

Data Semantic Associative Analysis and Synthesis[♥]

Sergei Levashkin⁺

⁺*Centre for Computing Research-IPN
UPALMZ, Ed. CIC, Mexico City, 07738 MEXICO
E-mail: sergei@cic.ipn.mx*

Victor Alexandrov^{*}

^{*}*Saint Petersburg Institute for Informatics and Automation-Russian Academy of Sciences
39 14th Line, Saint Petersburg, 199178 RUSSIA
E-mail: alexandr@mail.iias.spb.su*

ABSTRACT

An emerging area of digital data processing is the computer-based intelligent analysis and synthesis of information flows/streams. These include in particular processing of audio, hypertext, image, text, video, mixed, etc. data that travel over the Internet. The final goal of most current research efforts is to build the Semantic Web in the sense defined by the World Wide Web inventor Tim Berners-Lee. One of the main problems for success of this project is how to automatically extract (and then interpret) semantic components from unordered data flows/streams. This problem was already stated by Gestalt school in early 1920s in case of the human visual perception. Note that an efficient human-machine interaction is the stumbling block of modern computer technologies. In this chapter, we are going to discuss the foundations of semantic associative data analysis and synthesis to approach the solution of these fundamental problems, whose solutions are indispensable for the Semantic Web development. Data semantic associative analysis and synthesis consists of two main components – semantic-mind data analysis and object-oriented data integration (synthesis). They will be described in this chapter.

*Think about meaning...
The words are becoming themselves.
(Carroll, L.)*

INTRODUCTION

We are observing day-by-day more and more simple procedures of data content acquisition, processing, and transmitting. However, these constitute only one part of the problem. The other part is that access to heterogeneous and independent databases should be equally simple. Unfortunately, identification of desired (useful) information by searching and filtering has become more and more complicated. That is why the treatment of differences in the structure and semantics of the data stored in repositories plays an increasingly important role in modern information systems. The first studies on interoperating information systems have been directed toward syntactic (i.e., data types and formats) and structural heterogeneities (i.e., schematic integration, query languages, and interfaces). As interoperating information systems increasingly confront more complex knowledge management tasks, the technology needed to deal successfully with these issues must focus on the semantics underlying the data used by those systems.

[♥] Work done under partial support of the Russian and Mexican Governments: RFFR (Russia) and CONACYT, SNI, SIP-IPN (Mexico).

The use of computers for data processing, storage, and transmission in a historic perspective includes the following subsequent stages:

- *Data (temporal series, matrices, etc.) representation and processing using traditional mathematical models. The search for computer-oriented mathematical models in signal and image processing.*
- *Generalization of the concept of data. Software development of storage, search, and reference for texts.*
- *Development of computer programs of translation from one language into another.*
- *Development of decision-making tools and ontology driven information systems.*
- *Continuous data integration (text, audio-video) leads to the subsequent stage of software development—intelligent agents such as Data Mining, Copernicus, etc.*

This was being thought about among researchers since the early 1980s. (Simon, J.-C.):“*The arrival of a third generation of machines made it possible to experiment with information (data) provided by instruments in many fields of observation: in visual images, in spoken words, in the fields of physics, medicine, economics, linguistics, etc. A new field of observation and study was transferred from philosophy to experimental sciences. This was a general and crucial phenomenon in the history of Man’s efforts to understand Nature: the telescope gave birth to astronomy and metallurgical, chemical, electrical and vacuum techniques gave birth to physics. There is a certain paradox in the fact that the computer, designed for commercial accounting and scientific computation, gave birth to pattern recognition and artificial intelligence*”.

For the pure mathematician, *data recognition* is a trivial problem that can be expressed formally as follows: Let X be a representation space, preferably a “nice” topological space, and let Ω , the interpretation space, be a finite set of names. Recognition or identification is a mapping $E: X \rightarrow \Omega$, to which certain properties are described and from which elegant theorems can be deduced.

This, however, is not where the problem lays: in practice, the question is one of constructing E , i.e., of providing operators or programs which given any x from X , enables us to decide automatically onto which ω from Ω the element is mapped.

An extended definition of E , to be held in full memory, is out of the question even for small-scale problems, because here we come up against the problem of computational complexity. In the search for usable operators, data recognition is continually confronted with problems of *information complexity* best defined by (Kolmogorov, A.); so many data recognition problems are exponential that we are constantly obliged to adopt less than optimal, polynomial solutions. Data recognition is first and foremost a battle against complexity.

The other guiding thread appears to us to be the *semantics* of the general data recognition problem, which varies according to the question under consideration. Is there in fact a general, universally applicable method for constructing a data recognition operator?

We must therefore treat each problem in a specific manner and search for any items of information that will enable us to construct the required operators. Our view is that information is to be found in the properties of

- the representation space,
- the interpretation space or spaces.

However, the solution of applied problems requires the bridge between the strong structure of mathematical concept of space and empirical properties of the data as input information that leads us to the following conclusion: *semantic information* is to be found in the properties of

- data representation;

- data interpretation as knowledge of subject domain.

This means that we use the following model of a recognition process: find an element from a finite set, which is equivalent to the unknown object. When the element is found, the object is attributed with its all known properties. Essentially, this is the principle of “identification by indistinguishability”, first formulated by (*Leibniz, G.W.*).

Information flow (IF) is binary digital data stored and processed by the computer. This is the basic element of digital technologies, i.e., Turing’s (machine) world. On the other hand, IF is data that obtain from the outer (to computer) world and that have very different representations—environmental monitoring, text, music, speech, images, etc.; thus, IF is a part of the human’s world. At present, the problem to adapt computer systems to human perception and cognition is apparent (MPEG-7), but yet not carried out. In the present chapter, we would like to discuss some crucial aspects (*Semantic-mind analysis* (SMA) and *Object-oriented data integration* (OODI) of IF)¹ of this problem, outline the difficulties encountered on the way to its solution, and guide the reader toward the bridge between computer and human worlds. All this can be observed within the general context of human-machine interaction as a semantic approach to computer data analysis and synthesis.

The goal of the present chapter is to sketch the general frameworks of SMA (analysis) and OODI (synthesis) in different types of information flows.

DATA STRUCTURE AND INFORMATION

The concepts of data, knowledge, and information have held the steadfast attention of scientists during centuries. Although, knowledge as a research topic is one of oldest in the history of science, we have only a general philosophical understanding without a strong formal definition (*Floridi, L.*)². Such a definition is required for knowledge-based computer systems (e.g., on-line and Internet education).

The main difficulty in formal definition of these concepts probably lays in the fact that man as an information carrier and translator has very specific structure of their representation and processing and thus cannot easily abstract and detect these concepts from a structured shell.

Known concepts of data, knowledge, and information explicitly show their direct link to the form of representation. “*A knowledge representation is a certain method in which the experience becomes structurally defined in wide-spread and common terms*” (*Minsky, M.*).

Research concerning knowledge structuring can be divided into two lines. The first is related to the study of individual intelligence, but it does not yet systemize it. Therefore, this research does not provide an entire description of knowledge representation in the human brain. On the other hand, collective intelligence is much more accessible for research because numerous forms of knowledge representation have been constructed over the centuries.

