

# Content and the Internet, Part 1

by

Del Jensen and Steve Carter

## Summary

The Internet is about *content*. Content being accessed, published, indexed, analyzed, secured, purchased, stolen, vandalized, etc. Whether the content is white-papers, on-line books, catalogs, real-time games, address books, streaming audio and video, etc. it is content that people and cyber-agents are seeking. The future of the Internet lies not in bandwidth or capacity, but rather the ability to retrieve relevant content. Technology that allows fast and accurate access to relevant content will be used by the masses of carbon and silicon Internet users. Not because it is a better mouse-trap, but because controlled access to relevant content will allow the Internet to thrive, survive, and continue it's explosive growth. Fast and accurate semantic access to Internet content will determine who rules the next Internet era.

This white paper describes the formal basis for semantic content characterization. The discussion herein lays the foundation for the development of a framework that provides a description of a formal metric space. This will allow the description, characterization, and manipulation of the semantics of any content source.

As the paper shows, semantic characterization of content will readily fit in the described metric

The Internet is about *content*. . .

space thus allowing the application of optimal mechanisms for measuring and comparing semantics.

## Introduction

Caught between the sheer (and ever growing) volume of content, the huge and rapidly increasing number of Internet users and a growing sophistication in the demands of those users, the current TCP/IP infrastructure and architecture is showing its inadequacies - it is a victim of its own success. One of the many strategies under consideration by the Internet community for redressing these inadequacies is to *build intelligence* into the network. Directory Services and Caching are two prime examples of intelligent network components. Adaptive routing with route caching is another example of an intelligent network component.

Yet another example of network intelligence that is receiving close attention these days is the characterization of content by its meaning (semantics). This is a very ambitious goal. Indeed, it is the kind of undertaking that consumes vast amounts of resources. Even so, the obvious advantages that accrue with even a moderately successful semantic characterization component are such that almost everyone is tempted to dip a toe in the water.

In this paper we explore some of the issues associated with semantic representations. We outline what might be some of the basic components of a *semantic server*.

## Content and Meaning

Whether the data expressing content is encoded as text, binary code, bit map or in any other form, there is a vocabulary that is either explicitly (such as for code) or implicitly (as for bitmaps) associated with the form. The vocabulary is more than an arbitrarily ordered list: an element of a vocabulary stands in relation to other elements, and the "place" of its standing is the *semantic value* of the element. For example, consider the scene in figure 1. We can easily parse the major visual elements in the image. In particular, consider the spoon. If we were to compare the spoon to something taken from another scene - say, a shovel - we might classify the two items as being somewhat similar. And to the extent that form follows function in both nature and human artifice, we would be right! Likewise if we compared the spoon to a ladle. All three visual elements - the spoon, the shovel and the ladle - are *topologically invariant*. In addition, each element can be transformed into the other two elements with relatively little geometric distortion.



We outline . . . the basic components of a *semantic server* . . .

What happens when we compare the spoon to a fork? Curiously enough, both the spoon and the fork are topologically equivalent. But if we compare the ratio of boundary to surface area we note a distinct contrast. In fact, the attribute (boundary)/(surface area) is a crude analog

of the *fractal dimension* of the element boundary.

### *Iconic Representation*

Fractal dimension possesses a nice linear ordering. For example, a space-filling boundary such as a convoluted coastline (or a fork!) would have a higher fractal dimension than, say, the boundary of a circle. Can we characterize the topology of an element in the same way? In fact, we *can* assign a topological measure to the vocabulary elements, but the measure may involve aspects of homotopy and homology that preclude a simple linear ordering. Let us suppose, for visual simplicity, that we have some simple, linearly ordered way of measuring the topological essence of an element. We can formally represent an *attribute space* for the elements as in figure 2.

As can be seen, *fork-like* and *spoon-like* resolve to different regions in the attribute space. In this case, one might adopt the standard Euclidean metric for  $\mathbb{R}^2$  and thus have a well-defined notion

