NASA Technical Memorandum 110358

# A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain

Michael W. McGreevy



July 1995


National Aeronautics and
Space Administration

NASA Technical Memorandum 110358

# A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain

Michael W. McGreevy, Ames Research Center, Moffett Field, California

July 1995


National Aeronautics and
Space Administration

**Ames Research Center**
Moffett Field, California 94035-1000

# A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain

MICHAEL W. MCGREEVY

*Ames Research Center*

## Summary

An objective and quantitative method has been developed for deriving models of complex and specialized spheres of activity (domains) from domain-generated verbal data. The method was developed for analysis of interview transcripts, incident reports, and other text documents whose original source is people who are knowledgeable about, and participate in, the domain in question. To test the method, it is applied here to a report describing a remote sensing project within the scope of the Earth Observing System (EOS). The method has the potential to improve the designs of domain-related computer systems and software by quickly providing developers with explicit and objective models of the domain in a form which is useful for design. Results of the analysis include a network model of the domain, and an object-oriented relational analysis report which describes the nodes and relationships in the network model. Other products include a database of relationships in the domain, and an interactive concordance. The analysis method utilizes a newly developed relational metric, a proximity-weighted frequency of co-occurrence. The metric is applied to relations between the most frequently occurring terms (words or multiword entities) in the domain text, and the terms found within the contexts of these terms. Contextual scope is selectable. Because of the discriminating power of the metric, data reduction from the association matrix to the network is simple. In addition to their value for design, the models produced by the method are also useful for understanding the domains themselves. They can, for example, be interpreted as models of presence in the domain.

## Introduction

The design of a computer system that is intended to support a complex and specialized sphere of activity, a domain, must embody a model of that activity in order to be effective. Designers rely on analysis of the relevant sphere of activity, a process called domain analysis, in order to obtain a model of the domain. A domain model serves as a framework for organizing a domain system, that is, computer hardware and software which gather, manipulate, and distribute information concerning the domain and its participants. Domain modeling requires the relevant aspects of the domain to be mapped to logical forms which can be applied to the design of system components, especially the design of software. Characterization of the structures and functions that are important in the domain requires the analyst to reliably learn about the important conceptual and tangible objects in the domain, their key attributes and essential actions, and other important relationships among these objects. This information can be integrated to form an explicit model of the domain for use by software designers. The model can also be used by those who wish to understand the organization and operations of the domain so that aspects of the domain itself can be improved.

Subjectivity in domain analysis can reduce the utility of the resulting domain model and can lead to misinformed system design and inadequate service to the domain. Further, the complexity of those endeavors which require computer systems does not yield to ad hoc analyses. Effective domain analysis requires objective characterization, along with supporting quantitative metrics, in order to provide useful models of domains to the designers of domain systems.

Verbal data can be a useful source of information about specialized domains of activity. Every important domain is built upon countless words in innumerable documents, and increasingly, many of these documents are accessible in digital form. The ability to automatically, objectively, and quantitatively model important patterns in selected documents in this large collection of verbal data would be valuable to domain analysts. In fact, verbal data analysis is a central theme of research in many fields, including database design, artificial intelligence, knowledge acquisition, computational linguistics, and object-oriented analysis (OOA). Work in these fields suggests ways that verbal data might be processed to better support analysis and modeling of specialized domains.

### Network Models and Verbal Data Analysis

Common to many of the fields which conduct verbal data analysis is the use of networks as a basic modeling form.

As discussed below, the entity-relationship data model, semantic networks, association-based models of expertise, on-line lexical databases, and object-oriented models all utilize networks to organize and represent entities and relations for verbal data analysis. Networks are built of nodes, directed arcs, and annotations. Nodes and arcs are assigned a variety of meanings, but typically, nodes represent entities (identifiable things or ideas), arcs represent relations (associations) among nodes, and annotations label, quantify, and otherwise describe the nodes and arcs. The meaning of the term "entity" as used in different fields varies. In this paper, it is used to denote a separable and identifiable thing, idea, action, attribute, or attribute value belonging to a domain. The term "object" is used to refer to separable and identifiable things or ideas, and it denotes objects in general, classes of objects, and specific instances of objects.

The entity-relationship model was created to provide a generalized data model that "adopts the more natural view that the real world consists of entities and relationships" (Chen, 1976, pg. 9). Thus, using this data model, verbal and other information relevant to a domain is modeled in the database as entities and relations. Accordingly, entities are grouped in "entity sets," which are classes. Attributes and their attribute values are associated with each entity. Relationships, which have their own attributes and attribute values, characterize a variety of associations among entities. While Chen differentiates his data model from several others, including one technically known as the "network" model, his model utilizes the general notion of a network, that is, nodes, directed arcs, and annotations. The essential point, for the purposes of this paper, is the fact that Chen's data model, which is considered seminal, specifies entities, classes, attributes, attribute values, and relations as the essential ingredients for modeling a "real world" domain, and these are represented using nodes, arcs, and annotations.

Semantic networks were developed as a graphical representation of "semantic memory" (Quillian, 1968), and have been adopted by the field of artificial intelligence as a graphical form of predicate calculus (Nilsson, 1980). In both applications of semantic networks, the networks consist of a collection of labeled nodes connected by labeled associative links. Most often, the network is created by hand in an attempt to represent the precise meaning of a one or several declarative sentences. The nodes can represent terms in the predicate calculus that are identified with individual words, especially nouns, or more complex structures. The associative links can represent predicates in the predicate calculus that are identified with verbs, or with categorical or organizational relations such as "is a kind of," or "is a part of," which are sometimes called functions or mappings. The nouns, verbs, and functions in a sentence are interpreted as corresponding to things, actions, and mappings in a domain of discourse. The goal of those who use semantic network notations has typically been to investigate the nature of human or artificial memory via a concrete representation. The focus has generally been on a very fine-grained, manual analysis of a small number of sentences.

While formal verbal data analysis is only one of many knowledge acquisition methods (Boose, 1989), the conceptual models produced by efforts to elicit expert knowledge are usually in the form of linguistic structures (Shaw and Woodward, 1990). Further, most knowledge acquisition methods depend upon spoken communication between domain experts and knowledge engineers, and interviews are perhaps the most common method of elicitation. Other methods include verbal protocol analysis (Ericsson and Simon, 1984), for which the expert introspects and "thinks aloud" during a task or in retrospect, and automated textual analysis. Automating the understanding of general knowledge about a specific domain from text, however, is a major theoretical and technical challenge. The bottleneck in natural language understanding is the lexicon, an area of intense research (e.g., Zernik, 1991; Miller, Beckwith, Fellbaum, Gross, and Miller, 1990; Grefenstette and Hearst, 1992; Schütze, 1993). While unrestricted text cannot yet be automatically understood in all its complexity, it is still possible to derive useful information from it using computational means.

Analysis of large volumes of natural language text is central to the field of content analysis. An enduring hallmark of content analysis is its emphasis on mapping verbal data derived from public media to socio-political thematic categories and a search for bias or social influence. Much productive effort has gone into the application and development of computing tools for processing verbal data, including Key Word In Context (KWIC) indices, mapping input to thematic categories, and statistical analysis of text (Dunphy, 1966; Krippendorff, 1980; Weber, 1990).