Several examples of the social experience of mankind show a high degree of influence on the final result of man’s activity with regard to the choice of one form or another of knowledge representation. A typical example in science occurs when two scientists making the same discovery cannot understand this fact because they use different forms (notations) of knowledge representation. To avoid such situations, it is important to represent any information in compact graphic form that reflects its basic essence. The representation must provide a possibility to determine certain details of information flow using as tools that are as simple as possible. We would like to highlight here that this way provides good results not only

¹Formal definitions of SMA and OODI will be given in subsequent sections.

²The notions of data, information, and knowledge are fundamental in computer science, artificial intelligence, and the Web development.

in education or knowledge popularization but also directly in practice and science. This is probably why computer science specialists prefer computer programs that provide precise details of processing to long-winded ambiguous explanations or fogware.

Thus, research of the interrelation between information content and its structured representation is important from both theoretic and practical points of view.

Information

When investigating an object, anyone is stated (probably in an intuitive level) three questions:

- What is this?
- How is this related to something that I know?
- What are the characteristics of this object?

In other words, the human being attempts to identify the object, to order new information that corresponds to his early experience, and finally detects qualitative properties of the object. He partially obtains this information in an empirical manner by analogy with other objects in his Eigen-system of representation. This is an example of the developing system of knowledge representation.

Problems of identification, ordering, and semantics appear in any research but not necessarily in this order. As a rule, these problems are repetitively applied to define (and modify) the solutions of each problem obtained in the previous step. We shall consider these three problems as follows and attempt to define their meaning within the context of structured representation of information.

Identification

Identification is the oldest human activity and at first glance represents sufficiently well a simple procedure of name assignment of objects, processes, and concepts. If $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ is a representation of original object, X is the representation space and $\Omega = (\omega_1, \omega_2, \dots, \omega_p)$ the set of names, then identification is mapping ξ from the representation space to the set of names:

$$\xi : X \rightarrow \Omega.$$

Of course, the procedure of mapping should be constructive. However, in attempting to apply such a procedure we encounter several problems:

- Should we assign to any object a name? (*the problem of Leibniz' "identification by indistinguishability"*);
- How novel and informative should a concept or process to assign it a name be? (*the problem of label*), and
- How should new objects, concepts and processes be named for further easy use (*the problem of name structuration*)?

These three problems graphically illustrate that identification is not a mechanical procedure of name assignment (although we should perform such mechanics for pointer construction, for example). Thus, in subsequent sections we shall also discuss these three problems in relation to the concept of structure.

Leibniz' principle of identification by indistinguishability

Identification is related to classification, association, analysis, and synthesis of knowledge, i.e., it is a necessary link in cognition and probably is the only link: "*A hard classification implies the hardness of words. A soft classification implies an inflexible look at things*" (Dodgson, C.L.).

The first problem is essentially Leibniz' principle of identification by indistinguishability: “*Two objects are indistinguishable if all their properties are the same*” (Leibniz, *G.W.*).

The problem of identification was first strongly formulated by Leibniz in his Ph.D. thesis and led him to the concept of congruence: $ABC \cong ABX$. Congruence or universal characteristics according to Leibniz mean “... *man can draw while not being an artist...*” The method consists of the rule for identifying X as a set of points, objects, and things and their properties by the name and properties of known object C . At first glance, this method provides unlimited possibilities for learning and cognition. However, anyone who frequently uses the Internet for scientific information search constantly requires conceptual identification—comparison of things searched for with known information that can be represented in different forms and languages.

Moreover, if we use Leibniz' principle we always must take into consideration the following two limitations:

Limitation 1. Absolutely identical objects do not exist. An infinite number of properties characterize any real object. Therefore, we use only a subset of properties when comparing objects or rejecting others. A given problem or subject domain defines the choice of properties, which are essential for object comparison; hence, the same objects can be identical with respect to one subject domain and different with respect to another. Thus, the first problem is choice of identification properties, and

Limitation 2. Values of quantitative and qualitative properties can only be defined approximately, e.g., the weight is approximately 1 kg, the color is blue, etc. Thus, indistinguishable objects at the crude level of computing or measurement of parameter values can be distinguishable only after re-computing and re-measuring. Thus, the second problem is exactness of representation of object properties.

Sense of the principle of identification by indistinguishability is detection of object identity by known properties. Applications of this principle are different in different subject domains: in algebra—relations of equality and identity; in geometry—relations of congruence and similarity; in linguistics—synonyms; in statistics—coefficients of correlation, etc.

The problem of label

This problem can be illustrated as follows: *Having a name, the knowledge fragments become ‘frozen’ and ‘untouchable’ because the label can only be used for constant value. We must consider a world that is constructed from names such as house—from the bricks, which must be broken into parts and investigated separately to understand the whole.* That is the point! The reader encounters difficulties in understanding the main idea of a scientific paper that contains numerous abbreviations or introduces new terms for unjustifiably large number of concepts. Thus, the reader must usually break down and digest unusual concepts.

(Zipf, *G.K.*) analyzed the naming of new concepts and proposed the following law: name is a function of the frequency of use of a new concept in some limited social group; thus, professional language, dialect, slang, and other language subsets are generated. Naming is the desire to increase efficiency (velocity) of interaction in human society.

The appearance of new names inside a certain group is not sufficient for their introduction into a larger group; otherwise, the tendency to optimize transmission of messages can lead to the inverse result. For example, if each number has its own hieroglyph (digit) then number memorization and use would be difficult despite the fact that each number must be written, for example, shorter than in decimal notation.

Therefore, the problem of label should be solved taking into consideration the following two conditions³:

- Sufficient stability of the concept to be labeled, and

³In the history of natural language, the problem has been solved this way.

- Sufficiently frequent use of the concept so that efforts for memorizing a new name are lesser than its perception through description.

Problem of name structuration

This problem involves the assignment of names in such a manner that their associations to objects are simplified to the fullest possible extent. Let us consider a number. By thinking, we studied to perceive quantitative information. First, we perceived one stone and two stones as objects not linked to a unified scheme; thus, counting skills were developed step-by-step. At some time, it became possible to assign different and unsystemized names to various quantities, e.g., a dozen. More complex calculations that employed a wide range of numbers rendered such an archaic method loose. A positional number system with its strong structure of number description and generation of a description of any big or small number does not yet require such types of names. Thus, we use names like a million with the same ease as when we write 10^6 or 10 to the sixth.

Ordering

In Mathematics, the problem of structuration of object names in ways useful for applications and the problem of object ordering are solved by means of unified methods. Probably this is one of the causes of the high descriptive power of Mathematics. Although mathematical solution is quite natural, there is no similar correspondence between name and ordering in other sciences. To illustrate this thesis, let us consider computer-based numerical analysis.

A peculiarity of numerical analysis is the numerical model: one-to-one mapping between input and output data. Nevertheless, great difficulties can appear for ill-posed problems. However, these difficulties are exceptional. In most cases, strong mathematical description of the original problem provides adequate solutions. As a rule, input and output data as well as the model's parameters are objects of the same nature and can be strongly defined. This explains the elegance of the algorithmic model in numerical analysis. Moreover, admitted operations on objects represented by numbers are known a priori. Only interpretation of the result (assuming that all definitions are correct) can result in uncertainty of the numerical model.

Information analysis does not possess a mathematics-like axiomatic base. In particular, the set of operations on objects is not defined by ordering of natural numbers as in numerical analysis.