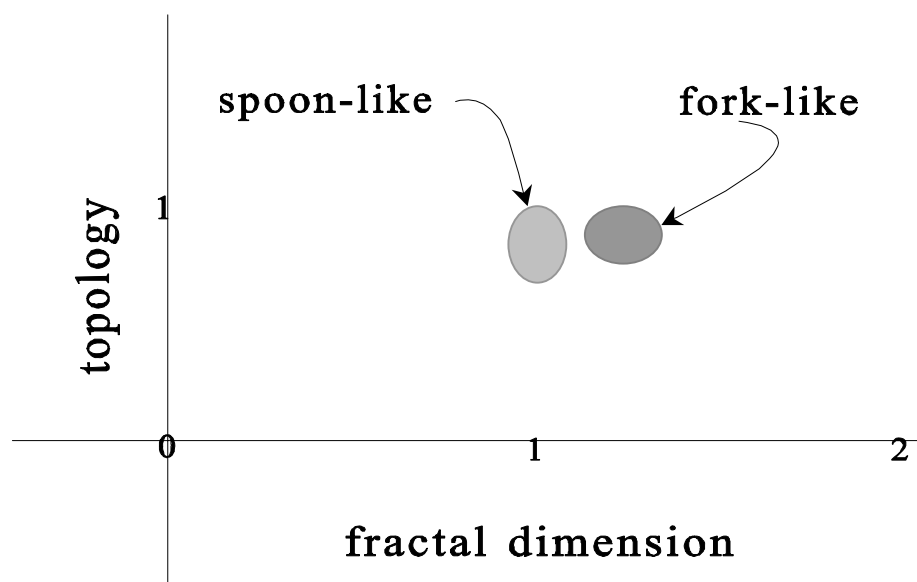
of *distance* in attribute space. Of course, one must buy into all the hidden assumptions of the model. For example, is the orthogonality of the two attributes justified, i.e., are the attributes truly independent?

Our example attribute space is a (simplistic) illustration of a semantic space, also known as a *concept space*. In our example, we were concerned with a vocabulary for human visual elements: a kind of *visual lexicon*. In fact, many researchers have argued for an *iconic representation* of meaning, particularly those looking for a representation unifying perception and language. They take an empirical positivist position that *meaning* is simply an artifact of the "binding" of language to perception, and point out that all writing originated with pictographs (even our "A" is just an inverted ox head!). With the exception of some very specialized vocabularies, it is an unfortunate fact that most iconic models have fallen well short of the mark. What is the visual imagery for the word "maybe"? For that matter, our own example iconic model has shown how spoons and forks are different, but how does it show them to be the same (i.e., cutlery)?

### *Propositional Representation*

Among computational linguists, a leading competitive theory to iconic representation is *propositional representation*. A proposition is typically framed as a pairing of *arguments* and *predicates*. For example, the fragment "a red car" could be represented propositionally as the argument "a car" paired with the predicate "is red". The proposition simply asserts a property (the predicate) of an object (the argument). In our example, stipulating the argument alone has consequences; "a car" invokes the existential quantifier, and asserts instances for all relevant primitive attributes associated with the lexical element "car."

How about a phrase such as "every red car?" Taken by itself, the phrase asserts nothing - not even existence! It is a null proposition, and can be safely ignored. What about "every red car has a radio?" This is indeed making an assertion of sorts, but it is asserting a property of the



semantic space itself; i.e., it is a meta-proposition. One can not instantiate a red car without a radio, nor can one remove a radio from a red car without either changing the color or losing the "car-ness" of the object.

At this point you should begin to suspect the preeminent role of the predicate, and indeed you would be right to do so. Consider the phrase, "the boy hit the baseball."

nominative: the boy  $\mathfrak{G}$  (is human),(is ~adult),(is male), (is ~infant), etc.  
predicate: (hit the baseball)  $\mathfrak{G}$   
verb: hit  $\mathfrak{G}$  (is contact),(is forceful),(is aggressive), etc.  
d.o.: the baseball  $\mathfrak{G}$  (is round),(is leather),(is stitched), etc.

We have transformed the phrase into two sets of attributes: the nominative attributes and two subsets of predicate attributes (verb and object). This suggests stipulating that all propositions must have the form (n: n ON, p: p OP), where N (the set of nominatives) is some appropriately restricted subset of  $2^P$  (the power set of the space P of predicates). We restrict N to avoid things like (is adult) and (is ~adult). In this way we can use the predicates to generate a semantic space. We might even hope to be able to semantically represent something like, "The movie *The Boy Hit the Baseball* hit this critic's heart-strings!"

Given that we can resolve propositions to sets of predicates, the way forward becomes more clear. If we were to characterize sets of predicates as *clusters* of points in an attribute space along with some notion of *distance*

between clusters, *we could quantify how close any two propositions are to each other!*. This is the Holy Grail.

. . . we could quantify how close any two propositions are to each other!

Before leaving this section we should note that another useful feature of the propositional model is *hierarchy of scope*, at least at the sentence level and below. Consider the phrase, "the boy hit the spinning baseball." The first tier proposition is "x hit y." The second tier propositions are "x is-a boy," and "y is-a baseball." The third tier proposition is "y is spinning." By restricting the scope of our view into the semantic space, we can focus on "hitting," "hitting spinning things," "people hitting things," etc.

