Theoretical foundations of Eaagle Mapping

Introduction

The initial mapping concept was formulated in December 1991, with a mix of mathematical, philosophical and sociological foundations. None of these three foundations are more important than the other two: every individual foundation has enriched the other two and enabled the whole approach to progress.

In the wake of ever increasing sophistication in data analysis techniques, we primarily focused on creating data synthesis mechanisms to reveal knowledge and meaning from complex information sets.

This approach, aiming at providing tools to understand and leverage complexity, requires the user's involvement (Observer's Modern Theory). In this sense, we strove to part from analytical approaches seeking to describe complexity and stressing the independence of the results vs. the user (classical Theory of Objectivity).

This document is focusing on the mathematical principles that allow synthesizing large sets of heterogeneous pieces of information and representing them as a map. Should one want to link these principles to an existing theoretical frame, we would mention the Theory of Individual Preferences Aggregation, or Theory of Collective Utility.

Mathematical Foundations

Mathematical findings underlying the mapping algorithms attempt to bring answers to the following well-known problems.

The Problem

- Taking a "A" ensemble (atoms or variables) and a "*B"* ensemble (lists or individuals) of (Bi, Ri) couples where Bi is a subset of "A" and Ri an order relationship on Bi;

Is there is an "R" order relationship on "A" that does not contradict Ri relations? (Meaning that "R" restrains to Bi or Ri)

If the answer is "no", how can we comment minimize the weakening of information pieces contained in "*B"* to obtain an "R" order "R" on "A"?

The "Tree" approach brings an answer to this problem. A solution is called "map" of "*B"* pieces of information on ensemble "A".

At the end of the 18th century, when writing his mathematical social studies (e. g. in: "On how to know the wish of plurality during elections" or "Essay on the applicability of probability analysis to the probability of decisions grounded in suffrage plurality"), French mathematician Condorcet faced the issue of "building" a general, consolidated opinion from heterogeneous individual opinions. We easily understand that ensemble *B* corresponds to a collection of individual opinions and that (A, R) would be the general, consolidated opinion that we are looking for. If this problem finds easily a solution for two individuals, it creates a problem for more than two persons: this is called "Condorcet's Paradox". In 1951 Arrow in his "Social choice and individual values" proved the impossibility to solve Condorcet's Paradox without any other information.

Statistics

Condorcet understood that in order to build a general opinion from individual opinions, i.e., in order to solve conflicting opinion situations, one needs to bring in and impose an external principle: he named this external principle the "majority" principle. Interestingly, Condorcet

also finely criticized this very majority principle, stating that one should in fact multiply choice processes in order to solve his paradox. After Condorcet, other experts developed increasingly sophisticated approaches mostly based on statistics, for example the factorial correspondence, principal component, multiple correspondences, multiple factorial and
relational factorial analyses.

All these methods require proficiency in mathematical techniques. They also require an important work to "prepare" the data, for example weighting, setting groups and creating typologies of variables, individuals or modalities. Also, they require a good command of axles inertia, factor senses, cloud transition formulas and significant situations (hierarchies, Gutman effect, etc.) in order to effectively interpret the results. In any case, effective results cannot be achieved without expertise and data pre-processing and defining "a priori" a data metric of the entire data set. In this type of approach - called "a priori processing" approach -, proximity is nothing else but an application of this metric.

The Game Theory

One could also try and look in the Game Theory for the mathematical foundations of the "Trees of Knowledge". Although a "Tree of Knowledge" could be compared to a "Pareto Extremum", techniques used in the Game Theory, e.g., linear algebra, mathematical analysis, linear programming, optimization, probability, etc., are very different from that used in the Trees of Knowledge.

Furthermore, a "Tree of Knowledge" cannot be compared to a "Game Tree". Let us say that we identify each (Bi, Ri) list as a player: the situation would still not be comparable because players would not play one at a time. Instead they would all play at the same time, as do musicians in an orchestra. Let us also stress that the Trees of Knowledge theory is not related to probability approaches because by definition these approaches tend to overwrite rare events and favor the most probable events. By opposition, the Trees of Knowledge theory is specifically interested in revealing "weak" or "emerging" signals, often caused by rare yet meaningful events or pieces of information.

A unique approach - neither statistical nor probability-based

Our approaches are radically different from the above mentioned approaches. Like contemporary scientific approaches that renounce to explain phenomena by the precise trajectories of participating elements, we do not seek a solution through synthesis of causes associated with the future of each piece of information. The key founding principle of the Trees of Knowledge theory is to reveal the general, collective opinion of a pool of agents (virtual or real) who express the lists: elements of "*B"*, in an "A" space. To do so, we made the decision to create this synthesis without taking any specific (Bi, Ri) into account but by examining them all together, and by focusing our interest on the successive phases of this collective expression. This is a systemic approach where every individual event impacts the whole system. Every individual event is not a well-defined pole inextricably linked to a set network of other poles; instead, every event cannot be separated from the whole pole complex. This is the reason why any variation in the start information set will impact the whole ensemble. Thus, any change in the information system consequently provokes a recalculation of the whole synthesis. As such, we easily understand why it is vital not to ground the calculation on any prior setup, and also why it is so important that it happens in real time.