Osgood (1959), for example, developed one of the first methods of computing the frequency of co-occurrences of important terms and usages within verbal data, a process he called "contingency analysis." The nature of the co-occurring entities, the granularity of the contexts in which co-occurrence was identified, and the multiplying factor, are important to note. Osgood identified themes which he grouped in such categories as: freedom, business, rugged individualism, youth, and other socio-political concepts. He considered one or more co-occurrences of two themes within an approximately half hour speech to indicate a single "hit," and the number of speeches (out of 38) containing a hit equaled the co-occurrence metric value

between those two ideas within the domain sampled by the speeches. He used the results to produce a network model of the ideas within the domain, which, at least on one occasion, he instantiated as a physical ball and stick model. The length of a stick was proportional to the frequency of co-occurrence, although some compromises were required because the dimensionality of the model was greater than three. At the time this work was done, the analysis was especially arduous, which partly accounts for the coarse granularity of the contexts (entire speeches), and the lack of precision of the multiplying factor (the number of speeches containing one or more co-occurrences). Osgood's essential contribution, however, was in his careful attention to the importance of co-occurrence relations among entities in verbal data. Although some researchers in computational linguistics also use co-occurrence information (e.g., Smadja, 1991), the main emphasis is on lexical co-occurrences such as "home run," not the conceptual relations of interest to Osgood.

The entities and relations of verbal data are of paramount importance to the disciplines of psycholinguistics and computational linguistics. For example, WordNet, a computerized dictionary based on psycholinguistic principles (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990; Beckwith, Fellbaum, Gross, and Miller, 1991), organizes tens of thousands of nouns, verbs, and adjectives according to a well-defined set of linguistic relations, the most important of which is similarity of meaning. Sets of synonymous nouns are organized into topical hierarchies based on hypernymy/hyponymy relations, which are also known as superordinate/ subordinate, generalization/ specialization, or "kind of" relations. Nouns are related to those in other synonym sets by holonymy/meronymy (whole/part) relations, and by antonymy. Verbs are grouped by major semantic category, such as: motion, possession, and communication, and they are interrelated by "entailment" (strict implication) relations. It is important to note the psycholinguistic validity, as indicated by the research underlying the design of WordNet, of using "kind of" and "whole/part" relations as the most important definitional relations among nouns in synonym sets, since these are the same relations emphasized in organizing classes/objects in object-oriented analysis and design. It is also important to observe that "general knowledge" (Miller, 1990) or "real world" relations among entities (Chen, 1976), such as environmental adjacency of objects which share a physical context or the logical adjacency of objects which share a conceptual context, are (intentionally) not captured by WordNet's definitional relations.

Network methods have been developed for analyzing "real world" relatedness among words which are prominent in domain vocabularies. In particular, "Pathfinder" networks (Schvaneveldt, Durso, and Dearholt, 1989) have been interpreted as models of expertise for application to skill level assessment (Cooke and Schvaneveldt, 1988), selection and training (Schvaneveldt, Durso, Goldsmith, Breen, and Cooke, 1985), user interface design (McDonald and Schvaneveldt, 1988; Roske-Hofstrand and Paap, 1986), and characterization of user interface designer expertise (Gillan and Breen, 1990). The data from which Pathfinder networks are created are typically derived from paired-comparison experiments, in which two words at a time are rated in terms of their relatedness. This provides a single relational weight per node pair for each subject. Relatedness data have also been derived using sorting methods in which words are assigned to groups based on relatedness, providing a relational weight of one among items in a pile for each subject performing the sort. Scores are summed across subjects to provide a relational metric based on agreement among subjects. Another data collection method is to note the sequence of command words or button presses in a user interface and then to apply a score of one to sequential adjacencies. As in the sorting task, summation of the scores across subjects provides an agreement metric. Unlike the sorting task, each subject can produce a relational value of greater than one, as when there are repeated transitions from one item to another. In the more commonly used paired comparison or sorting methods of measuring the relatedness between items, the context of the judgment is typically a scenario described at the beginning of the experiment. Further, the criteria of relatedness are usually unspecified. The resulting association matrix containing the relational weights is commonly reduced to a spatial distribution of related items via multidimensional scaling (Kruskal and Wish, 1978), or to a network of explicit pairwise relations via the Pathfinder network reduction algorithm (Schvaneveldt, Durso, and Dearholt, 1989).

In seeking to eliminate nonessential domain relations from their networks, users of the Pathfinder algorithm endorse the assertion that a practical network model of a domain must not include all possible relations among all of the entities in the domain. If it did, the model would be too complex for interpretation. Simon (1969) offered the "empty world hypothesis" as an explanation of the fact that simple models can provide useful representations of complex and important domains. His hypothesis implies that, due to the redundancy in most complex structures, there are far fewer than $N \times N$ relations of importance among N domain entities. "[F]or a tolerable description of reality only a tiny fraction of all possible interactions needs to be taken into account" (pg. 221). This suggests that in order to describe the salient entities and relations of a domain, one should first identify the domain entities of importance and then reduce the total number of possible

relations among them to those few relations which are of particular importance to that domain.

(It is helpful to be explicit about the number of possible relations among N entities. There can be $N \times N$ relations among N entities if the relationship R(A,B), with A and B being among the N entities, is distinct from R(B,A), if R(A,B) accounts for all relations from A to B and R(B,A) accounts for all relations from B to A, and if the reflexive relations R(A,A), R(B,B), etc. are included. If one is concerned with directed connections apart from reflexive ones, there are $(N \times N) - N$ or $N (N - 1)$ relations of interest. If one is only concerned with a directionless connection between A and B, and in addition, one is not interested in reflexive relations, then there are $N (N - 1)/2$ pairwise relations of interest among the N entities. More formally stated, the number of combinations of N items taken two at a time without replacement is $N!/(N - 2)!2!$, which is equal to $N(N - 1)(N - 2)!/(N - 2)!2!$, which equals $N (N - 1)/2$. If the reflexive relations are indeed of interest, this becomes $N(N + 1)/2$. Since in the general case it is not assumed that interactions between entities are directionless, and the relation of an entity with itself is not excluded, the maximum possible number of relations among N entities is considered, for the sake of discussion, to be $N \times N$, that is, N squared.)

**Domain-Based, Object-Oriented Software**

The object-oriented paradigm (Booch, 1991; Coad and Yourdan, 1991; Dillon and Tan, 1993) is particularly appropriate for mapping real-world domain models to software implementations (Fichman and Kemerer, 1992; Monarchi and Puhr, 1992; Laurini and Thompson, 1992). The object orientation, with its emphasis on objects derived from the vocabulary of the domain, is complementary to the procedural orientation, which emphasizes the order of events. The object-oriented domain model is especially useful for representing and interpreting the enduring structures of domains, integrating the logical and physical entities of importance into a coherent relational framework (Booch, 1991; Coad and Yourdan, 1991; Dillon and Tan, 1993; Graham, 1994). Further, the object-oriented approach is specifically intended to improve the isomorphy of the mapping from a domain to its software representation. Of particular importance in an object-oriented analysis is the identification of classes/objects, attributes of objects, attribute values, the actions associated with objects, and the relations among objects. The class relations among objects are represented in a superordinate/subordinate ("kind of") hierarchy, and structural relations among objects are represented in whole/part hierarchies. The relations between objects and their attributes, attribute values, and actions are implicit.

Chen (1992) asserts that before object-oriented designs can proceed effectively, users' mental models of their domains must be made available. It is difficult, Chen argues, to obtain mental models that are appropriate for object-oriented design because there are currently no objective and quantified methods for obtaining these kinds of models. Accordingly, new knowledge acquisition methods might be needed to obtain these specialized domain models. Kaindl (1994) compares object-oriented analysis with knowledge acquisition itself, and finds that they have much in common. In particular, they both require a process of discovery so that the domain of interest can be modeled, and the system requirements specified. Kaindl also asserts that networks of relations among objects are similar to the conceptual structures used by knowledge engineers. He suggests that textual documents which specify system requirements should be implemented in hypertext so they can explicitly represent that conceptual structure. Thus, he endorses the notion that networks of related domain entities are implicit in specification documents and that they can be made explicit using hypertext or object-oriented networks.