The concept of ordering in Information Science is often used in the narrow sense as a synonym of certain temporal unfixed object relationships. For example, people in a queue can be ordered in different manners: by height; professional or sporting interests, etc. In contrast with this pragmatic approach, we understand ordering as something implied from basic object content. For example, the meaning of the word six implies that the object with the name three precedes such an object (in linear systems); the meaning of the word father implies its following location of the object, son (in hierarchical systems). Of course, other examples can show situations in which these orderings do not hold. In this case, we can use the words as temporal labels and not their usual meanings. Note that not only numbers admit such (partial) ordering. The main difficulty here is that non-numeric information expressed by words does not admit one-to-one ordering due to the multiple meanings of each word. A possible solution is multi-dimensional ordering. The *Mendeleev Periodical Table* (Atkins, P. W.) is an outstanding example of this type of ordering. The Table shows how to take into consideration different properties of chemical elements and how to order each element.

The concept of ordered number could be established independently of its quantitative substance as the result of abstraction from qualitative differences of the equally arranged sets. Understanding independent number substances became clear only in the 19th century (Peano, G.).

To express solely the number's ordering substance, we could use letters without their quantitative meaning only by using alphabetical order. In this sense, the concept of natural numbers \mathbb{N} has been developed. In \mathbb{N} , each element is defined by its location in the series.

The French word *l'ordinateur* literally means *fixing order*, which is closer in meaning to modern computer use than the word in English, *computer*, which literally means *calculator* and mainly characterizes the prescription of computers in initial stages of the computer era.

V. Bush (Manhattan Project) was the first who noted this fact and who proposed a new approach to the order organization of concept indexing for search in the computer. In 1945, Bush published the semi-utopic work, "*As We May Think*", which remained during many years the most cited publication in the field of man-machine interaction. Amazingly, the work contains a description of *browser*—a system for text-graphics information search. This system, called *Memex*, included a large library of texts, photos, and drawings. Although Bush showed great forecasting talent, *Memex* was not a computer because it used microfilms and photo-elements. The main peculiarity of *Memex* system was the possibility to input relations among library elements. The corresponding mechanism was inevitably bulky but logical. If the user had two documents (each in single image) that he wanted to relate, he struck via a special relation a *name* button and this name appeared at the bottom of each image. To obtain the document related by a certain relationship to another, he simply entered its code via the same button. This we call *hypertext* at present.

Some time after D. Engelbart (developer of the first computer mouse) developed a system that linked hypertext with newly invented devices and graphics (*multimedia*).

In contrast, the search engines of the modern Internet use the principle of search by keywords. This principle is not efficient because an engine displays a great number of vague references from different sources related to each keyword. Among methods of identification in the Internet, the most interesting principle of information search belongs to the server www.altavista.com, which shows the complete route to the site (page) where a keyword is found. However, this approach is not a semantic (or meaningful) search. ISBN (book systematization standard) uses the subject-alphabetical principle of ordering but does not provide detection and construction of semantic-meaning concepts as inter-disciplinary and inter-topic relationships (*associative pointers*). Construct associative pointers means the establishment of relationships between a given and well defined, tree-like structure of unique pointers (*subject domain*) and a set of semantically meaningful concepts (*course domain*).

Problems of semantics and meaningful search for words in IF are considered in the subsequent section.

Similarity and confusion measures

They are main tools for the semantics quantification. Formally, the similarity (confusion) problem can be described as follows.

Let us look again at the ordering and similarity of the elements of finite sets E . A reflexive and transitive relation can be defined for the set F of all pairs of elements in $E \times E$, which is called a *partial order*. With $x, y, z, u \in E$, this is written $(x, y) \leq (z, u)$ meaning that x resembles y more than z resembles u . Such a relation does not necessarily apply over the complete set F because they may be pairs of elements that are really not comparable. If it does apply over the complete set it becomes a *total order*.

It is difficult in practice to set up such a partial order if the number of elements of E is large, and if it is possible it is difficult to make this order without running the risk of generating contradictions. In fact, the only practical way to establish a partial order is to define a numerical function of *similarity or dissimilarity (confusion)* that can be computed in terms of the attributes of every element of E : the *dissimilarity (confusion)* $\lambda(x, y)$ will be smaller the more closely x resembles y .

The same partial order can be generated by any of an unlimited number such functions. Some dissimilarity functions, however, may not be distances. Thus, by definition, a distance $d(x, y)$ satisfies:

- (a) $d(x, x) = 0$;
- (b) $d(x, y) = d(y, x)$;
- (c) $d(x, y) \leq d(y, z) + d(x, z)$.

Relation (a) can be satisfied by making $\min \lambda = 0$, which is simply a change of origin, and for (b) it is sometimes necessary to make λ symmetrical. Relation (c) may not hold, however, although it can be made to hold by adding a sufficiently large constant λ_0 to the values, whilst retaining $\lambda(x, x) = 0$. The corresponding partial ordering will not be changed. Thus the following general statement can be made: *it is always possible for a finite set E to make simple modifications that will transform dissimilarity into a distance measure without affecting the partial order.*

In early approaches (*Simon, J.-C.*), the measures of similarity between variables have been considered using *Kendall, Hamming, Russell & Rao, Jaccard, Kuzlinsky, Yule*, and other distances, and a *contingency table*. These result, however, insufficient for variables that take symbolic values.

In context of the *identification problem*, distance measures for objects and concepts have been proposed too. The idea of classification carries with it in the implication that a *descriptor* or *symbolic description* can be defined for each class and will in some sense be representative of the class; possible descriptors are *a symbolic description of the class; a feature represented as a relational structure*. If X is the representation of an *object*, we can define a *concept* A_i as an entity which is such that an ordering of the couples (X, A_i) can be established for $i = 1, \dots, k$. A concept is not necessarily associated with one particular class; this may be so, e.g. in the case of descriptors, but in other cases the concepts may overlap, as for example for probabilities. The “distance” $D(X, A_i)$ between an object and a concept can be used effectively as a measure of similarity or dissimilarity. In other words, the “object-concept” distance D is used as a characteristic function. The following functions have been used: *probability, fuzzy assignment, inertia and potential, nearest neighbors (q-NN)*, etc. The main problem with these approaches is that they often omit the context of a classification problem under consideration.

Let $f(X; A)$ be a measure of similarity or dissimilarity between the representation of an object X and a concept A . Such a measure limited to a single A would be of little interest; if for example, A were the symbolic description of a class, there would clearly be at least two classes, A and $\neg A$ (“not- A ”). However, it can happen that the concepts A do not coincide with the classes. Let $\{A_i\}$ be the set of concepts under consideration: the assumption that these concepts can be grouped together into a set implies that the set operations of union (\cup) and intersection (\cap) can be applied. The A_i are not necessarily disjoint, i.e. several can apply simultaneously to a single object, although in the very special case that each identifies one class they are clearly disjoint. Let $@$ be the algebra generated by the A_i and operations: \cap , \cup , and \neg ; the elements of $@$ form the *interpretation space*. If the concepts A_i are expressed by predicates, the operations are written as \wedge and \vee respectively and we have Stone’s theorem: *all distributive logic systems are homomorphic to a distributive lattice of subsets of a set*. We may recall that if p , q , and r are predicates then, \wedge is distributive with respect to \vee .

