SEMIOTIC TOOLS FOR ECONOMIC MODEL BUILDING

Ana Marostica
Facultad de Economía
Universidad de Buenos Aires

Fernando Tohmé
Departmento de Economía
Universidad Nacional del Sur
CONICET

ARGENTINA

e-mail: amarost@econ.uba.ar
amarost@infovia.com.ar

e-mail: ftohme@criba.edu.ar

Abstract:

Scientific researchers, when faced with real world data, try to detect the hidden relations and laws that are not readily apparent. This is the basic motivation for what is called "model building". Several techniques were developed in order to facilitate that work. Statistics provided ways to build efficient models by using the minimal amount of contextual information. Computational intelligence continued with this trend of automating the construction of models for restricted domains. In this paper instead, we claim that model building requires the active participation of researchers and their previous knowledge and intuitions. Contextual information must be taken into account to faithfully represent the real world phenomena. To improve this task, we need more expressive instruments. Semiotics, a discipline highly concerned with iconic reasoning tools, is the basis on which we will build the desired procedures for model building. The method we introduce here, with its great expressiveness, is extremely useful for economic model building. This is because in Economics (especially in Microeconomics) the heterogeneity of data and the different statistical methods applied lead to very different models. With more expressive methods, these differences will disappear or at least will become easy to detect where they come from. The method that will be applied here is a kind of semiotic data-mining technique for generating models. This semiotic engineering will be applied to the analysis of the degree of convergence among economies. This last issue has been a source of discussion for economic growth theory in the last years. Since several factors are involved, it seems clear that more than a mere statistical analysis is required to detect the relations between sources of growth and the rate of growth. A semiotic approach will be useful on this issue.

I.- Introduction

Science can be conceived, in a wide sense, as the activity of searching for *patterns* in the real world **[Hanson 1961]**. Although there is little doubt about this, there exist vast disagreements between epistemologists about how these patterns are discovered and about how to assess the evidence for scientific claims. Since the pioneering work of Thomas Kuhn, a consensus grew on the importance of the social aspects of the scientific enterprise on the final acceptance (or rejection) of theoretical constructs. How influential these aspects are is still

under debate, but for us the important realization is that the scientist is idiosyncratic, meaning that his own abilities, techniques, even prejudices matter. Of course, all these elements must lead to a body of results that still can be inter-subjectively assessed. In the long-run the systematic application of the scientific method must lead to that inter-subjective assessment. But it is also true that, in a short run, no theory or model can be completely freed from the idiosyncratic features introduced by its developer. The personal hallmark of a scientist in his theories and models is given mostly by his mistakes and errors. This is remarkably so in the use of statistical methods, where their blind application leads to a number of problems (some of which we note in the following). The exchange among scientists and their interaction with the real world leads to a self-correcting process that cleans the theory out of those mistakes in the long-run. Our approach applies, therefore, only to the task of scientific research in the short term.

The foregoing observations show that rather insurmountable difficulties exist for the automation of the task of science making, unless the automatic process is able to be idiosyncratic by itself (or at least it mimics the characteristic thought processes of its developer). The difficulties associated with the detection of patterns in reality show that it is hardly a matter of automatic curve fitting. This is well known by statisticians. As Savage and H. Simon (**[Simon 1968]**) put it bluntly, an approximate generalization is, according to any statistical test, indistinguishable from the form of a wrong generalization. An inductive inference, of course, would check, first, the data at hand, before making any hasty generalization. The point is that qualitative evidence is not easily translatable into quantitative forms that can be statistically supported.

Other (non-statistical) methods lead to similar problems. Even computational intelligence can only provide rough approximations to the task of theory or model building. In fact, systems like BACON (in any of its numerous incarnations) despite their claimed successes, are only able to provide *phenomenological* laws. That is, they are unable to do more than to allow for generalizations that involve only observable variables and constants. No deeper explanations can be expected to ensue from their utilization.

Our approach is based on a change of the focus of analysis. Instead of trying to design a fully automated artificial scientist, we think it is better to develop a more efficient human-machine protocol of exchange, in order to facilitate the task of finding patterns in reams of data.

We advocate here for a design founded on *semiotic data mining*. This technique, part of an incipient but rapidly growing new semiotic engineering[1], would provide the rules for a reasonable shared task of model building.

In the man-machine interaction that we try to regulate, the human side has a crucial task, not yet fully developed in the literature, consisting of the formation of concepts and the elicitation of qualitative relations. In fact, the human mind seems still to be unbeatable in the game of detecting patterns in disordered and noisy data. Of course, as it is well known in Combinatorics (as a result of Ramsey's theorem **[Graham et al. 1990]**), with enough elements many patterns can be found, some meaningful, others just casual. In any case, only the idiosyncratic aspects of the scientist's mind can make sense of the variety of possible structures that can be found.

The aim of this paper is to introduce the foundations of what we call *semiotic data mining(S.D.M.)*. We will give, as well, a gross overview of an architecture providing a flexible set of tools, expressive enough to allow an expert in a field to build models and theories in an interaction with databases.