The idea that a text document can be usefully transformed in order to specify the design of software was first put to the test by Abbott (1983). He transformed informal, written procedural descriptions into computer programs. As examples, Abbott wrote a description of a function to compute the number of days between two dates, and a high level description of a function to produce a KWIC index on titles. While he considered his task to be transformation rather than modeling, and emphasized a procedural rather than an object-oriented view, it is interesting that Abbott favored a linguistic approach and suggested an object oriented view of the problem to be solved. Abbott argued that there is an important correspondence between nouns and objects, and he used parts of speech as a starting point for the specification of variables, values, subroutines, and the like. This suggested the important potential of linguistic analysis for object-oriented modeling. It is, however, a serious weakness of Abbott's approach that he himself wrote the text which served as the basis of his designs. While this might be appropriate for the sort of examples he addressed, it does not seem to scale up for application to complex domains. For modeling of complex domains, the source text should not be generated by the software designer but by persons who are knowledgeable about, and participate in, the domain of interest. Apart from this concern for domain modeling, which is more demanding than the task Abbott placed on himself, the concept of deriving software specifications from verbal data was seminal, though not yet fully appreciated.

Abbott's language-based approach has had some influence on current thinking about object-oriented design. Dillon and Tan (1993), for example, suggest that an object-oriented analysis should begin with an informal verbal description that is written by the analyst. In their books on object-oriented methods, Booch (1991), Coad and Yourdan (1991), and Graham (1994) cite Abbott's work as a method worthy of consideration, while cautioning against over-dependence on it. The analysis of real-world user domains (as opposed to small algorithmic examples), they argue, requires a far more broad view than the mere underlining of nouns and verbs to gather possible domain objects and actions. Booch finds Abbott's method to be useful due to its simplicity and the fact that it forces the developer to work in the vocabulary of the problem space. He claims, however, that the method "definitely does not scale well to anything beyond fairly trivial problems" (pg. 143). Graham elaborates briefly on how the method might be improved by more attention to the various kinds of nouns and verbs. Coad and Yourdan also find merit in the Abbott's language-based approach, but suggest looking for key nouns and verbs in the widest possible variety of domain-related documentation, not just developer-generated domain descriptions.

### Symbolic Model of Presence in a Domain

Domain experts are immersed in the sensory and symbolic experience of their domains. This immersion is, in a very real sense, a combination of sensory and symbolic presence, and one can interpret a domain model as a model of presence. Not only do such models have the potential to be generally useful to the designs of all kinds of domain systems, they could also be particularly useful for the design of virtual environment systems (systems which surround users in computer generated places, for applications including visualization of scientific data, collaborative work among geographically separated researchers, and remote control of exploring vehicles).

Some organizing structure, formal paradigm, or model must be adopted to map the experience of presence in a domain to the unavoidably explicit formal model embodied in computer software for the creation of virtual environment systems. Typically, the sensory aspects of presence receive most of the attention. Certainly, it is very important to create sensory simulations as one major component of virtual presence. To design a virtual environment for a real world domain, however, a sensory model of presence is insufficient. A symbolic model of presence in the domain is also needed. A symbolic model addresses the cognitive dimensions of presence, representing the associative experience of presence, not merely its sensory aspects. While investigation of sensory presence involves visual, auditory, tactile, and other stimuli, investigation of symbolic presence involves entities and relations. Similarly, while virtual sensory presence demands generation of surrogate stimuli, virtual symbolic presence requires generation of surrogate logical and physical entities and relations. Thus, the meaningful things in an environment, whether concrete or abstract, their attributes and actions, and their interrelations are of particular interest in a symbolic model of presence.

There is some evidence that the symbolic component of presence is characterized by persistent engagement of the person present with metonymically related entities (i.e., those related by logical and physical adjacencies and associations) encountered in environments (McGreevy, 1993, 1994). This suggests that the necessary symbolic model could be built upon the logical and physical entities and relationships which are prominent in a given domain. It would therefore be useful, as a step in developing a model of presence, to objectively and quantitatively analyze and describe those domain entities and their interrelations.

## Method

A semi-automated method of verbal data analysis was developed which can be used to derive object-oriented domain models from interview transcripts, incident reports, technical reports, informal domain descriptions, and other domain documents. It is illustrated in figure 1. The method produces object-oriented networks whose nodes and relations are weighted according to their prominence in the domain, as represented by the analyzed text. In addition, descriptions of each node and relation are provided in an outline form ordered by the weights. The method was developed to address the need for a capability to quickly and accurately produce an objective and quantified object-oriented domain model in a form that is useful for domain system developers.

In a previous domain analysis (McGreevy, 1994), the key domain entities were derived quantitatively and objectively from the transcript of a field interview, but the relations were inferred by the analyst from a close reading of the text in the contexts of the important entities. The new method improves upon this approach by automating the relational analysis through a quantitative analysis of the contexts of important entities. Further, results are presented in a more interpretable graphical form, and entities and relations are described according to their prominence in the domain.

The first step of the method is to obtain appropriate domain text. The main criterion is that the text should be

generated by domain experts themselves. Ideally, the material will have been generated by the domain experts for their own purposes. If the analysis is in preparation for participant observation among domain experts, reports concerning the domain that are written by the experts of interest should be analyzed. If the domain experts do not write technical reports, as with pilots, it may be possible to obtain transcripts of on-the-job communications or incident reports. Once some initial insight into the structure of the domain has been obtained from analysis of such material, it may be possible to interview domain experts, whose answers would provide additional text for further analysis in order to test hypotheses about the model or the domain, or to refine the domain model.

Once the text is obtained, it must be made available in digital form and can be coded. If it is not already on-line it must be digitized to ASCII form, that is, plain digital text. The most direct way to do this is to use a scanner and optical character recognition (OCR) software. After hand-correcting the inevitable errors of the OCR process, the text can be coded, a process which is semi-automated. Coding reduces diverse forms of words to their root or basic forms, distinguishes between words with identical spellings that represent either nouns or verbs or which have multiple meanings, and links multiword terms of particular salience (such as "home run" or "New York").

Once the coding is done, the text is processed to produce a list of the unique words in the text, each with its frequency of occurrence, sorted in descending order of frequency. From this list, the most frequently occurring nouns, verbs, adjectives, adverbs, and (optionally) first person pronouns are identified for use as "probe terms." These words are initially considered to be the most important ones in the domain text, and the words that occur in their contexts are considered to be importantly related to them. The most frequently occurring articles, conjunctions, prepositions, nonaction and auxiliary verbs, and other thematically uninteresting words are marked as weightless (uncounted) place holders.

The key step of the method is then performed, in which the proximity-weighted co-occurrence metric values are computed for the most frequently occurring terms, called "probe terms" (PT), relative to the terms occurring within a small context or window surrounding each probe term, called "terms-in-context" (TIC). An appropriate context window size is the average sentence length. An example of the calculation of the relational metric is shown in figure 2(a) for a window size of six. The example calculation of the proximity-weighted co-occurrence relational metric values is for 12 terms in the context of one occurrence of one active probe term (PT). The active probe term is the one whose contexts are being processed. Other, nonactive probe terms may be in these contexts but they are considered as terms-in-context. The sequence w1 through w12 (including PT) represents a sequence of terms found in the analyzed text. PT is a probe term and w1-12 represent terms found in the context of PT, that is, they are terms-in-context (TIC). In this example, the context window size is six, that is, it contains six words at a time as it effectively slides along the entire text, moving forward one word at a time. The rows a-h represent eight positions of the context window as it scans the text.