#### *Hyponymy & Meaning Postulates - Mechanisms for Abstraction*

Two elements of the lexicon are related by *hyponymy* if the meaning of one is included in the meaning of the other. For example, the words *cat* and *animal* are related by hyponymy. A cat is an animal, and we say that *cat* is a hyponym of *animal*.

A particular lexicon may not explicitly recognize some hyponymies. For example, the words *hit*,

*touch, brush, stroke, strike* and *ram* are all hyponyms of the concept "co-incident in some space or context." We call such a concept a *meaning postulate*, and we extend the lexicon with the meaning postulate in order to formally capture the hyponymy.

Note that the words *hit* and *strike* are also hyponyms of the word *realize* in the popular vernacular. Thus we see that lexical elements can surface different hyponymies depending on the inclusion chain that is followed.

### *Topological Considerations*

We now turn our attention to the *metrization problem*: how do we determine the distance between two propositions? We do not propose a specific metric in this paper, but we do outline an approach in order to demonstrate the plausibility of doing so.

. . . how do we determine the distance between to propositions?

Many people begin by identifying a set  $S$  to work

with (in our case,  $S = P$ , the set of predicates), and define a *topology* on  $S$ . A topology is a set  $O$  of subsets of  $S$  which satisfies the following criteria:

- ! Any union of elements of  $O$  is in  $O$ ,
- ! Any finite intersection of elements of  $O$  is in  $O$ ,
- !  $S$  and the empty set are both in  $O$ .

The elements of  $O$  are called the *open* sets of  $S$ . If  $X$  is a subset of  $S$ , and  $p$  is an element of  $S$ , then  $p$  is called a *limit point* of  $X$  if every open set that contains  $p$  also contains a point in  $X$  distinct from  $p$ .

Another way to characterize a topology is to identify a *basis* for the topology. A set  $B$  of subsets of  $S$  is a basis if

- !  $S =$  the union of all elements of  $B$ ,
- ! for  $p \in b_\alpha \cap b_\gamma$ , ( $b_\alpha, b_\gamma \in B$ ), there exists  $b_\lambda \in B$  such that  $p \in b_\lambda$  and  $b_\lambda \subseteq b_\alpha \cap b_\gamma$ .

A subset of  $S$  is open if it is the union of elements of  $B$ . This defines a topology on  $S$ . Note that it is usually easier to characterize a basis for a topology rather than to explicitly identify all open sets. The space  $S$  is said to be *completely separable* if it has a countable basis.

It is entirely possible that there are two or more characterizations that yield the same topology. Likewise, one can choose two seemingly closely related bases which yield nonequivalent topologies. As the keeper of the Holy Grail said to Indiana Jones, "Choose wisely!"

The goal is to choose as *strong* a topology as possible. Ideally we are looking for a compact metric space. We look to

[We] choose as strong a topology as possible.

satisfy a minimum number of separability conditions such that the space  $S$  is guaranteed to be *homeomorphic* to a subspace of Hilbert space. One can then adopt the Hilbert space metric. Failing this, we impose as much structure as we can. To this end, consider the following axioms (the so-called "trennungaxioms").

- !  $T_0$ . Given two points of a topological space  $S$ , at least one of them is contained in an open set not containing the other.
- !  $T_1$ . Given two points of  $S$ , each of them lies in an open set not containing the other.
- !  $T_2$ . Given two points of  $S$ , there are *disjoint* open sets, each containing just one of the two points (Hausdorff axiom).
- !  $T_3$ . If  $C$  is a closed set in the space  $S$ , and if  $p$  is a point not in  $C$ , then there are disjoint open sets in  $S$ , one containing  $C$  and one containing  $p$ .
- !  $T_4$ . If  $H$  and  $K$  are disjoint closed sets in the space  $S$ , then there are disjoint open sets in  $S$ , one containing  $H$  and one containing  $K$ .

We should note that a set  $X$  in  $S$  is said to be *closed* if the complement of  $X$  is open. Since it isn't our intention to take the reader through the equivalent of a course on topology, we must be content simply to point out that the distinctive attributes of  $T_3$  and  $T_4$  spaces are important enough to merit a place in the mathematical lexicon -  $T_3$  spaces are called regular spaces, and  $T_4$  spaces are called normal spaces - and to state the following very beautiful theorem:

- Every completely separable regular space can be imbedded in a Hilbert coordinate space.

So, if we can demonstrate a countable basis for  $S$  that satisfies  $T_3$ , then  $S$  is metrizable. We denote the metrized space  $S$  as  $(S, d)$ .