This paper is organized as follows. In the next section we will present an exhaustive discussion on what S.D.M. is. In section 3 we will present the architecture that provides its implementation. In section 4 we will develop an example of its utilization in the context of economic growth theory. Finally, in section 5 we present our conclusions and prospects for further work.

II. Semiotic Data Mining: Foundations

The expression *data mining* represents any method to extract relations and patterns out of raw data. Any such method helps to generate models that represent the structure implicit in any amorphous database. Most data mining methods are based on the utilization of statistical methods, which accordingly , obtain probabilistic descriptions of the information at hand. Other methods are based on the utilization of neural nets, genetic algorithms or evolutionary programs. In any of those cases, similarly, a statistical characterization obtains. Symbolic methods allow to obtain rules that summarize implicit information.

Semiotic data mining is an alternative method, based on the application of semiotic methods to the task of extracting patterns from databases. To understand how it works we have

[1] Semiotic engineering means the application of semiotics to the design and construction of information systems. It emphasizes on expressiveness and content of the concepts involved. This top-down engineering helps to solve problems that quantitative methods does not solve.

first to characterize how semiotics treats data. In that sense, a datum has, beyond its obvious information-theoretic properties (which are relevant for the problem of efficient storage and transmission) a semiotic import already discussed by Peirce in his classifications of signs **[Maróstica 1997]**. More generally, according to this point of view, every set of data is a sign, which therefore can be classified according to Peirce's exhaustive taxonomy. The advantage of this approach is that there exists only an initial finite set of possibilities to match with the real world information. Once one of the possibilities is chosen it provides a neat statement of the kind of structure hidden in the data, although not necessarily a functional form. Even if the relation between data and structure, which we call an **analogue**, can be of different types we restrict our attention to those that can be expressed in a first-order language or *formal analogies*. This responds to the obvious fact that first-order logic constitutes the best known formalism in Computer Science. In that framework we will introduce our basic notions. Following the fundamental parts from**[Maróstica-Tohmé 1999] relevant here**:

Definition 1: given a first order language Γ a **structure** is $D = (|D|, N, F, G)$, where $|D|$ is a set of individuals; N is a function that assigns an individual to each constant of Γ; F is a family of endomorphic functions on $|D|$; while G is a set of predicates on $|D|$. An **interpretation** of any consistent set of well formed formulas of Γ, $T(\Gamma)$ obtains through a correspondence that assigns constants, function symbols and predicate symbols to constants, functions and relations in D. A **model** for $T(\Gamma)$ is an interpretation where every interpreted formula is true. The relationship between the set of well-formed formulas of G and the elements of the structure D is a **formal analogy**.

In words: a structure is no more than a database plus the relations and functions that are, implicit or explicitly, true in it. An interpretation is a structure associated to a certain set of well-formed formulas (when deductively closed this set is called a **theory**). If all formulas are true in the interpretation this structure is called a model. Since any scientific statement can be conceived as a mathematical expression, it seems that the previous definition of a structure is enough for our purposes. Then:

Definition 2: given a set of feasible structures $\{D^M{}_i\}_{i \in I}$ where I is a set of indexes, selected for verifying a set of criteria M, an **abduction** is the choice of one of them, say D*, by comparison with the available information.

This choice is not arbitrary. It is intended to find the best **explanatory** structure. The criteria represent all the elements that the scientist wants to find incorporated into the chosen structure. Given the criteria in M, the set of structures that verifies them is defined as follows:

Definition 3: a criterion m_j defines a set of structures in which it is verified, $\{D^j{}_i\}_{i \in Ij}$ (where Ij is a set of indexes corresponding to this criterion) . Then, $M=\{m_j\}_{j \in J}$ defines a set $\{D^M{}_i\}_{i \in I} = \cap_{j \in J} \{D^j{}_i\}_{i \in Ij}$.

The comparison of the structures with the data determines an order on $\{D^M{}_i\}_{i \in I}$:

Definition 4: given the database (a finite set of grounded formulas) D, and two possible structures D_j, D_l we say that $D_j \leq D_l$ if and only if $WFF(D_j) \cap D \subseteq WFF(D_l) \cap D$, where $WFF(\cdot)$ is the set of well-formed formulas of a given structure.

There exists a maximal element in the ordered set $\langle \{D^M_i\}_{i \in I}, \leq \rangle$:

Proposition 1: there exists a maximal structure D^* in the set $\{D^M_i\}_{i \in I}$ ordered under $\leq$.

Sufficient conditions for uniqueness of the chosen structure depend on the inclusion in M of certain methodological criteria. One of them is a *minimality* criterion:

Definition 5: a criterion of minimality m^{min} is such that given two structures D_i, D_l, where $WFF(D_i) \subseteq WFF(D_l)$ and $WFF(D_l) \not\subset WFF(D_i)$, it selects D_i.

Another criterion that could be included in M is one of *completeness with respect to the database*:

Definition 6: a criterion of completeness with respect to the database m^{com} is such that given two structures D_i, D_l, where $D \subseteq WFF(D_l)$ but $D \not\subset WFF(D_i)$, it selects D_l.