In window position "a" the probe term (PT) is not within the window, so the relational metric value for each word in the window is incremented by zero. In window position "b" the probe term is within the context window, so for each word within the window, w2 to PT, the relational metric value is incremented by one. In window position "c" the metric values of words w3 to w7 are each incremented by one. By position "h," the probe term is no longer within the window, so no values are incremented for that position. For the single occurrence of the probe term shown here, the proximity-weighted co-occurrence metric values of each word, relative to PT, are shown in the bottom row of figure 2(a). Thus, for example, the relational metric value of w5 is 4, indicating that, for this one occurrence, the relation between PT and w5 has a metric value of four.

Note that the context window scans the entire text, and whenever it contains the active probe term, the relational values of terms in its context are incremented in a manner similar to that shown here. Thus, for example, if there were two occurrences of a given probe term in the entire text being analyzed, and if term w5 were in the same position relative to PT both times, w5's total relatedness to PT in the text would be 4 + 4 = 8. This value can be indicated as: co-occ(PT,w5)=8, or R(PT,w5)=8. The method does not use reflexive relations, so co-occ(PT,PT) is set to zero.

Use of the sliding window method causes an asymmetry in the relations between some terms. Specifically, the relational metric value of one probe term, PT1, in the context of another probe term, PT2, is not necessarily equal to the relational metric value of PT2 in the context of PT1. This is because the context of one instance of an active probe term stops just short of another instance of the same probe term if they both occur within the same context window. Figures 2(b) and 2(c) illustrate this asymmetry. In the example, co-occ(age,flow) is equal to six while co-occ(flow,age) is equal to four. Note that the smaller of the two values is a result of the fact that potentially overlapping contexts are, in effect, prevented from fully doing so. In the example of figure 2(c), this seems reasonable, since "age" is already in the context of the first occurrence

of "flow" and "color" is already in the context of the second occurrence of "flow." Since cases with this extreme degree of proximity between two occurrences of the active probe term are rare, and since many contexts contribute to a final relational metric value, the degree of asymmetry is relatively small in most cases. The asymmetry issue is addressed in more detail in the discussion section.

The relational metric values which are produced can be displayed as an association matrix with the probe terms defining the rows, and the terms-in-context defining the columns. In practice, the probe terms themselves are also found among the terms-in-context. The cells of the association matrix represent the relations between the row and column terms, and each cell contains a relational metric value. A similar matrix can be produced by paired comparisons, sorting tasks, and sequential actions (McDonald and Schvaneveldt, 1988).

The overall process is one of identifying the most important or prominent terms in the verbal data and the degree of relatedness among the terms. This can be visualized as starting with a W × W matrix whose rows and columns are identical and which both contain every unique term in the verbal data. This can also be represented as a network of W × W relations (which includes reflexive relations) among W nodes. The method first identifies the N most important of the W nodes, where importance is assumed to be highly correlated with frequency of occurrence, and N is much less than W. Next, the method determines the strength of each of the relations among those N probe terms and M terms-in-context. In practice, M is much greater than N and some or all of the N terms appear among the M terms-in-context. If a square matrix is needed for application of data reduction algorithms, the matrix can be padded with zeroes to produce an M × M matrix, or a smaller square submatrix.

Several different approaches are available to reduce the complexity of the data in an association matrix. It is possible to apply multi-dimensional scaling (MDS) or cluster analysis methods to the association matrix to find overall dimensions of relatedness by computing a spatial arrangement of the terms and looking for patterns such as groups or other distributions (Kruskal and Wish, 1978). A different approach is to apply the Pathfinder algorithm (Schvaneveldt, Durso, and Dearholt, 1989) to the matrix, which preserves an interconnected network while reducing it to a simpler one with certain selectable features, such as the property that it approximates a minimal spanning tree. MDS, cluster analysis, and Pathfinder network reduction are especially useful for reducing association matrices having cell values that are rather uniformly distributed, as is common with paired-comparison data, but the relational metric values yield to more straightforward data reduction. The cell values (relational weights) produced by the proximity-weighted co-occurrence method have a large dynamic range, a small number of high values, and a rapid fall-off from the highest values. This enables the effective application of a threshold to select the most salient relations by selecting those few having the highest relational weights. This greatly reduces the complexity of the final network while retaining the top relations. This network provides the framework for description of the domain. As will be shown, Pathfinder networks do not preserve the relations with the highest relational metric values, but they provide useful supplementary representations of the domain.

The next step is to produce a description of every node and relationship in the final network. The nodes and relationships in the final network are listed in order of their importance to serve as an outline for a document in which each one is described, the object-oriented relational analysis report (see Appendix). All of the nodes are listed and defined in order of the total weight of relations in which they participate. This ranks the nodes on degree of relatedness, which can be considered as a measure of the importance of a node. Under the heading provided by each node, the importantly related nodes are listed, and each relationship is described.

To generate the descriptions, the important nodes and relationships are described in a weight-prioritized, pairwise relational analysis. It is during this stage of the process that the domain analyst must learn about the domain. To do so, a concordance/KWIC index (Thomson, 1992) is used to extract one probe term (PT) at a time from the original text, along with its contexts. These contexts are reviewed in terms of the relationship of that node with one term-in-context (TIC) at a time. The benefit of the relational metric method is that it provides the analyst with a focused, prioritized, and efficient outline of the most important entities and the most important relationships in the domain.

The products of the method include: an electronic database containing the most important relationships among the most important entities in the domain; a network model of the domain and optional supplementary networks; a Key Word In Context (KWIC) index in electronic form; and an object-oriented relational analysis report containing descriptions of each node and each relationship in the network model of the domain.

The domain model produced by this method illustrates Simon's empty world hypothesis. That is, the method preserves only the small number of truly prominent entities and relationships in order to produce a simple but potentially useful description of reality.

Software for processing the text includes a mix of commercial off-the-shelf software running on a personal computer, Unix utilities running on a workstation, specially written Unix shell scripts, and freeware from Internet. The network figures were generated from tabular data by a commercial software tool called KNOT (Knowledge Network Organizing Tool), from Interlink, Inc., Las Cruces, New Mexico.

## Domain to be Modeled

A current domain of interest to NASA is the Earth Observing System (EOS), an ambitious attempt to create a globally comprehensive capability to monitor and study the Earth's environment as an integrated whole. EOS is a complex domain which must be analyzed in order to provide designers with explicit and objective domain models so that they can effectively design EOS information systems. A recent study of the state of scientific visualization relative to EOS requirements (Botts, 1993a and 1993b) indicates that there remains a significant unmet need to effectively map user requirements to system designs.

The key bottleneck in visualization tools for EOS, according to the scientists responding to Botts' survey, is lack of adequate software. In particular, many respondents characterized their current visualization tools as inflexible, not extensible, difficult to learn and use, failing to provide integrated capability for both visualization and analysis, too costly, and not doing all that the scientists need to do. While there is not a single solution, it is clear that a commonality of these shortcomings is a failure of developers to fully address the needs of users. It is typical of developers to give very limited attention to the real needs of users (see discussion in McGreevy, 1994), and to instead concentrate on the challenges of implementation. Even some who propose to develop domain-oriented software (e.g., Tracz, Coglianese, and Young, 1993) do not make an adequate effort to discover user needs, but instead expect the user to concisely package their own requirements in a form that is immediately useful to the developers. As an alternative to this unrealistic approach, it would be helpful if reliable and valid domain models could be made available to developers so that they would have a correct understanding of the needs of users such as EOS scientists without either the developers or the users being required to perform the difficult and time-consuming chore of developing this model for themselves.

