The AvesTerra Formalism

A Hyper/Ultra/Uber-Graph Approach to Extreme-Scale Knowledge Representation for Accelerated Multi-Disciplinary Scientific Discovery

Organization: **Georgetown University**
Type of Business: Other Educational

Technical POC: J. C. Smart, Ph.D.
Office of the Sr. VP for Research
Georgetown University
300 Gervase, 37th & O St, NW
Washington, DC 20057
Phone: 202-687-4816
Email: smart@georgetown.edu

October, 2018

GEORGETOWN UNIVERSITY

Introduction

In support of collaborative computational reasoning over large, diverse sources of evidence, Georgetown University has developed a knowledge representation framework of considerable semantic expressivity and scale. As an enabler for scientific discovery, this framework provides a method for integrating evidence from many sources – including mathematical, computational, experimental, and observational – spanning multiple modalities and levels of abstraction. This framework was specifically formulated to automate the fusion of disparate data types and reasoning methods to unify understanding and eliminate a significant scientific research bottleneck. The resulting representation approach enable connection of heterogeneous forms of scientific information to greatly accelerate the research process. This approach resolves the tension between domain-specific tools that require tuning to specific problems and general-purpose abstract methods that require idealization of the data, providing a robust unified foundation upon which automated, knowledge-driven application systems can be built. This paper provides a formal definition of this framework.

AvesTerra Formalism

The AvesTerra concepts prescribe a distributed analytic ecosystem for examining data, communicating information, and discovering, representing, and enriching shared knowledge. The AvesTerra formalism begins with the establishment of a conceptual boundary to whatever it is applied. This *system* may be a microscopic organism, the human body, a community, a continent, or the planet Earth.

Within a system's boundary, *data* is regarded as a collection of observations about this system. That is, given a complex system S, the (often unbounded) set $D = \{d_1, d_2, d_3, ...\}$ contains data about S where each $d_i \in D$ is viewed as a "scientific measurement" of some aspect of S. For example, an element d may be spectral decomposition of a chemical, a gene sequence of an organism, an e-mail message of an individual, the text of community's legal statute, or the seismic motion of a tectonic plate. When viewed in this manner, associated with every datum d is a measurement precision, measurement accuracy, and a frame of reference (e.g. spatial-temporal perspective) in which the measurement was made (i.e. data was collected). These items may be either explicitly stated or implicitly assumed, and may not always be imminently calculable. This treatment is important, however, so that subsequent data collection and analysis about whatever system is being studied rests on a scientific foundation.

For any given system S, AvesTerra next prescribes the notion of a *knowledge representation*. A knowledge representation K is a model of system S, which may likely be time varying and dynamic depending upon the complexity of S. Associated with every knowledge representation K is an *ontology* O that formally defines the semantics of all element types of K. To aid complexity reduction, and often as a practical and computational necessity, a knowledge representation K and its accompanying ontology O generally represent a simplification of S which may purposefully ignore various characteristics of S that may be deemed irrelevant or of little consequence to the particular line of scientific inquiry.

AvesTerra provides a collaborative approach for building shared knowledge representations for systems of high complexity and large scale. These representations are derived by leveraging large, distributed, and heterogeneous data spaces. Specifically, AvesTerra was formulated to enable knowledge representation at extreme scale, where the complexity of the system may span a diverse range of application domains (e.g. our planet and all of its inhabitants, or a human

organ and all of its cells). The AvesTerra formalism may be applied recursively in a hierarchical fashion and then dovetailed, yielding an even broader ecosystem of integrated knowledge representation components that span multiple levels of abstraction.

Within a particular set of system domains, the knowledge representation K for system S is constructed through an ongoing sequence of observations (i.e. data) and the application of scientific *theories* (i.e. analytics) about such observations. These theories define a collection of transformation operations $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, ...\}$ upon K. Characterized here as knowledge enrichment, these transformation operations constitute the knowledge extraction process. In essence, these extraction operations transform a system knowledge representation from one state to another, thereby formally capturing what was extracted. For example, a specific theory γ_α might define an initial state based on a particular data subset as in

$$\gamma_\alpha(\varnothing, d_i, d_j, d_k, ...) \Rightarrow K$$

whereas another theory γ_β might transform a knowledge representation's current state into a new state given new data as in

$$\gamma_\beta(K, d_i, d_j, d_k, ...) \Rightarrow K'.$$

Other theories may transform a knowledge representation simply based on the value of its current state without involving any data observations, as in

$$\gamma_\zeta(K) \Rightarrow K'.$$

At an even greater level of complexity, a theory might transform a knowledge representation based on knowledge representation of other systems as in

$$\gamma_\xi(K, K_X, K_{Y, ...}) \Rightarrow K'.$$

AvesTerra embraces all of these forms of transformations/extractions and many variants.

