Journal of
Applied Sciences

A Topological Representation of Information: A Heuristic Study

A. Haouas, B. Djebbar and R. Mekki
Department of Computer Science, Faculty of Sciences,
Oran University for Science and Technology, 31000-Oran, Algeria

Abstract: The aim of the study is to design a representation of information that takes into account its non-quantifiable qualities. In this study, we then propose a topological one. This is achieved by the construction of an abstract topological model of information. We establish a mapping between information and abstract topological spaces. We then reformulate some fundamental concepts related to Information Theory such as sort-ability and track-ability. Finally, we give a criterion to extract-ability of information through the compactness property of its topological representation space.

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. This theory has been very fruitful in many fields, but only looking at the quantifiable aspect of information is insufficient for certain contexts such as semantics, allusion, ambiguity, etc. It is then natural to try and build an alternative theory that unifies the various aspects of Information Theory. One can cite Telecommunications, Computing, Cybernetics, Linguistic, Psychology and Biology (Cull, 2007). In each of these domains, Information is viewed under a specific angle. For instance, in Communication, we have the landmark work of C.E. Shannon who first wrote about the Theory of Information in his famous study, A Mathematical Theory of Communication (Segal, 2003), who describes information by measuring its entropy and manipulates it as a measurable content in terms of the probability of messages, which are represented as bit strings. In particular, it does not address the qualitative or semantic aspects at all. Shannon modelled information contained in a message as measurable quantity which is related to the probability of occurrence of the message and defined I, the information contained in a message, as:

$$I = \log_2 1/p$$

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

This approach, although ingenious, presents a certain dependence on the support (space) of information which is here the two state discrete spaces of bit strings. Besides, the space is not refined, in the sense that perturbations like noise and redundancy are not taken into account, so one cannot treat more than the pure concept of information (Thom, 1997).

Since the Internet coming, the notion of Information has radically changed (Järvelin and Wilson, 2003). The support now, is a virtual object and the invariant mean to consult document is the hypertext link. Information seeking has become a very challenge, 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 studies have been led without a formal model of the continuous aspect of semantics, the proximity of meanings and so on (Avery, 2003).

These considerations suggest to us that we should take a holistic approach to information where not only its quantitative aspect are taken into account but also its qualitative aspects such as the intuitive aspects of fuzziness and conciseness.

Since present aim is to deal with more than the measurable features only, we suggest the use of a form based representation in order to capture the now missed features of information. In a form based model, one can clearly see if information is uniformly distributed in time and space or not (Vàzquez et al., 2001). There may exist spots where it is true, credible or pertinent etc. and outside these spots information may undergo a certain

Corresponding Author: A. Haouas, Department of Computer Science, Faculty of Sciences,
Oran University for Science and Technology, 31000-Oran, Algeria
Tel: 0021341560339 Fax: 0021341560339

decay or loss. That is, this model allows us to treat the aspect of spreading of information and enables us to locate places where, it is most significant. These two main features will be the building blocks in present approach which uses the language of Topology to abstractly model information more generally.

Topology is a deep topic in mathematics and presenting it here is out of the scope of this study. However, we will now give a quick sketch that will permit the general reader to have a fair understanding of why this is the appropriate tool for our model.

THE BEHAVIOURAL PATTERN OF INFORMATION AND ITS LINK WITH TOPOLOGY

Some introductory notions on topology: Now we will 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 behaviour of phenomena and corollary the discontinuous one, also called catastrophe (Thom, 1977).

Catastrophes are interesting as they enable us to classify forms by acting as borders between classes and enable us to distinguish the non-categorical ones. This observation is at the heart of our model.

The information behaviour and its topological analysis: If we buckle a sheet of paper with a circle drawn on it, then this circle formally changes; but how far does one consider it as just a deformed circle? Answering this question leads us to consider the only properties that remain unchanged and when they cease to be preserved. We call this approach non-categorical classification: There is a conservation of the notion of proximity until the catastrophic deformation occurs.

We note that there is a natural correspondence (homology) between the qualitative aspects of information and general topological spaces. Note that the set consisting of the initial point (information accounting for the circle) and what surrounds it (information accounting for deformed circles) is an open set. This natural homology is effectively a mapping between these

two entities with which we retrieve not only the point where information is most significant (perfect circle) but also its spreading aspect (deformed circles).

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

The topological space is not necessarily metric, but if it is enriched with a metric then we can also measure and hence quantify information and hence capturing the measurable aspect of information.

Using the topological model, we shall now study some fundamental operations on information such as sort-ability and track-ability. We then give a criterion for extractability of information through the compactness property of the space.

The basic links: Information is carried by some structured space, a structure that is not categorical. Intuitively speaking, information contained in a geometric shape is conserved under deformation of space until a certain threshold deformation is hit. This means that the notions of interiority and neighbourhood are invariant and that they are the only aspects that remain unchanged. Having identified these invariants, we can now see the essential links that permit the linking of Information and Topology:

- Information 1 corresponds to a point P in an abstract topological space
- Spreading of 1 corresponds to a subset containing P
- Information related to 1 constitutes a topic that can be linked to the interior of a subset
- Proximity can be linked to the notion of neighbourhood

REPRESENTATION OF INFORMATION

The existence problem: We have formulated the abstract concept of information using minimal structure, thereby imposing very little limitation on it. Topology has allowed us to avoid specifying it in an over structured environment as in the purely geometric or algebraic models (Thom, 1997). With this approach, the question of possibility for information to be represented seems therefore soluble. In this model we should be able to answer questions like:

- Can information admit a faithful representation that reflects all of its features?
- Which types of spaces are enough for this?
- What type of structure do these spaces have?