In order to approach the human-computer interaction requirements for visualization and analysis systems within such a complex domain as EOS, it is necessary to select a very specific target. A first cut is to limit the scope to that within one of the many EOS Integrated Studies (EOS/IS) groups (Asrar and Dokken, eds., 1993). The volcanology group ("A global assessment of active volcanism, volcanic hazards, and volcanic inputs to the atmosphere from EOS") was selected, largely due to the author's desire to benefit from domain knowledge gained during earlier work in field geology in volcanic terrain environments (McGreevy, 1993; McGreevy, 1994). The scope of the interests of the EOS/IS volcanology group is world wide, involving a dozen or so lead investigators, so a further scaling down is required, at least for the development stage of the method.

The method of domain modeling described in this paper was originally developed to analyze interview transcripts, so one approach might be to interview one or more members of the EOS/IS group, and to apply the method to their answers. Since access to experts is a limited privilege (Jorgensen, 1989), the interviewer must do considerable preparation in advance. One way to do such preparation is to analyze more readily available verbal data in advance of the field work. Thus, the method can be applied at several stages of a domain study. In this paper, since the method is new, it was decided to first apply it to readily available textual materials.

To select appropriate verbal data, it was noted that several of the EOS/IS investigators had contributed to the development of a CD-ROM set containing Earth sciences data thought to be representative of future EOS data for a single volcano (NASA, 1992). On that CD-ROM, references to scientific papers were listed, some of which were authored or co-authored by members of the EOS volcanology group. One of these papers, "Combined use of visible, reflected infrared, and thermal infrared images for mapping Hawaiian lava flows" (Abrams, Abbott, and Kahle, 1991), was selected for the initial textual analysis because the paper described use of multispectral data to study volcanic terrain. The abstract of the paper is shown in figure 3.

Once coded, the paper contains 3480 total words arranged in 156 sentences, with an average sentence length of 22 words. There are 831 unique words in the text, of which 42 (including "a," "an," "and," etc.) are considered to be weightless spacers, leaving 789 unique words of interest.

Ideally, the method of analysis applied to this domain document, which produces an objective and quantified domain model of rather limited scope, can be developed and scaled up to address increasingly larger contexts, that is, multiple authors, studies, interviews, and other textual resources. This would allow it to address the needs of scientists in an entire Interdisciplinary Science group, such as the volcanology group, and perhaps volcanologists in general. Should the method be found to be useful, it could

be effectively applied to other EOS/IS groups, as well as to other domains of interest.

## Results

The results include: 1) an electronic database containing the most prominent relationships among the most prominent entities in the domain; 2) a network model of the domain and several supplementary networks; 3) a Key Word In Context (KWIC) index in electronic form; and 4) an object-oriented relational analysis report containing descriptions of each node and each relationship in the network model of the domain.

The core results consist of weighted relations between pairs of weighted nodes in a tabular database which shall be called the R-list. As an illustration of the R-list, the top 40 relations are listed in table 1. There are 9075 records in the R-list, representing 9075 relations among 789 unique terms. These were obtained by the use of 50 probe terms (listed in table 2) and a context window size of 22, applied to the coded version of the Abrams text. (Issues concerning the number of probe terms and the context window size are addressed in the discussion section.) Each record in the R-list represents a proximity-weighted co-occurrence relation between a pair of terms (nodes) in the text. There are generally many records containing a particular probe term (PT) or term-in-context (TIC) but there is only one record containing any given ordered pair (PT, TIC).

Figure 4 is a graph of the 9075 relational metric values in the R-list, sorted in descending order. Relations with zero weight are not included in the R-list. Clearly, only a small percentage of the total possible relations have large weights. This seems to be in accordance with Simon's empty world hypothesis, indicating that of all the many possible relations, only a very few are important for "a tolerable description of reality" (Simon, 1969, pg. 221). The method described in this paper depends upon Simon's hypothesis being true for the domains to be analyzed, and attempts to reduce the many possible relations to the few which really matter. The graph in figure 4 shows that the method identifies the desired small number of relations. In later parts of this paper, evidence will be presented to support the argument that these few relations do indeed capture the essence of the domain. Of the 622,521 possible relations among the 789 unique words in the analyzed text, the 789 reflexive relations (i.e., those between each word and itself) are not used, and, in this study, the method eliminated (zeroed) all but 9075 of the remaining 621,732 relations. Of these 9075 relations, it can be observed in figure 4 that few have relational metric values near the observed maximum of 314, and that most are nearer to the minimum of zero. For example, 1222 relationships have relational metric values greater than or equal to 25, 386 relations have values ≥50, 164 relations have values ≥75, and 84 are ≥100. The number of relations required for a domain model with an appropriate level of detail is yet to be determined.

The first alternative form of the results is the association matrix. The weights of the 621,732 nonreflexive relations among 789 unique terms can be represented as values in the nondiagonal cells of a 789 × 789 association matrix. The 789 cells in the diagonal, representing reflexive relations between each word and itself, are not used. Only 9075 of the remaining 621,732 cells contain nonzero relational metric values, and these 9075 relations can be represented as an association matrix having 50 rows and 789 columns. This matrix shall be called R-matrix. The choice of 50 probe terms accounts for the 50 rows. There are 789 columns because that is the number of unique terms-in-context found in the vicinities of the probe terms. As it turns out, the 50 probe terms picked up every one of the 789 unique terms in the coded text as a term-in-context. Of the 789 terms-in-context, 50 are probe terms found in the contexts of other probe terms. Thus, within R-matrix there is a 50 × 50 matrix of relations among the probe terms, a 50 × 739 matrix of relations between the probe terms and all of the other terms, and a variety of other submatrices, as discussed below. Table 3, for example, is a 21 × 21 submatrix containing the top 40 relations, which corresponds to the 40 item sublist in table 1.

Of all the possible submatrices extracted from R-matrix, only the 50 × 50 matrix of probe terms, or submatrices of it, have two different relational metric values between every pair of terms. That is, they have one value for probe term Y in the context of probe term X, and a different value for probe term X in the context of probe term Y. All of the other relations in submatrices of the R-matrix are unidirectional, that is, there is a single relational metric value for term-in-context Y in the context of probe term X.

The second alternative form of the results is the network. The analyzed text can be represented by a network containing 621,732 nonreflexive relations among 789 nodes. The R-list and R-matrix implicitly describe a subnetwork containing 9075 arcs and 789 nodes. This shall be called the R-network. Each arc has a nonzero relational metric value, and each node has a weight, its frequency of occurrence in the body of the text. The more heavily weighted nodes and arcs represent the more important parts of the network. It follows that simpler subnetworks can be derived from the R-network which still retain the most heavily weighted nodes and arcs, and thus could retain the important characteristics of the domain. These

subnetworks correspond to sublists of the 9075-record R-list, and submatrices of the 50 × 789 R-matrix.

