LA-UR 97-2398

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: **SEMANTIC REPRESENTATIONS FOR COLLABORATIVE, DISTRIBUTED SCIENTIFIC INFORMATION SYSTEMS**

AUTHOR(S): M. Kantor, Cliff Joslyn

SUBMITTED TO: External Distribution -Hard Copy

DISCLAIMER

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

LDRD IP PROPOSAL

TITLE:

Semantic Representations for Collaborative, Distributed Scientific Information Systems

PI: Marianna Kantor, CIC-3, 5-8310, mkantor@lanl.gov

TECHNICAL CATEGORY: Mathematics and Computational Sciences (MCS)

FUNDING REQUESTED: $150,000 for each of three years

Abstract

It is vital for Los Alamos to respond to the challenge presented by the ongoing revolution in Information Science and Technology. Distributed Information Systems (DIS) are having a profound affect not only in science, but in society in general. In view of their increasing role in the management of scientific information, in national security and intelligence, and certainly as objects of scientific inquiry themselves, these DIS need to be designed and studied from a scientific perspective.

The technological developments over the last ten years, the Internet and the World-Wide Web (WWW) in particular, have been breakthroughs, allowing for the construction of non-linear, hypertextually based, DIS. And yet most of these DIS are still constructed by hand, and have the properties and architectures of the prior paradigm based on books and libraries, with strictly hierarchical categorization designed with many hours of human effort.

Our broader vision is based on an organismal model where DIS are adaptable and evolutionary, scalable, highly connected, high dimensional, resilient, and admitting to many complementary views and orderings. The key development necessary to support this view is the representation of *semantic* information in DIS.

We propose a set of software developments and experiments which will both construct novel DIS with explicit semantic representations, and measure the semantic properties of existing DIS.

For DIS design, we propose an architecture called Semantic Webs (SW), where a binary multigraph representation relates a number of nodes according to a variety of semantic categories, each partially ordered. The ontological relations among the semantic categories allows a dynamic among them, and thus for the DIS to be self-modifying and adaptive, suggesting new links as a form of inference.

These structures will be implemented as Java add-ons in existing browsers. Semantic categories will be represented as hypertext links, with type indicated by anchor color. A database representation of all link types and instances will augment and help drive the browser. For initial target applications, we are considering a number of vital Los Alamos projects, including the ASCI Problem-Solving Environment (PSE), the Scientific Data Management (SDM) project, and the Nuclear Weapons Archiving Project (NWAP).

For DIS analysis, we propose a series of measurement experiments on existing, intranet-based networks. We will apply the proven multivariate analysis techniques of Multi-Dimensional Scaling (MDS) and MAximum Likelihood Continuity Mapping (MALCOM) to derive semantic categories from relatively low-dimensional Euclidean models of traversal patterns of these nets.

When applied to untyped webs, these models of traversal patterns among nodes are used to introduce typed link representations, and thus to move to SW structures. But when these methods are applied to networks which are initially typed, then models of traversal through link-type space are produced, which are in turn used to augment or modify those typed link structures. This is another way in which these structures become self-modifying and adaptive through interaction with their users.

Our first application here is the extension of prior internet measurement experiments which used a simple heuristic technique to derive semantic categories as single-dimensional clusters of an untyped linguistic corpus based on a stochastic matrix representation of its traversal patterns.

After initial success there, we will apply these methods to the Los Alamos e-Print Archive for Physics, Mathematics, and Non-Linear Science. This is a much richer corpus for web analysis, and is specifically oriented to scientific problems and texts. Using a combination of traversal history and keyword analysis, we can not only derive semantic categories in this corpus, but also assess the value of the existing, limited semantic structures provided.

1 Introduction

The ongoing movement in Information Science and Technology (IS&T) is revolutionizing not just publishing, media, education, and business, but social communication and organization in general. These effects are also of concern for Los Alamos:

- In **science**, where there is a great need to support collaboration and communication among interdisciplinary groups; information is highly "multi-modal" (e.g. text, mathematics, images, databases, and computer programs); instruments and simulations alike are producing huge databases; and distributed networks are well established as a primary means of communication.
- In **security** and **intelligence**, where national and international information networks are increasingly seen as strategic assets; information warfare is becoming a primary focus; and security and intelligence needs are focusing more on the informational than the physical.