Given $f(X; A)$ and $f(X; B)$, the *fundamental question* is: What are the values of $f(X; A \cap B)$ and $f(X; A \cup B)$? This can be answered in terms of two laws or operations: (i) an additive law \oplus , homomorphic with \cup , and (ii) a multiplicative law \otimes , homomorphic with \cap and distributive over \oplus , and the *answer* is:

$$\begin{aligned} f(X; A \cup B) &= f(X; A) \oplus f(X; B); \\ f(X; A \cap B) &= f(X; A) \otimes f(X; B). \end{aligned}$$

DATA SEMANTIC ASSOCIATIVE ANALYSIS AND SYNTHESIS: SEMANTIC-MIND ANALYSIS AND OBJECT ORIENTED DATA INTEGRATION OF INFORMATION

“*Semantics—the study of the meanings of words*” (Webster II). Therefore, semantics is the search for cognitive, associative object identification in IF. Keywords in text, sound-shapes in audio flows, segments in image flows (*data invariants*), etc., are the object-oriented data of IF.

“*Mind—the part of human being that governs thought, perception, feeling, will, memory, and imagination*” (Webster II). In other words, meaning is the capability to understand, to feel, to perceive, and to imagine, i.e., restoring an entire piece of knowledge by some segment or part. For example, the program *Guess a Melody* (Kosch, H.) identifies the musical piece by a few musical phrases; the-raster-to-vector conversion system *A2R2V* is oriented toward searching for names of the converted set of pixels, thus, object-oriented data integration to *Geographic Information System* (GIS).

On the other hand (See first paragraph of this section), semantics is the adequate, meaningful search for *words* in IF (e.g., Google search www.google.com as zero approximation), while the meaning is the extraction of the *subject domain* in IF (e.g., adaptation of a Physics textbook or articles in a specialized journal for the secondary school).

Subsequently, *Semantic-mind analysis* (SMA) of IF is the meaningful search for object names and definition of the subject domain as a set of found names (e.g., 32 letters require five bits of information by Shannon, while the same letters require fewer bits of information by Morse. The difference is that the Morse code took into consideration frequency of use of letters in text and named letters by symbols, such as point, dash). *Object-oriented data integration* (OODI) is input into a particular computer-based application of the output of SMA. This input suggests special SMA-output data organization, compression, storage, processing, etc., which is dependent on that application. For example, vector object integration to GIS is straightforward. Moreover, it is more desirable to store that object in GIS under the corresponding name only. In the following sections, we will use the abbreviation SMA/OODI for newly introduced concepts.

SMA/OODI: Metadata and MPEG-7

For the sake of simplicity with regard to the following explanations, let us now consider the concept of metadata. Metadata are aids (via HTTP, FTP, etc.) for network users to follow up information resources and optimize their primary and secondary uses (see, e.g., IEEE European Colloquiums’ Multimedia Database and MPEG-7). SMA/OODI can be also seen in the context of MPEG-7 technology as a system potentially adapted to process IF as metadata (*user-oriented processing*). It is motivated by the fact that at present, digital *audio-visual information* (AVI) can be accessed by anyone not only for consumption but also for yield. This converts us at least potentially into *content makers*. We can publish and transmit digital information yielded by us via the Internet. However, day-by-day more and more simple procedures of audio-visual content acquisition, processing, and transmitting constitute only one part of the problem. The other part is that access to existing data should be equally simple because of the huge amount of AVI yielded daily throughout the world. Unfortunately, identification of desired (useful) information by searching and filtering has become more and more complicated. Even if open-source resources are used for access in such specific area as GIS (open-GIS and similar), the problem remains due to data interoperability and homogenization. Thus, GIS-oriented people refer to *the Semantic Geospatial Web* (Egenhofer, M.).

Therefore, the problem of fast and efficient identification of audio-visual contents in IF is emerging. This problem has motivated the Multimedia Content Description Interface Project by MPEG, also known as MPEG-7. MPEG-7 attempts to define standards for description of different AVIs that include images,

image sequences, speech, audio, graphics, 3D models, and synthesized audio independently of representation formats (*Kosch, H.*).

Our general considerations in this chapter are aimed to follow up this research line and hopefully lead to better understanding of this research.

The law of progressive simplification

We visually perceive the outer world as an optical process of transmission of images to the retina and subsequently construct the scene model as the spatial-temporal structure of objects and their relationships by local analysis and synthesis. This allows for extracting image semantic relationships and advancing to the verbal level of representation of initial visual information by logical analysis (*Alexandrov, V.*).

Subsequent passage from visual perception to verbal (logical) cognition provides progress in many areas (Computer Science included), well illustrated in the following (*Toynbee, J.*): “... *In the history of handwriting, we observe not only the correspondence between the techniques of writing development and the form simplification but also these two tendencies are identical ‘de facto’ because the technical problem of writing as registrar and translator of human speech should have been also solved. This is a clear representation of the widest sphere of human language by means of maximal optimization of visual symbols, i.e., etherification is the law of progressive simplification...*”

Human reasoning, memory, and cognition as *self-substance* do not exist. On the contrary, they are simply *the names* (labels), synonyms that associatively reflect the result of the human’s brain functioning as continuous cognition, i.e., the tool (processor) of etherification. The human brain’s self-sufficiency is based on *the law of progressive simplification*. A particular case of etherification is *the principle of least effort* by (*Zipf, G.K.*)⁴.

Basic sources of information exchange and their types, and evolution of perception and thinking in the human world are shown in Table 1. If we look at this table from the point of view of computer-based systems (Turing’s world), some analogs are straightforward. Let us only highlight some of them, leaving to the reader to complete others as an exercise.

Suppose that you are a computer specialist and have never heard about cognitive science or have never learned any other science. Are you familiar with concepts such as genetic algorithm, environment, sampling, learning, self-learning, knowledge, etc., from your area? Of course you are. Then look again at the table.

Computer genetic algorithms and programming involve methodological background similar to the natural processes and laws that govern the transmission of genetic information. Thus, the concept of inherent knowledge can be spread and accepted in both worlds (Stage I).

Environment and environmental models are now commonly accepted concepts of Turing’s world. They are rather the computer media (*UNIX, Linux, Windows, WordNet, open GIS, Spatial Semantic Web*, etc.) in which computer programs are run than physical media. At present, computer experiments are common and available to nearly everyone. Moreover, newly invented computer environments encourage experiments that we call by analogy research instinct (Stage II).

Sampling is widely used in image processing and pattern recognition to imitate the characteristics and behavior of something already known from “parents”, “teachers” (sample set, seeds, etc.) to acquire new or identify old information from unordered data sets (Stage III).

⁴In other words, we adopt here a strongly nominalist point of view. It is rather our hypothesis and model than a fact that everyone should accept. The only aim of this “sound” affirmation is to guide the reader toward the computer-oriented model as quickly as possible.

Learning and self-learning are represented as powerful tools of new knowledge acquisition in both worlds. Moreover, learning and self-learning computer systems usually oriented to tune system parameters are most efficient in automatic modeling (Stage IV).

Knowledge-based computer systems applied to digital data processing are now revisited and aimed toward development of self-consistence virtual reality similar to human's imagination (Stage V).

Table 1. Evolution of perception and thinking (levels of information exchange) ordered by rows. We aim for this to show some analogs between human and machine worlds in this evolution.

| Stage | Source | Information process | Model | Transmission tool |
|-------|--------------------------|--|---------------------|--|
| I | Biological kind (Nature) | Transmission of genetic information | Inherent knowledge | Inherence, instincts of life and death |
| II | Environment | Self-learning | Environmental model | Research instinct |
| III | Parents, flock | Sampling | Behavioral model | Imitation |
| IV | Teachers, society | Transmission of exciting knowledge, learning | Model of life | Language |
| V | Cognition | Synthesis of new knowledge | Model of the world | Imagination |

TOWARD A PROBLEM-ORIENTED SEMANTIC ENVIRONMENT

In taking up again previously mentioned promptings within the context of SMA/OODI, we should note the following.