### Network Models Based on the Most Prominent Relationships in the Text

Figure 5 shows an example subnetwork which includes the 40 relations with the highest relational metric values out of the 9075 relations in the R-network. This network directly corresponds to the 40 records shown in table 1 and the association matrix in table 3. Note that there are 21 nodes in this subnetwork. Within this subnetwork the top five relations (with their relational weights shown in parentheses) are old-> flow(314), aa-> flow(299), young-> flow(296), pahoehoe-> flow(287), and age-> flow(271), where the first word in each pair is a probe term, and the second is a term-in-context. (Note: aa is a blocky form of lava and pahoehoe is a ropey form of lava.) The weights of the nodes included here are: flow(81), age(32), old(24), aa(19), pahoehoe(19), and young(13). From this very limited information, one can infer that "flow" is the most important node, and that the nodes "young," "old," "aa," and "pahoehoe" are importantly related to "flow" in this domain. A reading of the text confirms that "flow" is the most central concept, that determining flow age and distinguishing young from old flows is the main theme, and that differentiating the two kinds of flow textures, aa and pahoehoe, and determining their ages, is another of the main ideas expressed. Thus, from even such a tiny subnetwork as one consisting of five relations and six nodes it is possible to tolerably well describe the reality of the domain sampled by the text.

Similarly, the remainder of the subnet in figure 5 captures other key notions of the text. For example, a module centered on "data" is already beginning to emerge, showing the close relation between TIMS [Thermal Infrared Multispectral Scanner] and NS-001 [a multispectral scanner], which are the two sources of data, and the "data" node. Further, to use data, as indicated by the relation between "use_verb" and "data," is a central action that is repeatedly expressed in the text. The fact that the nodes "tims," "ns_oo1" and "use_verb" all converge on "data" indicates the centrality of "data" within this module, as well as the subordination of the other three nodes. In addition, the node "image" is closely associated with "data" because images in this domain are created from "data." Also, the characteristic of images which is most important in this domain sample is "color." Accordingly, the close relation between the nodes "image" and "color" clearly captures this idea. The fact that both "image" and "color" are directly associated with "flow" is in harmony with the fact that the images in this domain represent flows, and colors in these images differentiate one flow from another. The remaining nodes and links in figure 5 are also consistent with the main ideas expressed in the text.

By considering different numbers of the most heavily weighted relations, that is, those with the highest relational metric values, various other sublists can be derived from the R-list, producing submatrices of the R-matrix, and subnetworks of the R-network. For example, using a threshold value, T, applied to the graph in figure 4, the top R relations can be selected. For example, for T = 75, R = 164, as shown in figure 4. When the top R relations are selected from the R-list, the records obtained include not only the weight of each relation, but also the nodes involved and their weights (see table 1). This information can be used as an ordered list, or they can be used as an association matrix or a network diagram, to guide the next steps of the object-oriented domain analysis.

To obtain a large enough network for a meaningful test and demonstration of the method, a threshold value of T = 75 was used. The resulting network is the one which will be fully described in order to develop a domain model. This network, shown in figure 6, is based on all records in R-list having a relational metric value greater than or equal to a threshold value of 75. The network contains the top 164 relations in R-list. Participating in these relations are 53 nodes. The weights of nodes and relations are not shown in this figure in order to avoid visual clutter, but they can be obtained from tables 2 and 5. In addition, these weights are discussed in more detail below in the context of creating the object-oriented relational analysis report. Note that table 1 contains the top 40 of the 164 relations and 21 of the 53 nodes in figure 6. Further, the network in figure 5 is a subnet of the one in figure 6, and it, too, contains 40 of the 164 relations and 21 of the 53 nodes in figure 6.

Even without considering the weights of relations and nodes in figure 6, it is evident that "flow" is the central node of the domain, judging by the number of relations in which "flow" participates. It is also evident that the next most important nodes are "image" and "data." Without counting the number of relations, their weights, or the node weights, the nodes "age," "color," "old," "reflectance," "component," and "green" all seem to be important at a level just below that of "data." Specific attention is given to the weights and numbers of relations in the section below on generating detailed descriptions of the nodes and their relationships. For now it is sufficient to note that the subnetwork in figure 6 adds nodes and relations not contained in the smaller subnetwork in figure 5, and that those details correspond well to those obtained in reading the text. For example, the action "combine" is associated with "data" since the domain text describes the combination of NS-001 data and TIMS data.

The description of the nodes and relations in the network model of the domain (fig. 6) is provided in the object-oriented relational analysis report shown in the Appendix. Given the network in figure 6, the first step in generating this report is to calculate new node weights based on relatedness, as shown in table 4. Next, a list of the 164 most highly related node pairs is exported as plain text from the R-list database, and the metric values are normalized (divided by the observed maximum), as shown in table 5. Using this list as a guide, each node is defined, and the relationship between each node pair is described. It is helpful to consult domain glossaries for definitions of terms. To obtain descriptions of the relationships, the analyst reviews the original text. Figure 7 shows a screen image of the concordance/KWIC index as it appears while being used to search the original text for the node term, "flow." The window at the bottom shows some of the contexts around the term "flow" (the rest are available by scrolling) while the window at the top contains the full text context for any line selected in the bottom window. (In practice, the windows are made much larger on the computer screen, so as to display more of the contexts.) By using the pattern matching capabilities in the concordance program (Thomson, 1992), the contexts shown for "flow" or any other word can be limited to just those containing the second word in a node pair. For a small body of text, it is just as easy to print out all of the contexts of "flow" (or any other node) and to circle the occurrences of the second item in the node pair. A description of the relationship is then obtained by reading the contexts of the co-occurrences. The concordance/ KWIC index aids the analyst during the process of describing the relationship between each pair of nodes in figure 6, which are explicitly listed in table 5. The descriptions are shown in the Appendix.

A simplified, object-oriented network model of the domain can be derived from the network in figure 6 and the descriptions and weights in the object-oriented relational analysis report (Appendix). This network, shown in figure 8, shows only objects and inter-object relations. Table 6 shows all 164 relationships of figure 6 mapped to object relationships. That is, if one of the participants in a relationship is an attribute, attribute value, or action, rather than an object, its label is expanded to indicate the name of the object to which it belongs. For example, since "age" is an attribute of the object "flow," the label for "age" becomes "flow(age)." Similarly, since the attribute value "old" refers to the attribute (relative) "age," which belongs to the object "flow," the label of "old" becomes "flow(old)." When a relationship is between two nodes which refer to the same object, the relationship represents internal structure of the object, that is, an intra-object relationship. For example, the relationship between "flow(age)" and "flow(old)" is one which is internal to the object "flow." Otherwise, the relationship represents an external, inter-object relationship. For example, the relation-ship between "flow" and "image" and that between "flow(age)" and "image(color)" are both external, inter-object relationships. All relations between an object and its internals (attributes, attribute values, and actions) are summed, and the sum is used as a measure of the object's internal complexity. In addition, once all entities are identified as objects or assigned to objects, all relations between any two objects are summed and treated as a measure of overall inter-object relatedness. For example, the one relationship between "flow" and "band" is that between "flow(age)" and "band," whose relational metric value is 0.25, so that is taken as the value of the relation between "flow" and "band." To simplify the network even further, only one weight is shown for arcs representing either mutual or one-way relations. Thus, each relational weight in figure 8 is the sum of the individual relational weights in either direction.