Then we have the following result:

Proposition 2: if $M=\{m^{min}, m^{com}\}$ and the set of possible structures is otherwise unrestricted, the D^* is unique.

This result shows that a unique structure can be selected if the restrictions on possible structures obey only to methodological criteria like minimality and completeness. This result is at the same time deep and irrelevant for the goal of mining the database: if the only well formed formulas in the chosen structure are the ones drawn from the database it is not possible to provide more than a description (data fitting) of the available information. That means that if only methodological criteria are to be used, the result of the inference is the generation of a prototype, i.e. only a statistical inference is performed.

The last point makes it clear that specific criteria are required, and that their selection is matter of taste. The criteria selected represent the idiosyncratic characteristics of the model builder. On the other hand, it is clear that most of the scientific work is devoted to apply them to new sets of data. A scientist proceeds by building a skeleton for a structure (according to his criteria) and using the database to adjust the free parameters. Data is fit into the different structural skeletons and the results are compared with the hypothetical structures. Only those

accepted are candidates for further work. In other words: normal scientists only work with *familiar* problems.

As said, data fitting involves a certain form of analogical reasoning since only known structures are accepted. But this framework is not rich enough to handle a wider class of analogies. To deal with this objection we will extend our ordered set of structures $(\{ D^M_i \}_{i \in I}, \leq)$ to a more general ordering $\langle \{ D^M_i \}_{i \in I}, \Re \rangle$, where $\Re$ symbolizes a general binary relation. This general (analogical) relation may encompass all the relevant comparisons among structures, by their relative degrees of similarity in order to choose a preferred structure compatible with the database.

Notice that, if we assume that $\Re$ is a reflexive and transitive relation, Proposition 1 can be conveniently recast:

Proposition 1': there exists a maximal structure D^* in the set $\{ D^M_i \}_{i \in I}$ ordered under $\Re$.

This result allows us to preserve the conceptual framework of our previous section, including some derived results that are amenable to transformation into procedures. But before going into that, let us discuss the nature of $\Re$.

As said, $\Re$ represents a general analogical relation among structures. A representation is analog to its object if it has the following characteristics (a modification of **[Levesque 1984])**:

- For every element of interest in the real object there is a sign in the representation.
- For every simple relationship in the object there is a connection among signs.
- There exist one-to-one correspondences between relationships and connections, on one hand, and between elements and signs, on the other.

That is: a representation is analog to its object if it constitutes a kind of "picture" of the object. In the case of "objects" like states of real world systems, representations assume the form of abstract constructs that can be easily embedded in first-order structures **[Myers-Konolige 1996]**.

In the last case, the real world data can not just be reduced to grounded formulas, but also must include further descriptive structure. If, for example, we are dealing with information about the economic status of emerging countries, it is not enough to include in the database statements like "Argentina's annual growth rate in 1998 was 4% ". Expressions with universally quantified variables like "Countries with high levels of technical literacy have higher rates of growth" must also be explicitly included.

Once this point is made clear, we have to assume that we are not longer dealing with plain databases (like D) but with so called knowledge bases **[Russell-Norvig 1994]**. A knowledge base $K = \langle K^{g}, K^{u} \rangle$ consists of a set of well-formed formulas in a first order language where K^{g} is the set of grounded formulas (i.e. only involving constants) while K^{u} is the set of formulas with universally quantified variables.[2]

The analogical order $\Re$ represents any kind of relation between structures. It may be used to formalize the notion of similarity among theoretical constructs. Given the set of structures $\{ D^{M}{}_{i} \}_{i \in I}$:

Definition 7: given the knowledge base K, and two possible structures D_j, D_i we say that $D_j \, \Re \, D_i$ if and only if $D_j \cap K \subseteq D_i \cap K$.

Notice that when $K = D$ (i.e. when all the formulas in the knowledge base are grounded) it follows that $\Re \equiv \leq$. In other words, $\Re$ extends $\leq$. Moreover, Proposition 1' can be proven in the same way as Proposition 1 since $\Re$ preserves the property of being reflexive and transitive.

This line of reasoning assumes an **extensive** representation of the information an economic model builder faces. That is, that $\langle \{ D^{M}{}_{i} \}_{i \in I}, \Re \rangle$ is entirely available. But this is a strong assumption. An expert may know $\{ D^{M}{}_{i} \}_{i \in I}$ but the ordering depends on the knowledge base. Therefore, he has to search exhaustively until he finds the maximal structure (whose existence is ensured by Proposition 1').

The process of searching for the closest analogue to a given knowledge base responds to what can be called a **tychist / synechist** conception of logic. That is, one in which the signs that represent objects are only provisional since for every representation there exist facts that cannot be represented by it. In consequence, those signs have to change in time to adapt better to the real world. **Tychism** (which comes from the Greek word for "change") is Peirce´s idea that given any structure there exists always a meaningful fact that is not represented in it – a sort of dynamic doctrine of incompleteness-. **Synechism** is its dual notion: it states that for any new fact there always exists a structure that is able to encompass it **[Marostica 1997]**. We can describe this process of successive adjustments as follows:

Procedure 1

- Step 1: **Input** K (the knowledge base)

[2] Formulas with existentially quantified variables can be either grounded (eliminating the quantifier and replacing the free variable with a constant) or transformed into one with universally quantified variables using the equivalence $\exists x \, P(x) \equiv \neg \forall x \, \neg P(x)$.