Finally, consider  $\mathcal{K}(S)$ , the set of all compact (non-empty) subsets of  $(S, d)$ . Note that for  $u, v \in \mathcal{K}(S)$ ,  $u \cup v \in \mathcal{K}(S)$ ; i.e., the union of two compact sets is itself compact. We define the pseudo-distance  $\xi(x, u)$  between the point  $x \in S$  and the set  $u \in \mathcal{K}(S)$  as

$$\xi(x, u) = \min\{d(x, y) : y \in u\}.$$

Using  $\xi$  we define another pseudo-distance  $\lambda(u, v)$  from the set  $u \in \mathcal{K}(S)$  to the set  $v \in \mathcal{K}(S)$ :

$$\lambda(u, v) = \max\{\xi(x, v) : x \in u\}.$$

Note that in general it is *not* true that  $\lambda(u, v) = \lambda(v, u)$ . Finally, we define the distance  $h(u, v)$  between the two sets  $u, v \in \mathcal{K}(S)$  as

$$h(u, v) = \max\{\lambda(u, v), \lambda(v, u)\}.$$

The distance function  $h$  is called the *Hausdorff* distance. Since

1.  $h(u, v) = h(v, u)$ ,
2.  $0 < h(u, v) < 4$  for all  $u, v \in \mathbf{S}, u \neq v$ ,
3.  $h(u, u) = 0$  for all  $u \in \mathbf{S}$ ,
4.  $h(u, v) \neq h(u, w) + h(w, v)$  for all  $u, v, w \in \mathbf{S}$ ,

we can now form the metric space  $(\mathbf{S}, h)$ . The completeness of the underlying metric space  $(\mathbf{S}, d)$  is sufficient to show that every Cauchy sequence  $\{u_k\}$  in  $(\mathbf{S}, h)$  converges to a point in  $(\mathbf{S}, h)$ . Thus,  $(\mathbf{S}, h)$  is a *complete* metric space.

If  $\mathbf{S}$  is metrizable, then we claim it is  $(\mathbf{S}, h)$  wherein lurks that elusive beast, *semantic value*. For, consider the two propositions,  $\rho_1 = (n_1, p_1), \rho_2 = (n_2, p_2)$ . Then we can define the *nominative distance*  $|n_2 - n_1|$  as  $h([n_1], [n_2])$ , where  $[n]$  denotes the closure of  $n$ . We can define the *predicate distance* likewise. Finally, we might define  $|\rho_2 - \rho_1| = (|n_2 - n_1|^2 + |p_2 - p_1|^2)^{1/2}$ , or we might use "city block" distance  $|\rho_2 - \rho_1| = |n_2 - n_1| + |p_2 - p_1|$ . Or whatever appropriate metric.

The reader may recognize  $(\mathbf{S}, h)$  as the *space of fractals*. Some compelling questions come immediately to mind. Might we be able to find *submonoids of contraction mappings* corresponding to *related* sets in  $(\mathbf{S}, h)$ ; related, for example, in the sense of convergence to the same collection of *attractors*? This could be a rich field to plow.

### An Example Topology

Before leaving this section, . . . To do this, we exploit the notion of hyponymy and meaning postulates.

Let  $P$  be the set of predicates, and let  $B$  be the set of all elements of  $2^P$  that express hyponymy. We assert  $B$  is a basis, if not of  $P$ , then at least of everything worth talking about:  $P_0 = \chi(b: b \in B)$ . If  $b_\alpha, b_\gamma \in B$ , neither containing the other, have a non-empty intersection that is not already an explicit

. . . we demonstrate an actual topology on the set  $P$  of predicates

hyponym, we extend the basis  $B$  with the meaning postulate  $b_\alpha \sqsupseteq b_\gamma$ . For example, "dog" is contained in both "carnivore" and "mammal." So, even though our core lexicon may not include an entry equivalent to "carnivorous mammal," it is a worthy meaning postulate, and we can extend the lexicon to include the intersection. Thus,  $B$  is a basis for  $P_0$ .

Because hyponymy is based on nested subsets, there is a hint of partial ordering on  $P$ . A partial order would be a big step towards establishing a metric.

As we said before, it is not our intent to propose in this paper (Part I) a complete solution to the problem of extracting meaning from content. What we have presented is an outline of how one might approach the problem (of modeling an aspect of meaning). In Part II we propose a detailed model for the metric space  $(\mathbf{S}, d)$ , a critical first step toward realizing  $(\mathbf{S}, h)$ .

Here, reader, we pause in our journey through that shadow-land that is Content and Meaning. Were you to leave us now we would not blame you. It is no easy road.