This object-oriented network model of the domain (fig. 8) is a companion to the detailed information in the Appendix. It shows that "flow" is the most complex (i.e., elaborated) object in the domain, with the object "image" only elaborated about 37 percent as much. The object "data" is only 14 percent as complex as "flow." The relation between "flow" and "image" is by far the dominant one, with the next most important relation, between "image" and "component," being only 29 percent as important. In the context of "flow," the most importantly related objects, after "image," are "data," "aa," "pahoehoe," and "group." In the context of "image," the most importantly related objects, after "flow," are "component," "data," and "group." Another feature is that "image" is closely associated with a module consisting of "component," "band," and "reflectance." Further, "data" seems to form a module with "tims" and "ns_oo1." This network can serve as a summary framework of the detailed descriptions in the Appendix. Figure 6 shows both the intra-object and the inter-object relations.

### Comparison of Result Network with Pathfinder Networks

The network domain model shown in figure 6, which serves as the basis for the object-oriented relational analysis report (Appendix), could have been constructed by alternate means. Pathfinder networks can also be derived from the relational metric values in R-list. Three were created in order to compare them with the network domain model in figure 6, which is based on the top 164 relations.

The key equations and parameters for creation of Pathfinder networks are described by Schvaneveldt, Durso, and Dearholt (1989). The essential idea is to minimize the "path length" (which may include multiple arcs) between nodes. The maximum number of links in a path is set by a parameter, q. A second parameter, r, determines how relational weights contribute to the path length. The minimal Pathfinder network has an r value of infinity and a q value of n – 1, where n is the number of nodes in the network. One problem with large values of r is that as r increases, the proximity of nodes is determined by the weaker associations between the nodes, thus reducing the influence of the most important relations. A problem with any values of q larger than 1 is that as q increases, increasingly indirect (multilink) relations take precedence over direct relations. As r or q decrease, the number of links increases. When q = r = 1, every association in the input matrix appears as a relation in the output network. Thus, one must choose between an emphasis on less important or indirect associations and too many links in the output network. The minimal Pathfinder network (q = n – 1, r = infinity) is the most readable choice, and while it drops important direct relations and relies on implicit and very indirect relations, it does provide a genuinely meaningful model of a domain. That model is not, however, ideal for object-oriented analysis, as discussed below.

Of the three Pathfinder networks created from data in R-list, figure 9(a) is the most directly comparable to figure 6. Both networks contain nodes which are classes/objects, attributes, attribute values, and actions, and they each have about the same number of relations. To create the network in figure 9(a), the parameter q was set to 99 and the parameter r was set to infinity, so this is a minimal Pathfinder network having 155 relations among 100 nodes. It is derived from the 3020 nonzero-weighted relations among the 100 most interconnected nodes in R-list. Figure 9(a) has 9 fewer relations but 47 more nodes than figure 6. In cases where figure 9(a) has the same relations as figure 6, the relational weights are identical. Both networks contain the top 24 relations, for example. Because there are so many more nodes in this Pathfinder network, it represents more detail about the contents of the domain. To bring in so many additional nodes while keeping the number of relations down, the Pathfinder algorithm deleted many of the more important relations, which were considered to be redundant. Thus, for example, the important relation between "image" and "data" is deleted because a more heavily weighted multilink path can be traced via "flow." Still, the sparse and readable network in figure 9(a) contains an additional 47 nodes beyond that in figure 6, providing additional domain information. Further, the Pathfinder network contains reasonable semantic associations, attesting to the semantic coherence of the data and the utility of the method.

The strategy of deleting important links because they are "redundant" with multilink paths is the quality of Pathfinder networks which makes them undesirable for object-oriented analysis. In OOA, the analyst must identify collaborating classes/objects and internal class/object structure. If such important relationships as that between "data" and "image" are not made explicit in a network domain model, then the analyst will fail to appreciate those relationships. Another negative aspect of Pathfinder networks for object-oriented relational analysis is that as important relations are omitted, less important ones are retained. With Pathfinder networks, the only way to retain all of the important relations and none of the indirect ones is to settle for a dense network containing all of the many associations in the input matrix.

Pathfinder networks, while not ideal for object-oriented relational analysis, are still useful for reducing data in large association matrices to sparse and readable network representations. The many studies based on the Pathfinder method (see Introduction) attest to its usefulness in modeling domains. Thus, while networks based on a small percentage of the top relations, such as figure 6, are more directly applicable as the basis for object-oriented modeling of the most prominent relational structure of a domain, the Pathfinder networks provide useful supplementary information.

The Pathfinder network in figure 10(a) provides an additional example. It is one which is constrained to include only the 50 probe terms. The network is based on a Pathfinder analysis of the 1756 nonzero-weighted relations among the 50 probe terms contained among the 9075 relations in R-list. The algorithm used a q value of 49 and an r value of infinity, producing a network linked by a minimal set of 102 relations. This network domain model provides a useful supplementary view of the domain, as represented by its 50 most important nodes. These nodes are linked by a nearly minimal spanning set of relations. A similar supplementary view can be made, based on the top relations among probe terms, but one with the top 102 relations includes only 34 of the probe terms. To include all 50 probe terms in the top relations, a minimum of the top 495 relations would be required. Clearly, the Pathfinder network provides a much more readable network for a supplementary view containing all of the probe terms.

The Pathfinder network in figure 11(a) contains the 30 most important classes/objects in the domain and the 70 which are most closely associated with them. No attributes, attribute values, or actions are included. (The

notion of class/object was liberally interpreted for this network. Color is included, for example, because it could be considered to be a class/object by virtue of having a potentially complex internal structure and being widely reusable). The q and r values used by the Pathfinder algorithm were 99 and infinity, respectively. This reduced the 1454 nonzero-weighted relations among these 100 classes/objects to a minimal Pathfinder network containing 137 relations among the 100 nodes. This network usefully supplements the domain information contained in figures 6, 9(a), and 10(a). To do so, it uses many relations whose metric values are among the lowest, but it produces a highly readable, semantically interpretable network domain model which efficiently interconnects all 100 nodes.

The chief contribution of Pathfinder networks to object-oriented relational analysis is that they provide sparse and readable supplementary views of the semantics of the domain. The three Pathfinder networks produced from data in R-list (figs. 9(a), 10(a), and 11(a)) clearly demonstrate a flexibility of viewpoint that is unavailable when merely using the top R relations among N nodes. On the other hand, the method of using the top relations sacrifices none of the important relations to network efficiencies or alternative views of the domain. Thus, the two methods play complementary roles when applied to proximity-weighted co-occurrence data. Differences between the two approaches in terms of the numbers of nodes and relations have already been discussed above. Differences in their relational values are shown in figures 9(b), 10(b), and 11(b). Figure 9(b) compares the relational metric values of the relations used in figures 6 and 9(a). Figures 10(b) and 11(b) compare the relational metric values of relations used in figures 10(a) and 11(a) with comparable networks having the same numbers of top relations. Figures 9(b), 10(b), and 11(b) show that the Pathfinder networks contain relations having lower relational metric values (but have more nodes per relation) than comparable networks containing only the top relations. While it is true that the Pathfinder networks in figures 9(a), 10(a), and 11(a) do contain the top 24, 24, and 14 relations respectively, they omit many of the next most important relations. Thus, while Pathfinder networks are useful they must be interpreted with care.