- Step 2 : **Input** $MIN(K) \subseteq K$
- Step 3: **Input** $D^0 \in \{D^M{}_i\}_{i \in I}$ (the initial structure)
- Step 4: i:= 0 (iteration)
- Step 5: **If** $MIN(K) \not\subset D^i$ **then**
- **Choose(** D^j) (a model that fits better)
- i:= j **Goto** Step 4

  Else $D^* := D^i$ (accept as the better fit) ♦

Choose can have different specifications. But its behavior must be the following:

$$\textbf{Choose(}\ D^j) := \textbf{Select}\left(D^j \in \{D^M{}_i\}_{i \in I} \mid D^j \neq D^i, D^i \supseteq MIN(K)\right)$$

Where **Select**$(x \in X \mid C_1 \ldots C_n)$ is a procedure that chooses an element x out from a set X, such that x obeys a set of conditions $C_1 \ldots C_n$.

In order to ensure that the procedure tends to the maximal structure, it suffices to show that it is increasing. This is shown in the following:

Proposition 3: for every stage i of Procedure 1, $D^i \ \Re\ D^{i+1}$.

The key piece in this process is **Select**. To ensure the successful termination of the cycle it is necessary to give at least a sketch of how it works. The basic idea is that, given the description of a situation, several candidate analogues are generated (using the available resources) and then their "strengths" are compared. The strongest, representing the closest match is chosen. In formal terms, **Select** chooses a maximal element, according $\Re$.

Procedure 1 could be implemented. To do so in a useful way, a rich variety of structures is needed. That is, a **library** of structures should be compiled before making this process automatic. But the details of construction of such a library are far from trivial. In the first place, first-order languages cannot be used in their full generality since the evaluation of an expression like $MIN(K) \not\subset D^i$ may not be decidable. That is, there may not exist mechanical procedures to check out that relation. This is because an unrestricted first-order language allows infinite expressions while computations require finite time.

Typically, analogical structures use as data structures labeled diagrams like trees or graphs. Although this is rich enough for certain problems, it is apparently at the cost of leaving out of consideration lots of relevant contexts. Fortunately, we can show that this not quite so, since first-order structures and knowledge bases can be put in an equivalent form involving monadic, dyadic and triadic relationships, which are easily represented as diagrams.

The basis of our argument is the formal proof of a conjecture advanced by Peirce, known as the **Reduction Thesis**, which roughly says that from relations of arities 1, 2, and 3 exclusively, **all** relations (of all non-negative integers) may be constructed. Some previous definitions are in order. We follow here again **[Marostica- Tohme 1999]**:

Definition 8: Let (X, α) be a relational structure of arity n onto a base-set X, and (X, β) a relational structure of arity m on the same base–set. We define the **relative product** of (X, α) and (X, β), denoted by $(X, \alpha * \beta)$, as a relational structure of arity $n + m - 2$. It is the following subset of the Cartesian–product set of X^{n+m-2}:

$$\alpha * \beta = \{(x_1, x_2, \ldots, x_{n-1}, y_2, \ldots, y_m) \mid x_i, y_j \in X, i=1..n-1; j=2..m\}$$

such that there exists $u \in X$ verifying that $(x_1, x_2, \ldots, x_{n-1}, u) \in \alpha$, and $(u, y_2, \ldots, y_m) \in \beta$.

This construction can be recursively extended to relational structures built up from any number of basic relational structures. More precisely, from a collection $\{ (X, \alpha_i) \}_{i \in I}$ a relational structure $(X, \prod_{i \in I} \alpha_i)$ can be constructed, where, if the arity of each α_i is m_i, $\prod_{i \in I} \alpha_i$ is a subset of $X^{\sum m_i - 2(I-1)}$. Then, an inverse operation can be defined:

Definition 9: A relational structure (X, α) can be **relatively decomposed** onto a set $K = \{(X, \alpha_i)\}_{i \in I}$ of relational structures if (X, α) is the relative product of the relational structures in K.

This definition allows us to characterize the relational structures that can be decomposed into structures of arities less or equal than 3:

Definition 10: A relational structure (X, α) of arity n is **relatively reducible** onto X if and only if (X, α) can be relatively decomposed onto a set $K = \{ (X, \alpha_i)\}_{i \in I}$ of relational structures of X of arity smaller or equal than n.

Now we can establish that **every** relational structure of arity $n > 3$ is relatively reducible within any domain:

Theorem 1 (Reduction): Let (X, α) be a relational structure of arity $n > 3$. We restrict our attention to $X_0 = X \cap \alpha$. Therefore the relational structure (X_0, α) is of arity n (it is defined onto X by the n-tuples of α). Then (X_0, α) is relatively reducible onto a set of relational structures of a arity $n \leq 3$.

