues that the information is track-able and the source of information is necessarily unique (intuitively, the source of information is reachable if a crosschecking is possible).

Some fundamental neighborhood properties and their consequences

Some fundamental notions frequently needed in the information treatments and admitting an immediate transposition according to our purpose are briefly presented.

Of course, it can be pretended that these processings are exhaustive, but they can be briefly referred to in order to show that this approach is globally plausible.

1. Archiving problem (or historicity if we are in space-time):

If one considers that archives consist in gathering elements of information about an initial element of information, one can express what follows.

Archiving can be approached by the transitive property of the neighborhoods i.e. every neighborhood of a neighborhood of a point (respectively a subset) is itself a neighborhood of this point (respectively a subset). This is clear if we consider that archiving consists of summarizing all about an element of information.

2. The factorization of information is evidently retrievable in the fact that every intersection of neighborhoods of a point is a neighborhood of this point.

3. An informational subset does not present redundancy if as a topological space it *admits a minimal cover.*

A criterion of extractability of information

The problem of extracting information, well known as the Data Mining, can be characterized

by the compactness property of the associated space. Indeed, it can be defined that information is extractable if one can cover the associated topological space with a finite union of open sets.

It is recalled that a subset of a topological space is said to be compact if and only if it is separated and from any open cover of this subset one can extract a finite open sub-cover to this subset (Shannon, 1948).

The statement then can be made: Information can be extracted from a set if the associated space to this set is compact.

CONCLUSION

Representation has always been the problem first in approaching a phenomenon. This stage is the most fundamental in any process of formalization. It is not only necessary for apprehending phenomena, but also for their further treatments. One can cite for instance the varied representations of numbers and the resulting numerical systems. Each representation reflects a certain degree of abstraction. Easiness of treatments clearly increases according to the power of abstraction. One can also cite the actual problem of dynamic addressing in Operating Systems. This type of addressing is made with the help of pointers. Now, pointers are very hard to manipulate, so representation using Lambda Calculus - via the beta-reduction rule - permits this with elegance.

In this paper, the issue of representing the concept of information has been addressed. This entity is subject to many interpretations depending on the vision that is obtainable. It constitutes a spectrum of visions; from the fuzzy to the concise one. The aim of this study was to unify these conceptions. These considerations naturally led to their representation according to this purpose.

Now, there exists an abstract model, precisely the topological space that seems to have the fundamental properties needed for this representation among these properties able ones and the principal mappings induced, restricting this first approach to the heuristic aspect without insisting on the formal and rigorous treatments. The mathematical and abstract construction comes then after being convinced as for the global coherence of the project.

REFERENCES

Fabrikant SI, Buttenfield BP (2001). Formalizing semantic spaces for information access. Ann. Assoc. Am. Geogr. 91(2): 263-280.

Rajaraman A, Ullman JD (2003). Querying websites using compact skeletons. J. Comput. Syst. Sci. 66(4): 809- 851.

Schwartz L (1970). Topologie générale et Analyse fonctionnelle, Hermann Paris.

Shannon CE (1948). A Mathematical Theory of Communication. Bell Sys. Tech. J. 27: 379-423, 623-656.

Thom R (1997). Stabilité structurelle et morphogenèse, Inter Edition Paris.