THE FIT APPROACH AND THE TOPOLOGICAL REPRESENTATION REGARDING THE HANDLINGS

Now that we are starting to have clear understanding of the vague concept we call information, which in the model requires a pure structural aspect reducible to a point and the notion of fuzziness surrounding it, another series of questions pose themselves:

- What types of treatments can be supported by these structured spaces?
- What properties one must design in order to allow effectiveness of such treatments?

For instance predicting the evolution of stock markets using information is one of the hardest problems. It involves sets of temporal and spatial events which are closely related and often divergent. These facts are generally perceived intuitively and some of them are treated without measurable notions.

The diverse notions of measure in information processing can be required in some precise situations and they can be neglected in the contexts where the intuition takes precedence. A remarkable case is the when parameters evolve to give a stable state. One can mention for example, the case of 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 we will refer to them at the convenient moment.

Some instantiations and applications: It is presented some examples showing how information is intuitively treated by a topological view.

- **Here is an instantiation in the actual field of computing:** Let us take the example of deployment of information over Internet. The formalization can be given as follows:

The element of Information is in this case carried in an HTML document that constitutes the initial element of information and the sets of hypertext links give a set of HTML pages that constitutes a neighbourhood of this element

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.

We can interpret this behaviour as homeomorphic mapping from the linear space of purchasing frequency of items to the show space (Topologically viewed as a Space of Euclide). Let us recall that we can alternatively define a homeomorphism as a correspondence which is bijective, continuous and whose reverse is also bijective and continuous. Informally speaking, this type of mapping permits transformations or transports of objects without changing the notions of proximity between the elements of these objects.

- **The notion of importance of information topologically illustrated:** One can cite the criterion of web-pages indexing used by the searching 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 above definition, as follows:

- Each page i corresponds to row i and column i of the matrix
- If page j has no successors (links), then the ijth 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 will be expressed through the concept of density and can be therefore graphically specified (Munzner and Burchard, 1995).

Finally, we remark that this approach permits graphic depiction of information. 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 we take account of the current powerful visualisation techniques and the priority given to graphical aids (Fabrikant and Buttenfield, 2007).

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 we mean 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 is 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 with a discrete topology. The opposite case is the one that has a rough topology.

The Hausdorff property and some of its consequences: A necessary condition for the problem of locating information.

We cannot locate information if we cannot distinguish it 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, 1997). This forces the associated space to be a Hausdorff space. The consequences of this property are numerous, among which is sort-ability 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 neighbourhoods to perform inferences, but it is sufficient to work with a base of neighbourhoods (Bourbaki, 1974)

If this base is convergent and one has uniqueness in the convergence, then it ensues that 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 neighbourhood properties and their consequences: We now briefly present some fundamental notions which are frequently needed in information treatment and admit an immediate transposition according to present purpose. Of course, we do not pretend that these processings are exhaustive but we do refer briefly to them in order to show that this approach is globally plausible:

- 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 see that archiving

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

- The factorization of information is evidently retrievable in the fact that every intersection of neighbourhoods of a point is a neighbourhood of this point
- 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, we can define that information is extractable, if one can cover the associated topological space with a finite union of open sets

We recall that a subset of a topological space is said 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. We then can make the statement: Information can be extracted from a set if the associated space to this set is compact.

CONCLUSION

Representation has always been the core problem in approaching the solution of any problem. This stage is the most fundamental in any process of formalisation and mathematical study. It is not only necessary for understanding phenomena but also for their further study. In this study, we have addressed the issue of representing the concept of information with the main aim being to unify and generalize the different current models. Present abstract model, the Topological Space model, seems to have the fundamental properties needed for this global representation. At this stage, we have only heuristically checked consistency of this model with some fundamental properties. There are a lot of applications that one could examine but it is not the scope this study. Nevertheless, the evolution of future needs in semantics treatments will certainly go in this direction, due to the densification of information. One can for instance cite The Data Mining, the evolution of the Semantic Web, precisely the Web Ontology Language and so on (Plessers *et al.*, 2007).

REFERENCES

Avery, J., 2003. Information Theory and Evolution. World Scientific, ISBN: 9812384006.

Bourbaki, N., 1974. General Topology. Chapter I, 2311 Edn., Hermann, Paris, ISBN: 2-7056-1371-4, pp: 38.

Cull, P., 2007. The mathematical biophysics of Nicolas Rashevsky. Biosystems, 88: 178-184.

Fabrikant, S.I. and B.P. Buttenfield, 2007. Formalizing semantic spaces for information access. Ann. Assoc. Am. Geogr., 9: 263-280.

Järvelin, K. and T.D. Wilson, 2003. On conceptual models c and retrieval research. Inform. Res., 9: 163-163.

Munzner, T. and P. Burchard, 1995. Visualizing the structure of the world wide web in 3D hyperbolic space. Proceedings of the 1995 Symposium on Virtual Reality Modeling Language, December 14-15, San Diego, pp: 33-38.

Plessers, P., O.D. Troyer and S. Casteleyn, 2006. Understanding ontology evolution: A change detection approach. J. Web Semant., 5: 39-49.

Schwartz, L., 1970. General Topology and Functional Analysis. Edition No. 2294, Hermann, Paris, ISBN: 2-7056-5900-5, pp: 29.

Segal, J., 2003. Le Zéro et le Un. Syllepse Paris, ISBN: 2-84797046-0.

Thom, R., 1977. Structural Stability and Morphogenesis. Inter Edn., Paris, ISBN 2-7296-0065-5, pp: 123.

Vàzquez, P., M. Feixas, M. Sbert and W. Heidrich, 2001. Viewpoint selection using viewpoint entropy. Proceedings of the Vision Modeling and Visualization Conference, (VMV-01), November 21-23, Stuttgart, Germany, pp: 273-280.