This situation presents scientific laboratories like Los Alamos with both great challenges and opportunities:

- To adapt this emerging medium *for the benefit of* the sciences by developing new IS&T architectures and solutions, not only to support the traditional work of active scientists, but also to develop new platforms *in which* that work can be done.
- To study the medium *as an object of* scientific inquiry, modeling its properties and their use both in society generally and in more specialized domains.

The current components of the revolution in Distributed Information Systems (DIS) technology are well known, and include electronic networking and publishing in general, and the Internet and the World-Wide Web (WWW) in particular. But as scientists, we need a vision beyond existing technology to look generally at the Web as just one example of the class of possible DIS.

Our model for DIS moves beyond the limitations of the HyperText Transfer Protocol (HTTP) and the HyperText Markup Language (HTML), which currently support the WWW, by introducing representations of *semantic* information. This will result in significant scientific benefits:

DIS Architectures: Substantial improvements to collaborative scientific development environments.

Analysis Tools: New methods for measuring the semantic content and structures of information networks, and enabling their adaptive self-modification in the context of interaction with their user communities.

Target applications include such important Los Alamos projects as the ASCI Problem-Solving Environment (PSE), the Scientific Data Management (SDM) project [16], the Nuclear Weapons Archiving Project (NWAP) [15], and the Los Alamos e-Print Archive for Physics, Mathematics, and Non-Linear Science [14].

2 Distributed Information Systems for the Next Century

Current directions in WWW development are primarily market-driven, rather than scientifically directed. Furthermore, little scientific investigation to *measure* the properties and dynamics of the WWW is ongoing.

2.1 New Capabilities

HTTP and the WWW offer important new capabilities, including "distributiveness", or the ability for users to be physically distant from each other in time and space; and some very modest capacity for collaboration. But we must look to design DIS which either have, or have the capacity to grow to include, the properties we desire. These new required abilities include:

Collaboratability: For users, either actively or passively, to collectively construct shared information spaces, structures, and processes, ultimately to the point of supporting the representation of different levels and extents of true *consensus*.

Extensibility: For open-ended architectures to allow augmentation of DIS structures and processes.

Adaptability and Evolvability: For the DIS to adjust to the needs and patterns of use of its environment (users and other DIS processes); and beyond that, for the DIS to dynamically suggest or develop new information content and structures.

Reflexivity: For the DIS to act back on itself such that the *content* of structures affects the *process* of their construction, and vice versa, in order to support self-extension and self-organization of the DIS.

Resilience: For the DIS to be fault-tolerant and noise-accepting, to be flexibly responsive to conflicting, incomplete, or uncertain information.

None of these capabilities are supported by either current WWW technology such as HTTP, or any proposed DIS architecture specific to the above Los Alamos projects.

2.2 Semantic Representations

Current conceptions of DIS are derived from prior technology: texts and books. We traditionally approach the WWW, for example, as a vast electronic library, and are content with hierarchical categorizations similar to the Dewey decimal, library of congress, or encyclopedic systems. Simple, relatively inflexible, textual keyword searches of the databases like Yahoo stand in for "knowledge engineering" in the current Web.

But our guiding metaphor is not that of a library, but rather of an *organism*, or indeed of the brain itself. Although composed of relatively simple components, these are vastly complex structures: they have high dimensionality, connectivity, and non-linearity; and multiple views, orderings (partial or total), and complementary paths available. They do not just *store* information, but are dynamic, adaptive, resilient, and decision-making.

For decades, cognitive and information scientists have been exploring much richer representational systems and frameworks which allow or approach the kinds of DIS structures we have in mind. And in fact, there is a natural marriage between the capabilities for non-linear, flexible, hypertext representations of current DIS technology and such forms as semantic networks [9], and thematic relations [12].

Key to all of these challenges, and the essential component missing from current technology, is the ability to represent **semantic** information: not just *where* information is, but also *what* it is, and more importantly, what it is *about* and how it is *used*. This makes it possible, for example: for the DIS *itself* to relate thematic content in otherwise separately authored nodes; for the construction of views or orderings for particular purposes in otherwise unstructured "corpi" (collections of information nodes); and ultimately for the net to be able to augment, construct, or delete certain nodes and connections based on their meaning and use.