Such logic, where propositional inference does not exist and of which we do not know the mathematical equivalent, could be called "quantum logic" by assimilating each (Bi, Ri) to an "induction" possibility. In other words, we are not looking for a solution that would explain with certitude the dependency links between elements of "A", based on an analysis of *B* links. Yet, we can establish a stable solution, i.e., invariant in time and place, which defines the state of the links between subgroups of "A" - subgroups that should be as small as possible. Using the analogy of quantum physics, this "quantum logic" does not seek the effect or consequence on individual elements (particles) but rather on ensembles (paquets of particles) whose behavior can be known with precision even if we admitted the principle of incertitude on elements.

Since we work on finite ensembles, the language of expression could be – indifferently - that of hyper graphs, of topology or order structures. The solution space, because of its very existence (and not the reverse), creates a proximity that enables users to rapidly answer questions such as "What are the closest n to …?" It is a topologic space without any a priori metric, even if it is possible to establish an "a posteriori" metric (since the topology will be naturally separated).

Under this aspect and due to the very recurrent process of solution elaboration, we named this technique "recursive topology", because the solution space's topology is built recursively. This is this "recursive nature" that allows creating an IT, i.e., software solution.

Optimization principle

In fact, what we are really looking for by elaborating this topologic space is to create a shape whose meaning would best match the entire information set present in *B*. In other words, the projection of every Bi in the solution space should enable users to effectively find the pieces of information contained in (Bi, Ri). This is the very principle that the algorithm should respect in order to generate the solution space.

By refusing any metric or weighting as a prerequisite to solution building, the "maps" theory enables users to react very fast to any change in the start information set. In some ways, the objective of the representation is to respect the following principle: "near" opinions, i.e., little diverging, should have "near" representations topology-wise in the solution space. This proximity notion is more powerful than the usually admitted: indeed, it enables to get an idea of the relationship between two expressions by integrating the relationships between all the other expressions, and not based on the sole distance between these two expressions.

A specific solution

In the specific cases where Ri relationships are total orders, recursive generation of the parts of A, (A_k) (such as each Ak could generate, thanks to the (Bi, Ri), the remaining part of A non present in the A_n, for $n<k$), is at the heart of the algorithm that enables the building of the solution space. It is evident that to respect the optimization principle, the A_k have to be as numerous as possible and therefore be minimal (for example, if there was no contradiction between the Ri orders, the A_K parts would each boil down to one single element, and the solution space would be topologically equivalent to a stick: an ensemble of dots entirely sequenced).

Original theorems prove that if the Bi cover A, then the A_K constitute a partition of A, the links between the A_K related components create an arborescence, the (Bi, Ri) imply a fine structure on the A_K, all the pieces of information can be displayed as an image, topologically equivalent to a tree.

Comments

Therefore, the tree structure is not a goal conditioning information processing through algorithms. Instead, it results from the work of the algorithm that seeks to build a shape that will least contradict specific trivial shapes (sticks) implied by the (Bi, Ri). In no case whatsoever does the tree shape determine the work of the algorithm. Furthermore, this shape does not imply that there exists a tree-like structure does exist between the elements of A (not to be mistaken with the A_k) because specific A_K^j (see next definition page) can represent several elements of A that are not structured in the arborescence.

Other theorems help reduce the complexity of A_K calculations to the linearity pending on the cardinal of *B*, and in n*log(n), n being the cardinal of A. An algorithm of a totally different type manages the positions of every A element in order to generate the display. The similar formula also helps relocate in real time every element of A when there is an information system change on A or B.

Of course, no statistic element does participate the display's structuring. But it is possible to express quantitative data: every element in the display can take a specific color matching a quantitative occurrence.

To conclude, we would like to stress that the existence of different types of trees and more generally of different types of "maps", leads to the systematic study of a structure on all the maps on which operations can be defined: sum, difference, duality, etc. For example, the rich concept of duality, omnipresent in data analysis, finds here all its wealth and flexibility because links between variables (elements of A) and individuals (elements of *B*) can perfectly be reversed. For example, we can create a map of <u>both</u> products about which a given number of people have expressed their preference, <u>and</u> a map of individuals who expressed their preference about a given number of products.

A bit of technique

Let us say that A and *B* are (B_i, R_i) data. We are looking to create A_{ks} forming a partition of A. Links implied by the elements of *B*, restrained to every A_k imply in A_{ks} a connectivity structure

enabling us to define A_k^j related components. All the A_k^j create a partition of A that gets structured as a related arborescence.