Images or events of any nature involve *semantics*. Semantics is universal and context-interpretative in a certain finite space of meanings and events. This statement is based on the understanding (probably multi-valued) of any image or event of the outer world even if they do not yet possess a semantic context.

At present, organization of system analysis of data flows that aim to detect the interpretative in different contexts structured elements and a fast search for commonly accepted data structures experience great difficulties. This means that further significant development of data flow processing is impossible without a knowledge-based semantic-oriented analysis of data structures. The latest developments in the Internet graphically illustrate our statement. Of course, the knowledge bases themselves are not functional without special methods of preliminary analysis.

Here, correspondence of the problem-oriented environment to the level of the problem under consideration and the capacity to process meaning arises once again. For example, signal digital representation is solely based on the abstract concept of band capacity and does not take into account semantics: what is this signal—text, music, or image?

To illustrate certain principles with regard to *meaning processing*, we temporarily disregard the concept of generalized context.

If we denote something, then we assume that the meaning is understood, explicit. We do not go from sounds to images and from images to meaning: from the beginning, we are embedded in meaning, and being able to express it in sentences. Meaning prescribes possible denotations and conditions. Moreover, meaning is the object of the following sentence. If a certain name is assigned to a sentence, then clearly each name that itself denotes an object can be the object of the new name denoting its meaning: N1 addresses N2, which denotes the meaning of N1; N2 addresses N3, etc. The language for each of its

names must contain some name for the meaning of this name. Such infinite multiplication of verbal essences is called a *Freguier paradox*. In the meaning-object relationship, Freguier paradox locates the place of search for the meaning and indicates the moment of appearance of so-called *context-meaning dependence*. We use Freguier paradox as a methodology of the *dynamic structures* of data relationships in the problem-oriented environment. These relationships in difference with arbitrary environment are defined by means of associatively organized identification and allow eventual, i.e., meaningful structuration. This provides a new look at problems of efficient organization of the computer environment for data processing and understanding.

The first problem—compact representation of information concerning complicated systems — forces the search for new structured forms of knowledge representation. From our point of view, *self-similar recursive structures* provide adequate description tools. Moreover, self-similar recursive structures not only handle information but also manage information in a similar manner as genomic programs. One more important fact is their simplicity and regularity, although the result of application does not provide such as impression (*Alexandrov, V. & Gorsky, N. (1993) and Alexandrov, V. & Gorsky, N. (1991)*).

Secondly, the problem of memorizing the data structuration must be solved in a flexible, manageable manner so that signs that represent information segments are ordered by values, (*Levachkine, S. & Guzman, A.*).

The problem of development of computer knowledge-based systems is illustrated very well by (*Hofstadter, D.R.*): “*A computer does not have automatic sensitivity to the images it processes. Of course, we could not expect this. It only executes the program like an old saw. The computer does not tire by adding columns of numbers even if all numbers are equal. Men get tired. What is the difference? Obviously, the machine misses something that allows it to have unlimited patience for repeated operations. The missing detail can be described in a few words: this is the capacity for self-observation, contact with the outer world; this is the capacity to perceive an image of proper activity and carry it out at any level of abstraction. Meta-knowledge and knowledge are completely mixed between themselves in a unique flow and are mutually enriched. This renders self-observation as an automatic implication of the memory’s structuration. How is this amazing flow organized in the human brain?*”

We cannot answer Hofstadter’s question even approximately. Undoubtedly there is the need of hierarchical knowledge ordering, but *the hierarchy* should possess a special form that Hofstadter calls “... *complete mixture in unique flow*” (*Levachkine, S. & Guzman, A.*).

Natural organization of memory demonstrates once again the efficiency of a system approach with flexible inclusion of different tools for the problem of structuration of information. For example, (*Arbib, M.*) wrote: “*Many people have discussed the problem of whether the human brain is a sequential or a parallel computer? This is false opposition. Considering the eye’s motion, we observe some sequence of operations but understand that strong parallel computing is required within each time segment*”.

If we can successfully designate to machines the capacity to make decisions on an experimental basis, fully using any given insufficient data, furthering precision and extending these decisions with newly arrived additional data, then we would have a computer that does not require an explanation of *how-to-do*, but only *what-to-do*. Of course, the language of human-machine interaction should possess less logical and arithmetical depth than the internal language of the computer in this case.

FUTURE RESEARCH DIRECTIONS

Due to the chapter space limit, we restrict the future research directions to *Ontologies and hierarchies*, *Semantic mapping between different conceptualizations in a subject domain*, and *Web intelligence: Semantic services integration* topics.

Ontologies and hierarchies

Previous and recent investigations in information retrieval and data integration have emphasized the use of ontologies and semantic similarity functions as a mechanism for comparing objects that can be retrieved or integrated across heterogeneous repositories, the Web included. In this context, ontology is a type of knowledge base that describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. Ontology represents a certain view of the world, supports intentional queries regarding the content of database, and reflects the relevance of data by providing a declarative description of semantic information independent on the data representation.

New tools that can improve the retrieval and integration of information are emergent. A focus on hierarchies – a simpler, albeit very useful, version of ontologies is promising. Hierarchies are easier to understand, to implement, and their application to the extensions to searches, queries, and imperfect answers are straightforward. Ontologies promise longer mileage, although they are more complex to understand, to implement, and to apply.

A datum makes sense only within a context. Intuitively, we know that “computer” is closer to “office” than to “ocean” or to “dog.” A “cat” is closer to “dog” than to “bus station.” “Burning” is closer to “hot” than to “icy.” How can we measure these similarities? What wearing apparel do we wear for rainy days? *Raincoat* is a correct answer; *umbrella* is a close miss; *belt* a fair error, and *typewriter* a gross error. What is closer to an *apple*, a *pear* or a *caterpillar*? Can we measure these errors? How related or close are these words? An approach to modeling similarity between items from the same hierarchy or from different hierarchies can be based on an asymmetric, context-dependent measure called *confusion* (in using a qualitative value instead of the intended or correct value). The latter term is introduced to differentiate it from traditional approaches that used different kind of *distances* (i.e., symmetric, context-independent measures) for this purpose. The confusion’s asymmetry is given by definition and its context dependence by hierarchical structure. The concept of confusion allows defining the closeness to which an object fulfills a predicate as well as deriving other operations and properties among hierarchical values.

In environments with multiple information systems, independent systems may have their own intended models and, therefore, their own ontologies. In such environments, the general approach to data integration has been to map the local terms of distinct ontologies onto a single shared ontology. Then, the semantic similarity is typically determined as a function of the path distance between terms in the hierarchical structure underlying the single ontology. Other methods to assess semantic similarity within a single ontology are feature matching and use of information content. The feature-matching approach uses common and different characteristics between objects or entities to compute semantic similarity. Information content, on the other hand, uses information theory to define similarity measure in terms of the degree of informativeness of the immediate superconcept that subsumes the two concepts being compared.