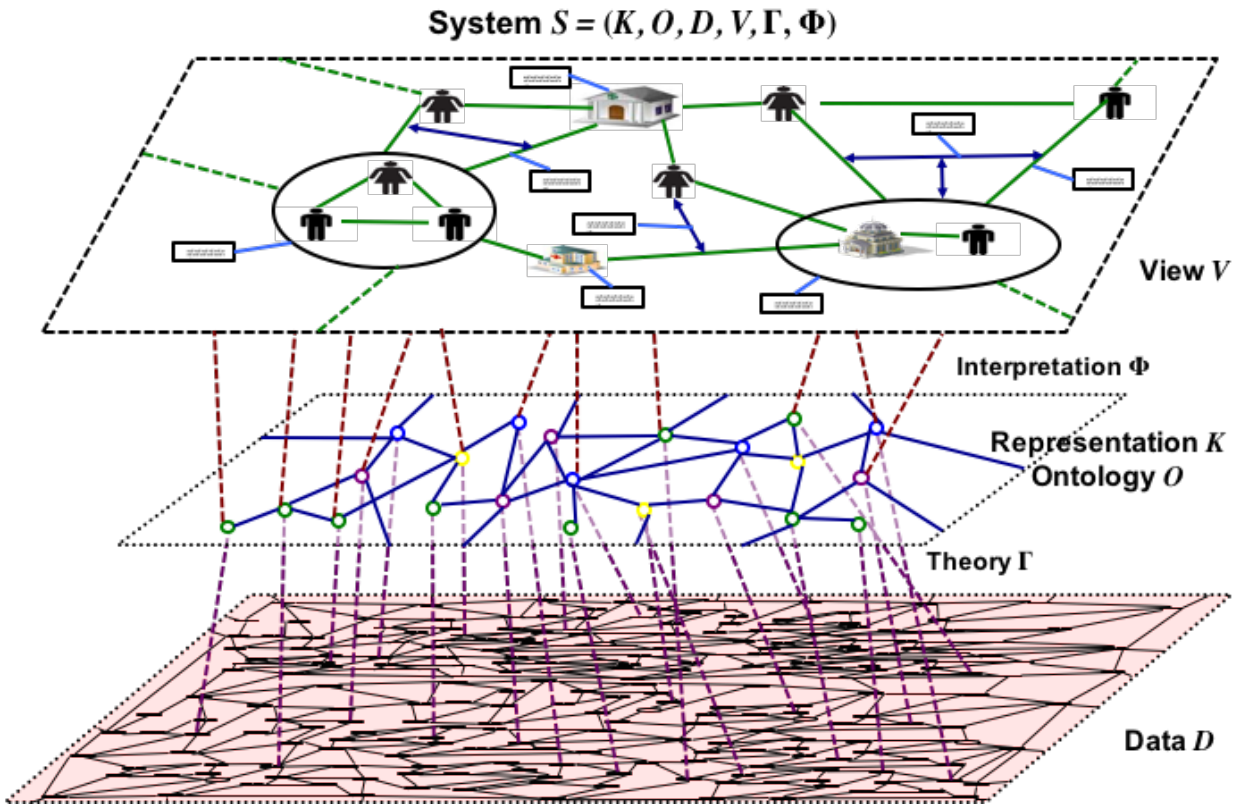
Given a knowledge representation K for system S, AvesTerra recognizes that there may exist multiple *interpretations* of K depending upon the specific domain. AvesTerra thus incorporates the notion of *views* $V = \{v_1, v_2, v_3, ...\}$ where each $v_i \in V$ is created via a function from the set $\Phi = \{\phi_1, \phi_2, \phi_3, ...\}$ that contains the various interpretations of K. Each interpretation function $\phi_i \in \Phi$ defines a mapping $\phi_i: K \rightarrow V$ that translates some subset of knowledge representation K into its respective view of K. This type of interpretation is again another form of knowledge extraction, but with the emphasis on enrichment of the end user(s) understanding, versus enrichment of the computer-based representation. AvesTerra imposes no particular mathematical restrictions on views or their respective interpretation functions. An interpretation of K may produce a view as simple as a range of scalar values or as complex as a multi-dimensional vector space. Alternatively, there may be a need for only one view with $K = V$ and Φ containing only the identity mapping. The view/interpretation formalism is provided so that the same system knowledge representation may offer varying perspectives in support of a wide range of end-user applications.