A relational structure (X, α) of arity n can be considered as a simple structure of information. On the other hand, we want to show that the converse is also true. It suffices to see that every closed well-formed formula can be seen as a relational structure. To do so, recall that we restrict our attention to either universally quantified or grounded predicates (this last convention is in order to facilitate reasoning with existential formulas). For each universally quantified predicate used in a first-order structure (D or K), say f, of arity n we associate a relational structure (X, α_f) where X is the set of constants which are potential arguments for all the predicates in the language, while $\alpha_f \subseteq X^n$ is such that for every $(x_1,...,x_n) \in \alpha$, $f(x_1,...,x_n)$ is true. On the other hand, for a grounded predicate of arity n, say $g(a_1,...,a_n)$, just define (X, α_g) in the same way, with the proviso that $\alpha_g \subseteq X^n$ contains a single element, $(a_1,...,a_n)$.

Then, each piece of information in a first-order structure can be reduced to a set of relational structures, which consequently, by means of the reduction theorem, can be represented as relational structures of arities 1, 2 and 3. Graphical representations are easily constructed for these kinds of relational structures by means of the so-called **entity-relationship diagrams** (E-R diagrams). They are hyper-graphs, i.e. each relational structure $\Lambda=\langle N, E\rangle$ consists of a set of nodes (N) while $E \subseteq 2^N$ is the set of edges that connect any subset of nodes (edges of trees only connect two nodes). In our case, nodes represent entities, i.e. either generic variables or constants of the language. Diamond-shaped boxes represent the edges, i.e. the relationships between entities. It is possible to represent Peircean diagrams, (i.e., (a) monadic diagrams or 1-diagrams, (b) dyadic diagrams or 2-diagrams, and (c) triadic diagrams or 3-diagrams), which are depicted in Figure 1.

Any first-order structure can be represented using these diagrams, connected between them, representing the logical (or in a wider sense, *semiotic*) relations among statements. Each structure then, can be compiled as a hyper-graph, which is easily storable in a library of forms on which Procedure 1 can be performed. Since all these hyper-graphs are assumed to be finite (since in practice either $|D|$ and $|K|$ are finite) a search for the best match ends in finite time. The procedure is as follows:[3]

Procedure 2

- Step 1: **Input** D (a given structure)
- Step 2: $d^*:=\langle D^u, D^g\rangle$
  (predicates are partitioned into those that include universally quantified variables and those grounded)
- Step 3: **For** $f \in D^u$
  $H:= H \cup(X, \alpha_f)$ (form a set of relational structures equivalent to the predicates with universally quantified variables)
- Step 4: **For** $g \in D^g$
  $H:= H \cup(X, \alpha_g)$ (add the grounded predicates)
- Step 5: **For** $(X, \alpha) \in H$
  $H^*:= H^* \cup \{(X, \alpha_i)\}$ (form a set of relational structures of arities 1, 2 or 3 applying the reduction theorem)
- Step 6: **Output**(H^*) (i.e. the hyper-graph corresponding to D) ♦

[3] Note that H^* constitutes a data structure formally equivalent to a hyper-graph.

III.- The Architectonic Features of S.D.M.

The previous section introduced our main ideas about a qualitative method for the detection of relations in data and knowledge bases. Since this is a semiotic approach it must be complemented by means of the pragmatic maxim advanced by Peirce in 1878 (**[Peirce 1960]**). It commands that to determine the exact meaning of scientific concepts we have to specify their experimental consequences. In other terms this means that for each theoretical approach a procedure to make it operative must be given. In our case it means that S.D.M. needs to be fully implemented by a protocol.

As we have seen, the first-order structure that applies to certain context is selected because its structure fits well to the state of affairs in that context. As we discussed previously, there exists always a graphical representation for the internal relationships among statements. On the other hand, concepts can also be organized, this time hierarchically, according to their respective semiotic relations. This organization remains the same for all the cases in which its structure is applied.

This hierarchical ordering of concepts is based on the idea that some of them depend on others. Charles Peirce's idea of semiotic trichotomies provides a foundation for the construction of this ordering. First of all note that in this formulation, each piece of information is considered to be a *sign* that has an object and a meaning. Each sign is assumed to have different relationships to the sign itself, to its object and to its meaning. Under the first relationship we have the categories of *qualisign*, *sinsign* and *legisign*; under the second we have *icon*, *index* and *symbol*. Finally, under the third, we have *rheme*, *dicent sign* and *argument.*

No matter the exact meaning of those classifications, it is clear that they reflect Peirce's idea (already seen at work in the Reduction theorem) that three levels are enough to categorize any type of information. Each kind of relationship is either of a first, second or third level (meaning three different levels of abstraction). Moreover a first determines only a first; a second determines a second or a first, while a third determines either level. This determination is expressive enough to allow representing not only classifications but also empirical determinations (see **[Marostica 1997]**).

Since we showed that any first-order structure or knowledge base could be reduced to more or less involved graph-theoretic representations, it suffices to consider the relationships between signs and their objects. In other words, we classify items in the diagrammatic representations as being *for* some real world objects.