The use of a single ontology does ensure complete integration across heterogeneous information systems. However, this type of ontology is costly, if not impractical to obtain, since users and information systems are forced to commit to this single ontology and compromises are difficult to maintain when new concepts are considered. Using another approach, which considers scalability issues in building an ontology, some studies create a shared ontology by integrating existing ones. Ontology integrations need to treat overlapping concepts and inconsistencies across ontologies. Like semantic heterogeneity in the database field, ontology mismatches occur when two ontologies have terms denoting categories, components of category definitions, or ontological concepts that are the same.

A strategy for ontology integration is the mapping of local ontologies onto a more generic ontology. The use of semantic interrelations is another approach for ontology integration. Once ontologies have been

integrated, similarity measures are applied to compare concepts in much the same way as a similarity evaluation is done within a single ontology.

In general, current methods that compare concepts from different ontologies are based on an a priori integration of local ontologies through a top-level ontology or through terminological relations that are defined manually or semi-automatically. However, more interesting and promising is a computational approach that compares concepts from unconnected and independent ontologies without constructing a priori a shared ontology. Its approach to modeling similarity is based on a matching process that uses available information from various ontology specifications (i.e., synonym sets, distinguishing features, and semantic relations of entity classes). Such similarity modeling establishes links among ontologies while keeping each ontology autonomous. This is a weak form of integration because it does not allow deep processes, that is, it cannot be used for making inferences about the relationship among other entity classes within a given ontology and cannot guarantee computations that require particular components of the entity class representation. It provides, however, a systematic way to detect which entity classes are most similar to each other and, therefore, which entity classes are the best candidates for establishing integration across the ontologies (*Levachkine, S. & Guzman, A.*).

Semantic mapping between different conceptualizations in a subject domain

A subject domain on the Web is characterized by vagueness that makes defining universally accepted ontology an onerous task, especially for semantic disambiguation of the concepts in that domain. This is compounded by the lack of appropriate methods and techniques where the individual semantic conceptualizations can be captured and compared to each other. Therefore, multiple user conceptualizations require personalization where user diversity can be incorporated. The most promising approach is the applications of commonsense reasoning to elicit and maintain models that represent users' conceptualizations. Such models will enable taking into consideration the users' perspective of the real world and empower personalization algorithms for the databases and queries processing. Intelligent information processing over the subject domain will be achieved if different conceptualizations are integrated in a semantic environment and mismatches between different conceptualizations outlined. A formal approach for automatic detection and resolution of mismatches between a user's and an expert's conceptual models is needed. This should allow users with different conceptual models, to meaningfully exploit the Web data repositories produced by different expert groups and entities and provide a unifying framework for search querying. Moreover, the resolution of mismatches will permit "automatic translation" between a user's query and the ontology of a given database. For instance, the user queries about "watercourse" will be translated to "rivers", "creeks" or "navigation way".

Web intelligence: Semantic services integration

Ontologies constitute the centrepiece of the knowledge retrieval, sharing, and reuse mechanisms either on the web or among agents.

Dynamic services development is based on existing infrastructures (e.g. *J2B*), service coordination mechanisms (e.g. *eCo*, *BizTalk*, *ebXML*), and semantic interoperability (e.g. *OWL*, *RDF*). From a syntactic point of view, web services are very promising and experienced tremendous growth because of their reliance on well-known standards. From a semantic point of view, there are several limitations because web services can currently be discovered based only on keywords (e.g. *UDDI*). Therefore, the ability for run-time discovery, a requirement for automatic service composition, is limited, though W3C develops ontological description of the web services (*OWL-S*).

An application of the ontology merging models to the semantic integration of services (provided by agents and current web services) in a subject domain is another challenging task. With OWL domain ontologies and OWL-S, there is a much greater chance of semantic interoperability for web applications within a technologically well-founded environment supporting high level, dynamic composition of

heterogeneous online agent services and their integration to web services. To do this, three basic elements are required: (i) a model for agent services that is compatible with the model of web services, (ii) ontological description enabling effective service composition, and (iii) service level interoperability.

The problem of agent service specification and composition and its integration with web services on the Web should be solved here, namely: a) an agent service model that will adopt the current standards (when and if possible) and b) ontology-driven service composition model for agent services. These lead to a new model for agent services and their composition inspired by works on workflow modeling techniques and web service orchestration for business process techniques.

CONCLUSION

Psychographic visual images as tool of communication appeared much earlier than language i.e., the abstract verbal form of the semantic-meaningful representation of the outer world. We attempted to highlight this fact by citing (*Toynbee, J.*). Development of traditional mathematical models has only led to numerical data processing. By inertia, similar approaches have been relegated to the computer data processing (e.g., a huge number of useful-less, pixel-oriented image processing approaches: the amount of data required to apply such approaches is often equivalent to the amount of the original data). There is no need to prove that recognition of the known painting requires less information than an attempt to recognize the unknown. Modern computer technologies (protocols, formats, etc) empirically demonstrate the necessity to identify different types of images, including cartographies, paintings, photos, chart-flows, etc. Morphologic classification could be useful here, but experiences great difficulties in computer analysis due to its weak formalization.

In the present chapter, we recall and show one possible approach to the problem: detection of semantic-meaningful components from information flows. These components can be different with regard to dependence on the problem under consideration.

In Table 1, we attempted to exhibit evolution of human processing of information flows and to put forward some analogs with machine processing. From our point of view, these analogs are quite correct because efficient human-machine interaction is the stumbling block of modern computer technologies.

The problem herein discussed is of great current interest (remember the now-popular image-processing slogan *Back to the intelligence!* since IJCAI 2003 www.ijcai-03.org). We conjecture that the most promising line of progress toward the solution of this problem lies in successively increasing automation of the separate links of the approach considered herein. We suggest as *the main principle* of such automation, the maximal use of data-semantic content. Indeed, semantic information can be optimally organized and effectively processed by a computer system.

The articles in the additional reading section illustrate our approach by describing the *computer-based systems* that encapsulate basic elements of semantic analysis and synthesis of crude data. These systems constitute symbolic language descriptions of objects of information flows rather than the traditional programs of data treatment and represent some advances on the problems discussed in the *Future research directions* section.

ACKNOWLEDGEMENTS

The authors of this chapter wish to thank Prof. Jean Serra <http://cmm.enscm.fr/~serra/aaccueil.htm> for very useful comments and especially for reminding us of Saint Augustine's words⁵. Our special thanks to Prof. Adolfo Guzmán-Arenas <http://www.cic.ipn.mx/aguzman/> for very fruitful discussions.