Early DIS theorists were seriously concerned with semantic representations. More significantly, the tools and capabilities for such representations are well within reach. They are addressed to a certain extent in such early DIS standards as HyTime [8]. But in their zeal and success, early DIS implementations like HTML neglected developments in this direction, compromising long-term conceptual and architectural integrity for immediate commercial rewards.

3 Proposed Work

We propose to develop specific capabilities to construct systems with these attributes. The proposal has two complementary and interacting components:

Construction: Development of a software architecture to support construction of novel DIS with explicit semantic representations.

Analysis: A set of experiments to measure the properties of both these novel systems and existing information networks in terms of their semantic content.

In the following sections we outline our direction for both of these components, in the context of their application to the kinds vital, ongoing lab projects mentioned above.

3.1 DIS Design

First we need the ability to construct DIS with explicit representations of semantic content. The simplest and most direct architecture is that of a Semantic Web (SW) [7]: a set of **nodes** (usually primarily textual) with a number of sets of **links** within and among those nodes, where each set of links represents a different **link type**, and each link type corresponds to a given **semantic category**.

SW extend semantic networks. Mathematically, an SW is a binary multirelation on a set of nodes, with each relation representing a given semantic category. Furthermore, each relation is partially ordered, thus restricting each link type to be a directed acyclic graph. This allows each link type to be structured as a "loose hierarchy" [6], providing a sufficient degree of ordering to make the link types represent interesting structures, while still maintaining sufficient flexibility.

Graphically, imagine a map of a hypertext corpus connected by directed arcs representing links, and further that the links are labeled or colored. Thus nodes may be connected to each other in a variety of ways, each distinct way representing a distinct semantic category.

Obvious semantic categories include the standards used in semantic networks [9], for example *is-a, has-parts* and *cases*. But here we take "semantic category" more broadly to include at least:

Current WWW Links: The "standard" links in today's WWW, indicating a general subject relation, a sub-document relation, links to various web and `ftp` sites, etc.

Meta-Data: Information about the storage format and types of information stored in certain web nodes.

Argument Flow: Link types such as *supports, disputes, claims, refutes*, etc.

"Guides": Linear "cuts" describing a view through a corpus held by an author or group of authors.

Qualification Information: Such as the confidence, certainty, or precision of the information.

Scientific Support: Links to databases of experimental data, graphics, visualizations, etc.

Chronological: Tracing the temporal flow of a discussion, similar to the flow of articles within a subject thread in a newsgroup.

Navigational: Standard "buttons" on web pages such as up, down, next, previous, etc.

Publication Information: Such as bibliographic citation, links to authors other works, keyword links, links to similar publications.

Typed links are a significant feature of most mathematical hypertext models [8] and are of current interest in knowledge engineering [1, 11]. Typed links were specified in early HTML designs through the `LINK` tag, but were never implemented. Current standards discussions are aiming to revive this capability for a specific set of link types. Thus typed links are essentially unsupported in the current WWW.

As described, these semantic categories cover a broad range of "ontological" levels, from the general (A causes B) to the specific (the paper A cited the paper B). So a crucial next level is the representation of ontological relations among the semantic categories themselves. For example, let $P = (A$ cites $B)$, and $Q = (B$ precedes $A)$. An SW with P represented as a typed link, and augmented with knowledge that (P implies Q), should be able to construct the typed link Q.

More generally, a complex set of ontologically relations is available. When fully developed, this approaches current work in knowledge engineering [4, 10] and Object-Oriented design. These complex levels of semantic representations will give DIS the ability to truly approach the desired properties listed above, in particular by allowing new connections to be generated, and the net to become self-modifying.

We propose Java language add-ons to HTTP parsers and browsers to represent link types, and thus semantic categories, explicitly. At first a fixed menu of link types will be available, and indicated through colored anchors in a browser. This will be augmented by a database representation sufficient to support the more complex knowledge engineering structures.

Preliminary discussions are underway for potential collaboration with each of the PSE, SDM, and NWAP teams. A potential example of this architecture in action is shown in Fig. 1. Here a portion of an intranet system similar to those used in the PSE or SDM is represented as an SW. Nodes correspond to the different types of data used, for example descriptions of physical systems being modeled; mathematical problem types, frames, or specifications; data sets derived from instruments; data sets derived from simulations; and software codes. Links are arcs among the nodes, and are typed by their labels.