As said, the classification of signs according to their objects distinguishes three types of signs: icons, indexes and symbols. Icons are the closest (the real graphic) representations

of their objects, while indexes summarize information about several cases (that is why inductive reasoning is indexical **[Maróstica 1998]**). Symbols, of course, stand for the highest degree of abstraction and facilitate formal reasoning.

In each field of knowledge concepts can be put in any of these three categories. In our framework this means that every constant or variable can be labeled according to its nature. This translates into the well-formed formulas since those that have symbols as their arguments inherit the "thirdness" from their arguments. The same is of course true for formulas whose arguments are indexes and icons. If a formula involves disparate different kinds of signs, it inherits the category of its argument of lowest category.

With this characterization at hand, we can see how the hyper-graphs can be transformed into trees. Algorithmically:

Procedure 3

- Step 1: **Input** H (a given hyper-graph)
- Step 2: $h^* := \langle h^e, h^r \rangle$ (elements of the hyper-graph are partitioned into entities and relationships)
- Step 3: **For** $e \in h^e$
  $H^e := H^e \cup \langle e, \lambda \rangle$ (form a set of labeled entities, attaching to each entity its corresponding label –icon, index or symbol)
- Step 4: **For** $r \in h^r$
  $H^r := H^r \cup \langle r, \lambda \rangle$ (form a set of labeled relationships, attaching to each relationship the lowest label corresponding to the entities that it relates)
- Step 6: **Output**($H := \langle h^e, h^r \rangle$) (i.e. the hyper-tree corresponding to h) ♦

In fact, Procedure 1 can be conveniently rewritten to handle, instead of first-order structures, hyper-trees (i.e. hyper-graphs without cycles). Notice that now the rather mysterious requirement of matching the minimal part of a structure is transformed into matching a significant part of a hyper-tree. By construction this means an "upper" fragment, i.e. a sub-hyper-tree where entities and relationships have a high-valued label. In other words:

Procedure 1'

- Step 1: **Input** $H^0 \in \{H^M{}_i\}_{i \in I}$ (the initial hyper-tree)
- Step 2: **Input** H (the hyper-tree representing the knowledge base)
- Step 3: $i := 0$ (iteration)
- Step 4: **Input** $MIN(H^i) \subseteq H^i$
  (the upper fragment of the hyper-tree that suffices to accept it as an analogue)
- Step 5: **If** $MIN(H^i) \not\subset H^i \cap H$ **then**
- **Choose(** H^j) (another hyper-tree)

- i:= j **Goto** Step 4
  Else H* := H^i (accept as the better fit) ♦

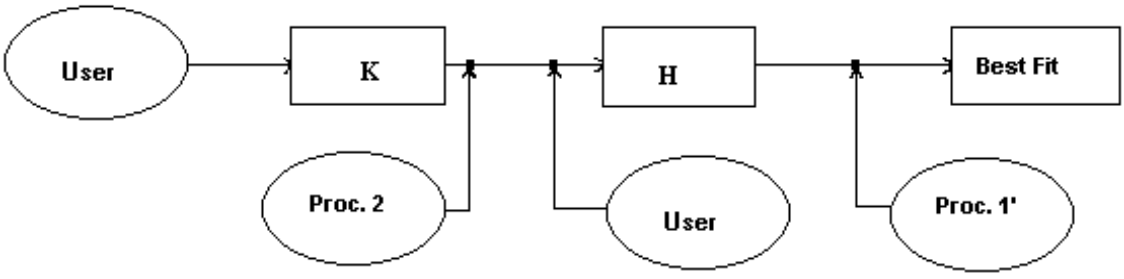
Therefore, any system intended as an implementation of S.D.M. must be based on a man-machine interaction according to the following protocol:

1. Input the knowledge base.
2. Build a hyper-tree representing all the available information.
3. Apply Procedure 1' to obtain the model that best fits the data.

It is rather obvious that step 2 involves the active participation of the model builder. In fact, determining the labels corresponding to the formulas in the knowledge base (after the application of the Reduction Theorem) is matter of idiosyncratic reasoning.

The idea here is that the scientist provides the initial set of information, which contains both general statements as well as facts. Procedure 2 operates on those statements building an hyper-graph of relations of arity three or less. The human side is called again to classify the information items and to label them. This results in the generation of a hyper-tree. Procedure 1' takes care of the last activity, namely to find a best match between the meaningful part of the hyper-tree and an already accepted model. Then, as a result a meaningful simplified representation emerges, which can be used for further scientific work.

The following figure depicts this protocol. We follow here the following convention: boxes indicate structures, ellipses represent a kind of "operators", that is, either human or automatic procedures performing actions. Finally, arrows indicate the direction of the flow of control while marked dots represent the points of intervention of the operators.



A final word on the critical intervention of the user. The main goal of it is to perform the label-attachment Procedure 3. There is where semiotics is invoked, i.e. a part of a cognitive task that intends to clarify the model-builder's set of ideas.

IV.- Model Building in Economics

There is little doubt about the role that idiosyncratic reasoning play in the hard sciences. In fact, its relevance in the field is very restricted beyond the aspects that are considered exogenous to our protocol. It is certainly present in the design and definition of the knowledge base. But it is of little use in the label-attachment procedure, since it exists a common understanding among scientists about which is the right classification of concepts.