⁵ *Nec ipsa tamen intrans (in memoriam), sed rerum sensarum imagines, illic praesto sunt cogitationi reminiscenti eas. Quae, quomodo fabricatae sint? Quis dicit? conf. X-8*

REFERENCES

- Alexandrov, V. & Gorsky, N. (1991) *From humans to computers: cognition through visual perception*, Singapore, Singapore: World Scientific Publishing Co. Pte., Ltd.
- Alexandrov, V. & Gorsky, N. (1993). *Image representation and processing: a recursive approach*, Dordrecht-Boston-London: Kluwer Academic Publishers.
- Alexandrov, V. (1999). The eye and visual perception, *Journal of Optical Technology*, 66(9), 809-816.
- Arbib, M. (1964) *Brains, Machines and mathematics*, New York, USA: McGraw-Hill.
- Atkins, P. W. (1995). *The Periodic Kingdom*. HarperCollins Publishers, Inc.
- Bush, V. (1945). As We May Think. *Atlantic Magazine*, July 1945.
- Carroll, L. (1898). *Alice's Adventures in Wonderland*. UK: Macmillan.
- Dodgson, C.L. (1895). What the Tortoise Said to Achilles. *Mind*, 4 (1895), 278–80.
- Egenhofer, M. (2002). *Toward the Semantic Geospatial Web*. In Voisard, A. Shu-Ching Chen (Eds.), *Proceedings of the 10th ACM international symposium on Advances in geographic information systems GIS'02* (pp. 1 - 4), November 8-9, 2002: McLean, Virginia, USA.
- Engelbart, D. (1995). *Boosting Our Collective IQ: A Selection of Readings*.
<http://dougengelbart.org/pubs/books/augment-133150.pdf>
- Floridi, L. (2010). *Information: A Very Short Introduction*. Oxford, UK: Oxford University Press.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an eternal golden braid*, New York, USA: Basic Books.
- Kolmogorov, A. (1965). Three approaches to the definition of the notion of amount of information, pp. 184-193. In A.N. Shirayev (Ed.), *Selected Works of A.N. Kolmogorov, Volume III: Information Theory and the Theory of Algorithms*. Dordrecht-Boston-London: Kluwer Academic Publishers.
- Kosch, H. (2001). *MPEG7 and Multimedia Database Systems*.
<http://www.sigmod.org/publications/sigmod-record/0206/5.kosch2-new.pdf/view>
- Leibniz, G.W. (1666). *Dissertatio de arte combinatorial*. 1666, A VI 1, p. 163 *Sämtliche Schriften und Briefe* (Berlin: Akademie Verlag, 1923-)
- Levachkine, S. & Guzman, A. (2007). Hierarchies as a new data types for qualitative variables, *Journal of Expert Systems with Applications*, 32(2007) 899-910.
- Minsky, M. *The Society of Mind* (1988). New York, N.Y.: Simon and Schuster.
- Peano, G. (1889). *Arithmetices principia, nova methodo exposita* (The principles of arithmetic, presented by a new method), pp. 83–97. An excerpt of the treatise where Peano first presented his axioms, and recursively defined arithmetical operations.
- Simon, J.-C. (1984) *Patterns and operators. The foundations of data representation*, New York, USA: McGraw-Hill.
- Toynbee, J. (1996). *Comprehending history*, Moscow, Russia: Progress.
- Webster II (1984). New York, N.Y., USA: Berkley Books.
- Zipf, G.K. (1949). *Human behavior and the principle of least effort*, Cambridge, MA, USA: Addison Wesley Publishing Co., Inc.

ADDITIONAL READING SECTION

B

- P. Bautista. Solving arithmetic problems using words: Computing with Words. M. Sc. Thesis, CIC-IPN. In Spanish (2009).
- B. Bergamaschi, S. Castano, S. De Capitani di Vermercati, S. Montanari, and M. Vicini, “An intelligent approach to information integration”. In: Guarino, N. (ed.), *Proc. 1st Int. Conf. Formal Ontology in Information Systems*, Springer-Verlag (1998) 253-268
- T. Berners-Lee, J. Hendler, O. Lassila. (17 May 2001). The Semantic Web. *Scientific American*.
<http://www.scientificamerican.com/article.cfm?id=the-semantic-web>
- N.T. Bhin, A.M. Tjoa, and R. Wagner, “Conceptual multidimensional data model based on meta-cube”, *Lecture Notes in Computer Science*, Vol.1909 (2000) 24-31

Y. Bishr, "Semantic aspects of interoperable GIS", Wageningen Agricultural Univ. and ITC, The Netherlands (1997)

A. Botello, "Independent Databases Query Resolution by Partial Integration", PhD Thesis in progress, CIC-IPN, Mexico

M. Bright, A. Hurson, and S. Pakzad, "Automated resolution of semantic heterogeneity in multidatabases", *ACM Trans. Database Systems*, Vol. 19 (1994) 212-253

A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures". *Proc. North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, Pittsburgh, PA.

C

R. de Caluwe, "*Fuzzy and uncertain object-oriented databases: concepts and models*", World Scientific, 1997

C. Collet, M. Huhns, and W. Shen, "Resource integration using a large knowledge base in Carnot", *Computer*, Vol. 24 (1991) 55-62

F. Colorado. *Mapping words to concepts: disambiguation*. M. Sc. Thesis, CIC-IPN. In Spanish. (2008).

E

J. Everett and D. Bobrow, "Making ontologies work for resolving redundancies across documents", *Comm. ACM*, Vol. 45, No. 2 (2002) 55-60

F

P. Fankhauser and E. Neuhold, "Knowledge based integration of heterogeneous databases", In: Hsiao, D., Neuhold, E., and Sacks-Davis, R. (eds.), *Proc. IFIP WG2.6 Database Semantics Conf. Interoperable Database Systems*, Vol. DS-5 (1992) 155-175

G

A. Gangemi, D. Pisanelli, and G. Steve, "Ontology integration: Experiences with medical terminologies". In: Guarino, N. (ed.), *Proc. 1st Int. Conf. Formal Ontology in Information Systems*, Springer-Verlag (1998) 163-178

E. Godínez. *Development of a system for thematic analysis of scientific knowledge*. M. Sc. Thesis, CIC-IPN. In Spanish. (2009).

A. Goni, E. Mena, and A. Illarramendi, "Querying heterogeneous and distributed data repositories using ontologies". In Charrel, P.-J. and Jaakkola, H. (eds.), *Information Modeling and Knowledge Base IX*, IOS Press (1998) 19-34

N. Guarino, "Formal ontology, conceptual analysis, and knowledge representation", *Int. J. Human and Computer Studies*, Vol. 43 (1995) 625-640

N. Guarino, C. Masolo, and G. Verete, "OntoSeek: Content-based access to the web", *IEEE J. Intelligent Systems*, Vol. 14 (1999) 70-80

N. Guarino, "Formal ontology in information systems". In: Guarino, N. (ed.), *Proc. 1st Int. Conf. Formal Ontology in Information Systems*, Springer-Verlag (1998) 3-15

A. Guzman-Arenas, "Finding the main themes in a Spanish document", *J. Expert Systems with Applications*, Vol. 14, No.1/2 (1998) 139-148

A. Guzman-Arenas and J. Olivares-Ceja, "Finding the most similar concepts in two different ontologies", *Lecture Notes in Artificial Intelligence*, Vol. 2972 (2004) 129-138

A. Guzman-Arenas and S. Levachkine, "Graduated errors in approximate queries using hierarchies and ordered sets", *Lecture Notes in Artificial Intelligence*, Vol. 2972 (2004) 139-148

A. Guzman, J. Olivares, A. Demetrio, and C. Dominguez, "Interaction of purposeful agents that use different ontologies", *Lecture Notes in Artificial Intelligence*, Vol. 1793 (2000) 557-573

A. Guzman-Arenas, A.-D. Cuevas. "Knowledge accumulation through automatic merging of ontologies." *Journal Expert Systems with Applications* 37, 1991-2005 (2010).

A. Guzman-Arenas, A. Jimenez, "Obtaining the consensus and inconsistency among a set of assertions on a qualitative attribute." *Journal Expert Systems with Applications* 37, 158-164 (2010).