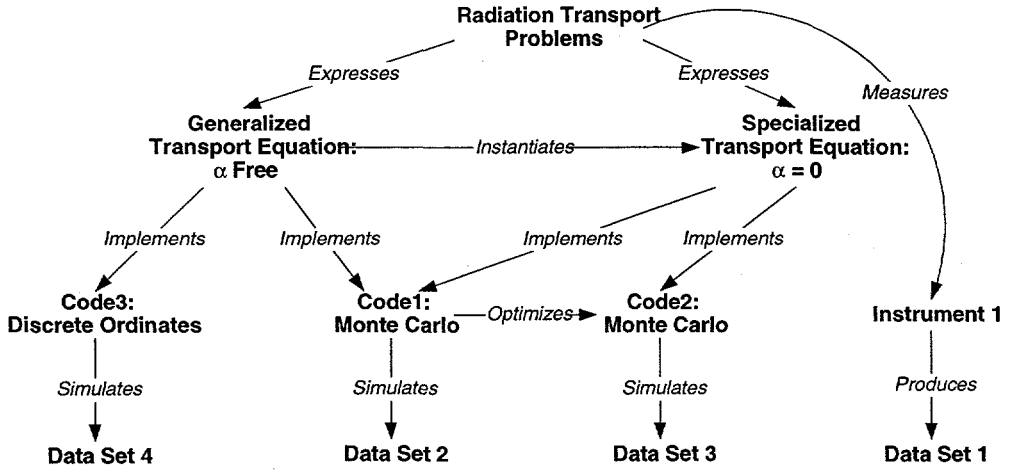

Figure 1: A portion of the PSE or SDM as a semantic web.

This type of representation is extremely valuable for management and navigation as the system is scaled up to include hundreds of problem domains, codes, and data sets which are related in a number of ways. First, the sheer complexity of these interelated components is more effectively captured when their relations are specifically encoded

within the DIS itself. For example, the information that some mathematical frames are specializations, some codes are optimizations, etc., will be available to the users and managers.

But in addition, this system can result in the generation of new information, for example the ability of a particular code to be used for a new problem type, or the superiority of one implementation over another.

3.2 DIS Analysis

In the design phase, semantic categories are *constructed on* information networks, resulting in a net which is semantically structured from the beginning. But in the analysis phase, a semantically unstructured net is assumed, and by examining the patterns of use of this corpus in an experimental context, semantic categories and link types will be *induced from* the net.

The best example to date of this kind of work is that of Bollen and Heylighen [2, 13]. They used a small corpus of the hundred most common English nouns in newspapers, with weighted links having small, randomly distributed values. A WWW page displays a word, following by an list of linked words in weight order, so that the weights represent the likelihood of being presented with candidate links. The users then navigate among these words, and their collective traversed paths were tracked. Well-visited paths are rewarded and less well-visited paths punished.

The result is a stochastic matrix with a specific structure. Clustering on this matrix then reveals disjoint regions of semantic coherence. In addition, this method is an example of true adaptability in a DIS: this ability *reject* information is a missing feature of today's databases.

Our approach will move beyond this application in both the analysis tools and the web corpus studied. First, we have confirmed that we will be able to collaborate with Dr. Heylighen and his colleagues. We will use their data, but with more sophisticated methods to produce relatively low-dimensional Euclidean models of the semantic space of the corpus. In this context the Euclidean dimensions of the model are interpreted as semantic categories, and the relative positions of the nodes as their participation in those categories. From this data, link types will then be induced based on multidimensional clustering.

An example is shown in Fig. 2. In Bollen and Heylighen's one-dimensional clustering, "influence", "control", and "power" were in one cluster, and "change" and "time" each in distinct clusters. Our approach not only allows terms like "control" to participate in multiple clusters, it also adds information from the Euclidean metric to characterize semantic distance.

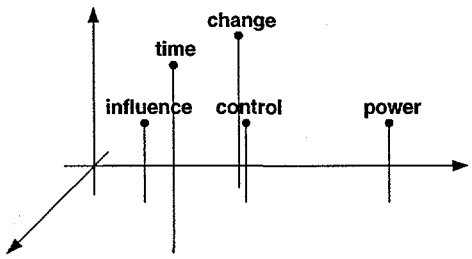

Figure 2: Euclidean model of a portion of the noun-based corpus.