Therefore, for the physical sciences, there can exist automatic labeling procedures. They may just proceed by following the temporal ordering of statements (using the customary representations in the form of differential equations). Or they could create a hierarchical ordering rooted on the most basic statements (alternatively the most exogenous) down to the most derivative (or more endogenous).

But in the social sciences things are different. There exist far more disagreements that depend only on the idiosyncratic characteristics of the scientists. On the other hand, only economics allow (until this point) exact (quantitative) treatments. Even so, problems may arise there for a variety of reasons:

- Theoretical concepts may not have a precisely defined empirical correlative (e.g. What is the variable that can be used as an indicator of "human capital"? The number of high-school graduates or the figures of enrollment in universities or the number of white-collar jobs?).
- There may exist differences between schools of though about which are the basic concepts (e.g. for a mainstream economist the source of value is given by the scarcity of resources and the preferences of the individuals, while for a Marxist it is the amount of labor put into the production of commodities).
- As said in the Introduction, a qualitative approximation may fail to pass the customary inductive statistical tests, even if it is intuitively correct (e.g. when do outliers be considered as such and when as legitimate outcomes in a random experiment?).

It is clear that all these problem call for the active participation of the model-builder. A little toy example may help to show how this intervention could help to clarify issues in an original knowledge base.

Let us consider the problem of economic growth, i.e. the process by which the income of a country varies in time.[4] The standard model of the last half-century was the model of Solow, which indicates that in an economy with a constant population, the per-capita amount of capital k(t+1) available at a period t+1, is the result of the preservation of the per-capita capital of the past period, k(t) –less depreciation- plus the amount of per-capita income saved in t, sy(t):

$$k(t+1) = (1-\delta)\, k(t) + sy(t)$$

where y(t) = f(k(t)) represents the per-capita amount of income produced by means of a concave technology f (i.e. a function that exhibits diminishing returns to scale), s the savings rate and δ the depreciation rate.

The point is that this system reaches a steady state k* that verifies that

$$k^*/y^* = s/\delta$$

where a slight change in the savings rates amounts to a higher steady state, while an increasing depreciation lowers that value.

A problem with this approach is that it predicts that the growth process will eventually stop, and second that if all countries share the same technology and follow the same savings practices, they will converge to an absolutely identical state. Of course no one of this predictions makes much sense in the light of the cross-country evidence.

To seek a way out of these problems, it has been advanced the idea that countries that invest more in "human capital" (labor that is skilled and can furthermore promote innovations) are more prone to grow indefinitely. The idea has been to augment Solow's model by incorporating an additional type of saving, representing the investment in education. According to this approach, it has become possible to explain the long-standing process of growth of countries that created a strong industrial system based on advanced technology.

Suppose now that an analyst faces the problem of building a model representing a growth process in a country A. Suppose, furthermore, that our model-builder wants to introduce human capital as an important factor of the process. Then the knowledge base could contain the following universal statements (where we assume the universal quantifiers)

$$K^u = \{ [y(t) = k(t)^{\alpha} h(t)^{1-\alpha}], [\, k(t+1) - k(t) = sy(t)\,], [\, h(t+1) - h(t) = qy(t)\,], [dlny/dt = s^{\alpha} q^{1-\alpha}\,]\}$$

Enclosed in brackets we find statements (which, being mathematical expressions, could be alternatively as first order expressions) that the model-builder assumes should apply

[4] This example is based on the discussion in **[Ray 1998]**.

necessarily. The first one indicates that income is in fact a function with decreasing returns to scale ($\alpha < 1$) of both per-capita physical and human capital (k and h respectively) at each point of time. The second formula postulates that physical capital increases due to a fraction of income saved, while the second shows basically the same for human capital. The final formula (actually derived from the first three) defines the growth rate of the economy as a function of both types of savings rates.

Now consider the grounded statements, involving observable data about the behavior of A's economy during the last ten years:

K^g = {[average proportion of national budget in elementary and secondary education(A) = 0.05], [average proportion of national budget in universities(A) = 0.005], [average investment/national income(A) = 0.3], [y(1)=100, y(2) = 120, y(3) = 120, y(4) = 140, y(5) = 150, y(6) = 170, y(7) = 190, y(8) = 200, y(9) = 220, y(10) =250] [estimated logarithmic participation of industry in the economy = 0.4]}

That is, it indicates two candidates to fulfill the role of q (the first two statements), a proxy for s and the sequence of values of the national income for the last ten years.