The more complex task is identifying the A_k. This is done recurrently (A_{k+1} is calculated like A_k, after having deleted the elements of A_k from the (B_i^{K-1}, R_i) lists; in turn, these lists turn into (B_i^K, R_i)) until there are elements of de A left. The principle is thus fully unveiled when one understood how to create A_1.

Definition of A_1

Let A^1 be the ensemble of the first elements in (B_i, R_i) lists.

We say that the X ensemble "generates" the Y ensemble thanks to the (B_i, R_i) lists if the elements of Y follow the elements of X in the lists that contain the elements of X.

A_1 is the smallest subgroup of A^1 that generates A^1. In the exceptional case where there would be several candidates, one will activate a decision process on the number of lists then on the number of elements of A present in the lists.

Comments

1-If the theory may sound simple, the calculation technique is less simple because it is complex to determine a "smaller ensemble" while trying to optimize a solution. This is one of the key proprietary features of the Eaagle software solution.

2-As we can see, the determination of A_1 does use any statistic type of information on the lists or variables: consequently, nothing opposes to the fact that very rare elements appear right at the bottom of the tree. This situation would happen if necessary in order to faithfully reflect the information contained in B.

3-Yet, the statistic importance of an element will increase the chances that this element appears at the bottom of the tree because a high level of frequency might cause a rich "procreation", a large diversity of the lists in which the element appears. If it were not the case, the statistic distribution would not reveal a great information wealth (Shannon): as a result, it is logical that this element be not privileged.

4- If statistic information does not participate to the creation of the tree, it does not mean that statistics are totally absent. All statistic pieces of information are directly accessible and visible through the element color code. The map analogy works well here: the shape of a map does not does tell us anything about the territory (-color does-); instead, the map's shape

informs us on the relationship with a "general" level, usually the sea level. In terms of the trees, the analog level will be the level of constraints on the structuring.

5- As we can see the algorithm subjects information to 2 phases of structuring:
   a) Identification of the A_k, through the "generation" principles
   b) Identification of the A_k^j through the connexity principle

If information was either very repetitive or totally contradictory, these two principles applied in all mathematical "purity" might very well reveal trivial or very varied structures. In this case it would be entirely possible to weaken the constraints imposed by either one principle by making their defining parameters vary (and obtain a more "readable" shape and hence to reveal more intelligible meaning). Yet, this is obtained through a more important incertitude margin on the way the algorithm "respects" processed information. This incertitude margin is digitized, and we can then assert that the tools to adjust the algorithm enable us to measure the error delta that exists between a shape that gives sense to a set of information and this information set itself.

The notion of incertitude is fundamental to understand the scientific approach and the practical application of the "Trees of Knowledge". Indeed, the constraint to obtain a nontrivial solution is that all or part of the elements of information (elements of A) be shared by specific Bi. A totally precise identification of an element linked to a Bi would lead us to conceive them all as different (Leibniz's principle of the imperceptibles). The belief that knowledge can be extracted for an information set implies the possibility to create sense or meaning from this information set, hence to admit the sharing of specific elements of A, hence to accept a principle of incertitude the elements of A's determination features. To illustrate, we can create sense or meaning on skills only if we admit that the exact identification of a person's skills has no sense!

Benefits of our approach:

Work on data preparation is minimal because it is not necessary to pre-process, pre- format or weight the variables.

The individual/variable duality is very easy to activate since the absence of weighting eliminates the need for transposition calculations.

Contextualization is not an issue since it is - by definition - the space in which the problem is solved. Indeed, there is no need to predefine the space in order to create the synthesis mapping: as a result, there is no need to create a universal model or matrix on which to position specific axles. The variables topological space is created by the interactions between all the elements or individuals. As such it does not pre-exist any of them and it changes with each variation. It also means that the notion of independent distance of two elements with regard to the other elements does not make sense. One more, contextualization is based on given up the idea of a predefined, universal model. Space and individual define
and mutually influence each other (it sounds like relativity structurally speaking).

Total absence of pre-existing metric allows weak signals (i.e., statistically rare yet structurally coherent elements) to appear and not be swallowed by very frequently repeated
signals. As a result, very rare phenomena can clearly appear on the map.

The existence of a topologic, non-metric space does not prevent the emergence of a metric from this very space, and proximity calculations are of course possible. Similarly, the space will be able to support and reveal all and any statistical information on the variables, specifically every related to usage. It is then possible to integrate in real time the effects of the Information System's leverage by users. This is what allows measuring involvement and
integrating the user in the very system.

Speed of synthesis visualization of an Information System (from 5 to 10 seconds for a 1 Mo IS on a Pentium 90) enables users to simulate changes and assess impact of the changes on the IS. Thus, it is possible to simulate transformations in minutes and assess their impact on the whole information set distribution. The tool's speed of reaction to any changes is what allows mastering and managing complexity; without speed, we could not react fast enough to
ever changing evolution of our surrounding reality.

It is clear that this performance level is obtained thanks to computer systems. Yet, it is more clearly due to the extreme simplicity of incoming information structure and to the reduced complexity of the algorithm.