The first method attempted will be Multi-Dimensional Scaling [3]. Given a dimensionality N and a set of inter-point distances, MDS produces rotationally invariant Euclidean N-space representations. However, MDS uses only pair-wise information about points.

A more sophisticated approach which can represent more context-dependence is the MAximum Likelihood Continuity Mapping (MALCOM) system [5], a time-series analysis technique which estimates the probability of sequences of categorical data in order to predict future data. MALCOM models the probability of sequential symbol production as a point moving through a low-dimensional Euclidean space along a "smooth" path (one with no Fourier components above a given cutoff frequency).

These constraints (a smooth path in a Euclidean space) provide MALCOM with significant advantages over similar methods, such as N-grams, which require vastly more parameters to be estimated. MALCOM has had considerable success analyzing categorical databases derived from such diverse applications as speech output, for modeling speech production; and medical histories, for detecting insurance fraud.

Second, beyond the simple body of English words used by Bollen and Heylighen, we have confirmed our ability to use the Los Alamos e-Print Archive for Physics, Mathematics, and Non-Linear Science [14] as a test corpus. Since 1991 this very succsessful DIS project has resulted in the construction of a large corpus of scientific papers. It is used daily by over 50,000 working scientists, and is a rich source of information about the semantic content of a number

of scientific fields. Both MALCOM and multidimensional scaling will be applied to sequences of nodes traversed in the e-Print database, resulting in a semantic mapping of this space.

Finally, we propose application of both MDS and MALCOM to analysis at a higher level: constructing models of traversals through *link type* space rather than *node* space. This must be carried out either on constructed, typed nets as described in Sec. 3.1, on nets with induced semantic categories as described here, or on nets with minimal semantic representations (for example, navigational buttons only).

By constructing models of the structure of the link space, existing semantic typologies can be analyzed and modified. Iteration through this process will affect a very important feedback cycle between the design and analysis of the semantic structures, and facilitate another route to the self-modification of these DIS.

References

[1] Baron, L., Taguesutcliffe J, Kinnucan Mt, And Carey T: (1996) "Labeled, Typed Links as Cues When Reading Hypertext Documents", *J. American Society For Information Science*, v. 47:**12**, pp. 896-908

[2] Bollen J., and F. Heylighen: (1996) "Algorithms for the Self-Organisation of Distributed, Multi-user Networks", in: *Cybernetics and Systems '96*, ed. R. Trappl, Austrian Society for Cybernetics, p. 911-916.

[3] Dillon, W.R., and M. Goldsten: (1984) *Multivariate Analysis*, Wiley, 1984.

[4] Heylighen, Francis: (1991) "Structuring Knowledge in a Network of Concepts", in: *Workbook of the First Principia Cybernetica Workshop*, ed. F. Heylighen, pp. 52-57, PrincipiaCybernetica, Brussels, NY

[5] Hogden, J.: (1997) "Speech Processing Using Maximum Likelihood Continuity Mapping", Provisional Patent Application for the University of California.

[6] Joslyn, Cliff: (1991) "Hierarchy, Strict Hierarchy, and Generalized Information Theory", in: *Proc. ISSS 1991*, v. **1**, pp. 123-132

[7] Joslyn, Cliff: (1996) "Semantic Webs: A Cyberspatial Representational Form for Cybernetics", in: *Cybernetics and Systems '96*, ed. R. Trappl, Austrian Society for Cybernetics, pp. 905-910.

[8] Newcomb, Steven R; Kipp, Neil A; and Newcomb, Victoria T: (1991) "HyTime: Hypermedia/Time-Based Document Structuring Language", *Communications of the ACM*, v. **34**:11, pp. 67-83

[9] Sowa, J.F., ed.: (1991) *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, Morgan Kauffmann, San Mateo, CA.

[10] Uschold, M., and Gruninger, M.: (1996) "Ontologies: Principles, Methods, and Applications", *Knowledge Engineering Review*, v. **11**:2, pp. 93-136.