V.P. de Gyves and A. Guzman-Arenas, "A distributed digital text accessing and acquisition system: BiblioDigital", *Lecture Notes in Computer Science*, Vol. 3061 (2004) 274-283

V.P. de Gyves, Adolfo Guzman, S. Levachkine "Extending databases to precision-controlled retrieval of qualitative information", *Lecture Notes in Computer Science* Vol. **3563** (2005) 21-32

I

J. J. Ibarra-Vargas, "*Spatial Analysis in Conceptual Spaces*", Thesis, *CIC-IPN*, Mexico, Spring 2009 (in Spanish)

J

J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proc. Int. Conf. Computational Linguistics (ROCLING X)* (1997)

K

V. Kashyap and A. Sheth, "Semantic heterogeneity in global information systems: The role of metadata, context, and ontologies". In: Papazoglou, M. and Schlageter, G. (eds.), *Cooperative Information Systems: Tends and Directions*, (1998) 139-178

W. Kim and J. Seo, "Classifying schematic and data heterogeneity in multidatabase systems", *Computer*, Vol. 24 (1991) 12-18

L

J. Lee, M. Kim, and Y. Lee, "Information retrieval based on conceptual distance in IS-A hierarchies", *J. Documentation*, Vol. 49 (1993) 188-207

D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems*. Addison-Wesley, 1989

S. Levachkine and A. Guzman-Arenas, "Hierarchies measuring qualitative variables", *Lecture Notes in Computer Science*, Vol. 2945 (2004) 258-270

S. Levachkine, A. Guzman-Arenas, and V.P. de Gyves, "The semantics of confusion in hierarchies: Theory and practice", *Common Semantics for Sharing Knowledge*, Kassel University Press (2005), Germany

D. Lin, "An information-theoretic definition of similarity". *Proc. 15th Int. Conf. on Machine Learning (ICML 1998)*

Ping Luo, Hui Xiong, Guoxing Zhan, Junjie Wu, and Zhongzhi Shi, "Information Theoretic Distance Measures for Clustering Validation: Generalization and Normalization", *IEEE Transactions on Knowledge and Data Engineering*, Vol. **21**, No. 9, 1249-1262 (2009)

M

F. Martinez-Trinidad and A. Guzman-Arenas, "The logical combinatorial approach to Pattern Recognition, an overview through selected works", *Pattern Recognition*, Vol. **34** (2001) 741-751

F. Mata, "Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching". *Lecture Notes in Computer Science* Vol. **4853** (2007) 98-113

F. Mata, Serguei Levachkine, "iRank: Integral Ranking of Geographical Information by Semantic", Geographic, and Topological Matching. *Lecture Notes in Geoinformation and Cartography*, Springer-Verlag (2009) 77-92

F. Mata, "*iRank*: Ranking Geographical Information by Conceptual", Geographic and Topologic Similarity. *Lecture Notes in Computer Science* Vol. **5892** (2009) 159-174

R.B. McMaster and K.S. Shea, "*Generalization in Digital Cartography*", the Association of American Geographers, 1992

E. Mena, V. Kashyap, and A. Sheth, "OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies", *Proc. Int. Conf. Cooperative Information Systems (CoopIS 1996)* (1996)

N.E. Montoya-Irbe, "[Using Semantic Similarity Measures Processing Unstructured Information](#)", MS Thesis in progress, *CIC-IPN*, Mexico (in Spanish)

M.A. Moreno-Ibarra, Serguei Levachkine, Miguel Torres, Rolando Quintero, Giovanni Guzmán, Semantic Similarity Applied to Geomorphometric Analysis of Digital Elevation Model. *Lecture Notes in Geoinformation and Cartography*, Springer-Verlag (2009) 149-163

M.A. Moreno-Ibarra, Ph.D. in Computer Science, Centre for Computing Research, [Applying Similarity between the Systems of Geographical Objects to the Generalization of Vector Data](#), PhD Thesis, *CIC-IPN*, Fall 2007

N

N. Noy, R.W. Ferguson, and M.A. Musen, "The knowledge model of Protégè-2000: Combining interoperability and flexibility", *Stanford Medical Informatics Technical Report*, Stanford Univ. (2000)

O

J. Olivares-Ceja and A. Guzman-Arenas, "Concept similarity measures the understanding between two agents", *Lecture Notes in Computer Science*, Vol. **3136** (2004) 182-194

R

W. Renteria-Agualimpia, "[Precision-controlled Retrieval of Qualitative Information from Data Repositories](#)", MS Thesis, *CIC-IPN*, Mexico, Fall 2009 (in Spanish)

P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language", *J. Artificial Intelligence Research*, Vol. 11 (1999) 95-130

P. Resnik, "Disambiguating noun groupings with respect to WordNet senses", In: Armstrong, S. *et al.* (eds.), *Natural Language Processing Using Very Large Corpora*, Kluwer Academic Publishing: Dordrecht (1995) 77-98

I. Reyes-de los Santos, "[Personalization of Geographic Information Retrieval](#)", MS Thesis in progress, *CIC-IPN*, Mexico (in Spanish)

S. Ross, *A first course in probability*. New York: Macmillan, 1976

A. Rodriguez and M. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", *IEEE Trans. Knowledge and Data Engineering*, Vol. **15**, No. 2 (2003) 442-456

S

G. Sarabia-Lopez, "[Information Searching and Ranking in the Spatial Databases, using Hierarchies](#)", MS Thesis, *CIC-IPN*, Mexico, Fall 2008 (in Spanish)

L. Sarmiento-Castaneda, "[Intelligent Query Processing in Unstructured Data Repositories](#)", MS Thesis in progress, *CIC-IPN*, Mexico (in Spanish)

A. Sheth, "Changing focus on interoperability in information systems: From system, syntax, structure to semantics". In: Goodchild, M., Egenhofer, M., Fegeas, R., and Kottman, C. (eds.), *Interoperating Geographic Information Systems* (1999) 5-30

A. Sheth and V. Kashyap, "So far (schematically) yet so near (semantically)". In: Hsiao, D., Neuhold, E., and Sacks-Davis, R. (eds.), *Proc. IFIP WG2.6 Database Semantics Conf. Interoperable Database Systems*, Vol. DS-5 (1992) 283-312

A. Smeaton and I. Quigley, "Experiment on using semantic distance between words in image caption retrieval", *Proc. 19th Int. Conf. Research and Development in Information Retrieval (SIGIR 1996)* (1996)

T

A. Tversky, "Features of similarity", *Psychological Rev.*, Vol. 84 (1977) 327-352

V

P. Visser, D. Jones, T. Bench-Capon, and M. Shave, "Assessing heterogeneity by classifying ontology mismatches". In: Guarino, N. (ed.), *Proc. 1st Int. Conf. Formal Ontology in Information Systems*, Springer-Verlag (1998) 148-162

E. Voorhees, "Using WordNet for text retrieval", In: Fellbaum C. (ed.), *WordNet: An Electronic Lexical Database*, Cambridge, Mass.: The MIT Press (1998) 285-303

W

P. Weinstein and P. Birmingham, "Comparing concepts in differentiated ontologies", *Proc. 12th Int. Workshop Knowledge Acquisition, Modeling, and Management*, (1999)

Z

L.A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic". *Fuzzy Sets and Systems* **90**, 111-127. (2007).

R.W. Zagal-Flores, "[*Ontologies Alignment using Boosting Algorithm*](#)", MS Thesis, CIC-IPN, Mexico, Fall 2008 (in Spanish)