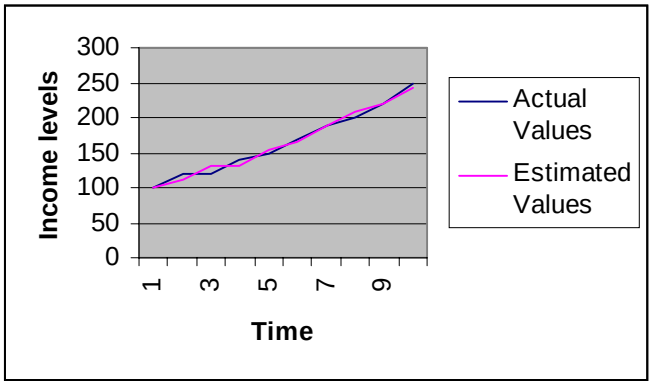
The goal of the model-builder is to determine if the hypothesis of human capital is valid enough to justify the observed rate of growth. Now it becomes necessary for him to provide a labeling of items of information. To make a long story short let us assume that:

- **Third rank**: income, growth rate of the economy
  where income is represented by the series y(1)...y(10), of the values of A's income for ten years, while the growth rate is represented by the following differential equation that indicates that the (long run) proportional variation of income depends on the savings rates of physical and human capital: $dlny/dt = s^{\alpha}q^{1-\alpha}$.
- **Second rank:** average investment/national income, average proportion of national budget in elementary and secondary education, estimated logarithmic participation of industry in the economy.
- **First rank:** average proportion of national budget in universities.

Notice that this labeling (only of the most relevant statements) indicates that for the model-builder, the most fundamental pieces of information are the theoretical definition of the growth rate of the economy as well as the income levels effectively observed. In a second level he puts what he wants consider as the proxies for s, q and α, respectively. Finally, in the lowest level he includes what he may consider an alternative for q, in case his first option fails.

Then, when given the opportunity to declare (asked by Procedure 1') what is the accuracy he wants of the model, *MIN*(H^i) , he can equivalently say that he will accept the model only if it fits with the values he has given with less than 0.1 of difference. This is because he can check out with his assumed values whether effectively, the participation of industry in the economy is consistent with the model. Otherwise, the alternative q must be checked. If this does not work then he must go back to his desk to think things through again.

Fortunately, as the following graph shows, his hunches worked well, both in assessing the right values for α as well as for the closest proxy for q:



V.- Concluding Remarks

We have laid out in this paper a proposal for the application of semiotic engineering in the field of economic data mining that complements the more traditional statistical data mining methods. We claimed that this provides grounds for a better understanding and solving many problems that were of difficult to handle with less expressive tools.

Our research has been limited to an exposition of the foundations of the method we advocate for. The scheme for an interactive architecture was also presented and applied to analyze a relevant economic problem. Our example shows the potential of this new approach.

This research must be seen as an investigation in the broad area of knowledge engineering. That is, on the foundations of methods for the use of knowledge in computational systems. In a further step, these ideas should be applied to the design of the appropriate software systems.

Future research will be concerned with extending this to other areas in economics and finance, as well as to developing the full extent of semiotic engineering in these fields.

References

[Devlin 1993] Devlin, K. The Joy of Sets, Springer, NY 1993.

[Graham et al. 1990] Graham, R. – Spencer, J. – Rothschild, B. Ramsey Theory, Wiley & Sons, N.Y. 1990

[Hanson 1961] Hanson, R. Patterns of Discovery, Cambridge University Press, Cambridge 1961.

[Levesque 1984] Levesque, H. Foundations of a Functional Approach to Knowledge Representation, *Artificial Intelligence* 23(2), 1984.

[Maróstica 1997] Marostica, A. **A Nonmonotonic Approach to Tychist Logic**, in Studies in the Logic of Charles Sanders Peirce, N. Houser et al. (eds.) Indiana University Press, Bloomington IN 1997.

[Maróstica 1998] Marostica, A. **Semiotic Trees and Classifications for Inductive Learning Systems** in Semiotics 1998, J. Deely (ed.) University Press of America, NY 1998.

[Maróstica 1999] Maróstica, A. **Peircean Diagramms and Programming Languages**, presented at the VII Semiotic Conference, Dresden 1999.

[Maróstica- Tohmé 1999] Marostica, A. - Tohmé, F. **The Role of Automated Semiotic Classifications in Economic Domain (I)** , Presented at Computation in Economics and Finance `99, Boston 1999.

[Marty 1990] Marty, R. L´Algèbre des Signes, John Benjamin Publishing Company, Amsterdam 1990.

[Myers –Konolige 1996] Myers, K. – Konolige, K. **Reasoning with Analogical Representations in Diagrammatic Reasoning**, B. Chandresekaran et al. (eds.) Diagrammatic Reasoning, AAAI Press/ MIT Press, Menlo Park CA 1995.

[Peirce 1960] Peirce, C.S. **How to Make Our Ideas Clear** in C. Hartshorne et al. (eds.) Collected Papers of Charles S. Peirce vol. 5, The Belknap Press of Harvard University Press, Cambridge 1960.

[Ray 1998] Ray, D. Development Economics, Princeton University Press, Princeton 1998.

[Russell-Norvig 1994] Russell, S. – Norvig, P. Artificial Intelligence: a Modern Approach, Prentice Hall, Englewood Cliffs NJ 1994.

[Simon 1968] Simon, H. **On Judging the Plausibility of Theories**, B. Van Rootselaar and J. Staal (eds.) Logic, Methodology and Philosophy of Science III, North-Holland, Amsterdam 1968.