[11] Wang W, And Rada R (1995): "Experiences With Semantic Net Based Hypermedia", *Int. J. Human-Computer Studies*, v. 43:**3**, Pp. 419-439

[12] Wilkins, W., ed.: (1988) *Thematics Relations*, Academic Press.

[13] `http://pespmc1.vub.ac.be/adapnet.html`, `http://pespmc1.vub.ac.be/ADHYPEXP.html`

[14] `http://xxx.lanl.gov`

[15] `http://www.lanl.gov/projects/nwap`

[16] `http://www.lanl.gov/projects/SDM/sdmHome.html`

Cliff Joslyn

Postdoctoral Associate, CIC-3, Los Alamos National Laboratory. Founding editor of the Principia Cybernetica internet project for the development of collaborative, distributed, hypertext corpi on information science and cybernetic philosophy (`http://pespmc1.vub.ac.be`).

Research Interests: Uncertainty modeling, data mining, Systems Science, General Information Theory, qualitative modeling, semiotic systems theory, and cybernetic philosophy.

Education:

- PhD 1994 in Systems Science, SUNY at Binghamton. Dissertation: *Possibilistic Processes for Complex Systems Modeling.* Advisor: George Klir
- MS 1989 in Systems Science, SUNY at Binghamton.
- BA 1985 in Mathematics and Cognitive Science, with High Honors, Oberlin College. High Honors: Cybernetics and Cognitive Science.

Experience:

- NSF Postdoctoral Research Associate, Software and Automation Systems Branch, NASA Goddard Space Flight Center, 1994-1996.

  Research projects: Possibilistic Qualitative Model-Based Diagnosis and Trend Analysis of Spacecraft Systems; Data Analysis and Systems Modeling Environment (DASME) for mixed Discrete EVent (DEVS) and discrete-time modeling. C++, UNIX, X/Motif.
- Graduate Fellow, NASA Goddard Space Flight Center, 1991-1994. Possibilistic Processes for Complex Systems Modeling
- Instructor and Teaching Assistant, SUNY at Binghamton, 1987-1991. Systems Science Departments and Continuing Education.
- Software Consultant: ABB Environmental, Portland, Maine, Summer 1994.
- Computer Consultant: 1987-present, small business information systems design and development.
- Computer Manager: Pryme-Line Distributors, Binghamton, New York, 1988-1991.
- Software Engineer: Computer Consoles Inc., Reston, Virginia, 1986-1987.

Selected Publications:

- Heylighen, Francis, and Joslyn, Cliff: (1993) "Electronic Networking for Philosophical Development in the Principia Cybernetica Project". *Informatica*, v. 17:3, pp. 285-293
- Joslyn, Cliff: (1995) "Semantic Control Systems", *World Futures*, v. 45, pp.87-123
- Joslyn, Cliff, and Luis Rocha: (1997) "Towards a Formal Taxonomy of Hybrid Uncertainty Representations", *Information Sciences*, under review.
- Heylighen, Francis and Joslyn, Cliff: (1995) "Systems Theory", in: Cambridge Dictionary of Philosophy, ed. R. Audi, pp. 784-785, Cambridge U. Press, Cambridge MA
- Joslyn, Cliff: (1996) "An Object-Oriented Architecture for Possibilistic Models", in: *Computer-Aided Systems Technology*, ed. T. Oren and G. Klir, pp. 80-94, Springer-Verlag, Berlin, in series: Lecture Notes in Computer Science # 1105
- Joslyn, Cliff: (1997) "Semiotic Aspects of Control and Modeling Relations in Complex Systems", to appear in: *Control Mechanisms for Complex Systems*, ed. Michael Coombs, in series: *Santa Fe Institute Studies in the Sciences of Complexisty*, Addison-Wesley, Redwood City CA.
- Joslyn, Cliff: (1996) "Semantic Webs: A Cyberspatial Representational Form for Cybernetics", in: *Cybernetics and Systems '96*, ed. R. Trappl, Austrian Society for Cybernetics, pp. 905-910.
- Joslyn, Cliff: (1996) "Hybrid Methods to Represent Incomplete and Uncertain Information", in: *Proc. 1996 Interdisciplinary Conference on Intelligent Systems: A Semiotic Perspective*, ed. James Albus and Alex Meystel, pp. 133-140, NIST, Gaithersburg, MD.