The blending of these various concepts is depicted in the figure below. The bottom layer of the diagram is a representation of the vast data space D for system S. The middle layer of the diagram contains the system knowledge representation K created as a result of the theory set Γ. At the top of the diagram is the view space V created as a result of the interpretation set Φ that results when applied to knowledge representation K.

In summary, AvesTerra mathematically portrays a complex system S as the 6-tuple

$$S = (K, O, D, V, \Gamma, \Phi).$$

That is, the process strives to create a unified knowledge representation K with ontology O of a complex system S that explains phenomena D via theory Γ, presenting results V via interpretation Φ. The scientific method provides a mechanism for testing the validity of theory Γ and the utility of knowledge representation K through experimentation, offering an evolutionary (or revolutionary) path to new models, new theories, new interpretations, and thus ultimately new discoveries.



AvesTerra Formalization

AvesTerra Knowledge Representation

The AvesTerra formalism uses an extremely powerful form of knowledge representation based on a highly generalized mathematical graph structure. In this formalism, a knowledge representation K is defined as a 3-tuple $K = (E, R, A)$ where $E = \{e_1, e_2, e_3, ...\}$ is a set of entities, $R = \{r_1, r_2, r_3, ...\}$ is a set of relationships between entities in E, and $A = \{a_1, a_2, a_3, ...\}$ is a set of attributes that are associated with elements of E and R. AvesTerra supports traditional (binary) directed and undirected graph relationship structures where $R \subseteq E \times E$. AvesTerra, however, allows much richer relationship structures including hypergraphs, ultragraphs, and ubergraphs, with the latter two forms constituting new constructs that emerged as a result of the DARPA SIMPLEX/AvesTerra Phase-I and Phase-II effort.

With the hypergraph formulation, R is generalized so that relationships may be n-ary. That is, $R \subseteq E \times E \ldots \times E$ or $R \subseteq \mathcal{P}(E)$, where $\mathcal{P}(E)$ is the power set of E. Ultragraphs generalize the notion of a relationship still further, allowing relationships between not only entities, but also with other relationships. Thus, $R \subseteq \mathcal{P}(\mathcal{P}(\ldots\mathcal{P}(E)))$. Finally, Ubergraphs generalize this one step further, allowing the relationships to be recursive. That is, an entity $e \in E$ may have relationships with other relationships, that have relationships with others, and so on, that ultimately may contain e itself. In simple terms, this is accommodated by allowing relationships to behave and be represented as entities themselves where $E \cap R$ need not be the empty set. Thus, at full generalization, entities and relationships mathematically become indistinguishable, and are differentiable only within the context of an accompanying ontology.

To accommodate most if not all contemporary underlying data organization, storage, and retrieval techniques including relational databases, object systems, indexing systems, graph systems, etc., the AvesTerra formalism includes a general attribute structure that enables collections of attributes to be affixed to any element of the knowledge representation. That is, an attribute $a \in A$ may be affixed to any entity $e \in E$ or $r \in R$. Instances of such attributes may range from a simple attribute taxonomy with accompanying scalar or aggregate values to rich sets of entities and relationships that have already captured within the representation. With this structure, any arbitrary graph, hypergraph, ultragraph, or ubergraph representation of a system can be constructed and appropriately annotated via the AvesTerra formalism.

References

J. C. Smart, et. Al, "AvesTerra: A Reference Architecture for Global-Scale Information Sharing and Analysis - Application Programming Interface (API)," Department of Computer Science, Georgetown University, October 2014; http://avesterra.georgetown.edu/sites/avesterra/files/documents/avesterra_api_may_2016.pdf

J. C. Smart, et. al, "The **FOUR**-Color Framework - A Reference Architecture for Extreme-Scale Information Sharing and Analysis," Department of Computer Science, Georgetown University, October 2014; http://avesterra.georgetown.edu/tech/4cf.pdf

J. C. Smart, "Privacy Assurance," International Engagement on Cyber, Georgetown Journal of International Affairs, 2011; http://avesterra.georgetown.edu/tech/privacy_assurance.pdf

J. C. Smart, "Rapid Information Overlay Technology (RIOT) - A Unifying Approach for Large Scale Analytics," Intelligence and Information Systems, Raytheon Company, July 2010.

J. C. Smart, "A Model for Extreme-Scale Knowledge Representation," NSA Report, January, 2002.

J. C. Smart, "High Yield Intelligence," NSA Cryptologic Quarterly, Summer/Fall 2002, CQ-EC-2002 02/03.

J. C. Smart, S. D. Pritchard, "The NSA Reference Model," NSA Cryptologic Quarterly, Winter 2000/Spring 2001, CQ-E05-00-04/01-10.

J. C. Smart, "Dependency Visualization for Complex System Understanding," Ph.D. Thesis, University of California, Davis, September, 1994.