The threshold method for selecting the top relations is effective and the Pathfinder network reduction method is not required because of the distribution of the relational metric values. These weights, produced by the proximity-weighted co-occurrence method, have a large dynamic range, a small number of high values, and a rapid fall-off from the highest values. This enables the effective application of a threshold to select the most salient relations by selecting those few having the highest relational weights. This reduces the complexity of the final network while retaining the top relations, providing a framework for description of the domain. Pathfinder network reduction is more appropriate for reducing association matrices having cell values that are more uniformly distributed, as is common with paired-comparison data. Figure 12 illustrates the difference between the relational weights in a typical association matrix based on paired-comparison judgments (from Schvaneveldt, Durso, and Dearholt, 1989) (the upper graph) and the highest 700 of 9075 relational metric values computed from the analyzed text according to the proximity-weighted co-occurrence method developed in this paper (the lower graph). The large dynamic range, small number of high values, and rapid fall-off from the highest values distinguishes the relational metric data from the paired-comparison data. The Pathfinder method is only necessary for analyzing relatedness values which are not strongly differentiated. This includes all of the paired comparison data in the upper graph of figure 12, and those relations in the lower graph which have relational metric values below a threshold of around 75, that is, for relations beyond the top 164 relations which produced figures 6 and 8, and the object-oriented relational analysis in the Appendix.

## Discussion

A key innovation introduced in this paper is an automated method of calculating a relational metric, based on proximity-weighted frequencies of co-occurrence among terms in domain text, and the use of that metric to characterize the relational structure of the analyzed domain. Another innovation is the generation of link-weighted networks based on the relational metric values derived from verbal data, whereas other researchers have derived such networks from paired comparisons, sorting tasks, or sequential activities such as typing commands or pushing buttons (see Introduction). A third innovation is the processing of verbal data in such a way as to generate an object-oriented domain model for the purpose of implementing software. This most directly builds on the ideas of Abbott (1983), but improves the method by processing domain-produced verbal data, and by objectively and quantitatively deriving weights and rank orderings for the domain terms and the relations among them. This use of domain-produced data to generate a domain model is the fourth innovation introduced in this paper. A fifth innovation is the quantitative method of network reduction based on object-oriented principles, as when the network domain model in figure 6 is reduced to class/object relations in figure 8, by applying the information in the object-oriented analysis report (Appendix) and the relational metric values in table 6. This method provides some

of the "information hiding" needed for representations of complex domains, and it could be extended to reduce the class/object network to a network of modules.

### Efficacy and Repeatability of the Method

For the method of domain analysis presented in this paper to be useful, the resulting model must represent the prominent structural characteristics of the domain, as contained in the analyzed text, in an explicit, objective, and quantified form that is stable and reproducible. The efficacy and repeatability of the method are discussed below, as appropriate, for the key components of the method: coding of the original text, determination of the prominent domain entities and selection of probe terms, determination of the relative prominence of relations, creation of domain networks, and description of key domain entities and relations, including identification of domain classes/objects and assignment of attributes, attribute values, and actions to classes/objects.

**Coding**– The degree of coding done in this study was minimal, especially as compared with traditional content analysis methods in which the primary task of the coder is to assign terms to socio-political categories. In this study, the coder differentiated nouns from verbs, and mapped them to base forms. Where there was ambiguity of meaning or part of speech, the coder assigned tags to eliminate the ambiguities. The most variation of coding in this method is likely to occur when multiword entities are identified. One solution is to be conservative in the identification of multiword entities. At the extreme, none would be identified. This, however, would force terms like "Mauna Loa" into two separate but closely related nodes. At the other extreme, any adjective could be permanently linked with its noun, which has the effect of artificially lowering the apparent prominence of the noun. A reliable and effective linking of multiword entities would only join terms such as "Mauna Loa." Overall, the coding done in this study was minimal and rather mechanical, making it very repeatable. That the coding supports the effectiveness of the method can be seen in the fact that meanings were clarified and no spurious meanings were introduced.

**Probe terms**– Selection of probe terms (PT) is objective and repeatable. The probe terms selected in this study include the nouns, verbs, adjectives, adverbs, and first person pronouns that were most frequently used in the analyzed text. Words which were not used as probe terms include: pronouns referring to things, nonaction and auxiliary verbs, conjunctions, prepositions, articles, and numbers. The purpose of using the most frequently occurring terms is to capture an objective, overall characterization of the prominent entities in the domain as represented in the text. It might also be appropriate to supplement this view by using additional sets of probe terms comprised of those frequently occurring terms which are also focused on particular themes, but this remains for a future study.

The effectiveness of initially using the most frequently occurring terms as the most important terms in the domain sample, and thus as probe terms, must be demonstrated. One argument in favor of the effectiveness of using frequency of occurrence as a measure of importance is that fact that it is the most fundamental metric used in content analysis studies, and has been for decades (Krippendorff, 1980). Similarly, ethnographers routinely infer the importance of domain concepts or "native terms" from their presence, frequency, and context in verbal data (Fetterman, 1989; Jacobson, 1991). Another argument in favor of the effectiveness of using frequency of occurrence as a measure of importance is that the results of this study show that a small number (50) of the most frequently occurring terms (the probe terms) are closely related to all of the other terms in the analyzed text. That is, every one of the 789 unique terms appeared in the context of one or more of the 50 probe terms. The ability of 50 probe terms to span the entire text argues in favor of accepting them as including or being among the most important terms in the domain text. This also indicates that 50 is a large enough number of probe terms to span the domain sample.

In contrast to using the most frequently occurring terms in domain documents as probe terms, terms for deriving Pathfinder networks from paired comparison studies have been identified by more arbitrary means. For example, Cooke and Schvaneveldt (1988) used 16 terms taken from chapter headings of an introductory computer science textbook as probe terms, and Schvaneveldt and his colleagues (1985) used 30 "basic concepts" in air combat without elaborating on their origin. McDonald and Schvaneveldt (1988) cite a sequential adjacency study whose probe terms were 49 Unix commands used by at least five of nine experienced Unix users.

The issue of how many probe terms should be used for analysis of a body of text needs further investigation. In this study, the use of 50 probe terms was suggested by the distribution of frequencies of occurrence of the unique words in the analyzed text. The distribution of candidate probe terms (that is, words other than articles, prepositions, and other such words) is shown in figure 13. The most frequently occurring words are relatively few in number and appear much more frequently than the others, so these were used as probe terms. The cut-off point was chosen to be the middle of the "knee" of the curve. At this point, the frequency of the lowest ranking probe term, "visible," is only 8, which is less than 10 percent of the frequency of the most important probe term, "flow,"

which occurs 81 times. These criteria are somewhat arbitrary, however, and the issue remains as to where to draw the line between probe terms and nonprobe terms, and the effect of that decision.

Several constraints may influence the decision, including the need to have a sufficient number of probe terms to adequately model the domain, the cost of processing probe terms, and the need to limit the number of paired comparisons to be made in a parallel experiment. If it takes 5 sec to make a relational judgment between a pair of terms (McDonald and Schvaneveldt, 1988), then 50 probe terms compared with 789 terms-in-context (which include the 50 probe terms) would require (50*49/2) + 50*739 = 38,175 judgments taking 53.02 hr. If the comparisons were limited to paired comparisons of probe terms, it would require 50*49/2 = 1225 judgments taking 1 hr and 42 min. Paired comparisons among 60 probe terms jumps to 2 hr and 28 min, and 70 probe terms would require 3 hr and 22 min. The cost of processing probe terms has already decreased significantly by improvements to the software, but the still-significant cost of processing additional probe terms must be weighed against the benefit. The key benefit is the quality of the domain model produced by a given number of probe terms. A method to obtain a quantitative measure of this might be to find the relationship between the magnitudes of the relational metric values of the top relations, and the number of probe terms required to obtain these top relations. When inclusion of additional (less frequently occurring) probe terms provides no additional important relations, this would indicate that a sufficient number of probe terms had obtained a model of the desired complexity. This method is being developed and applied in a subsequent study.

**Relational metric**– The proximity-weighted co-occurrence method of determining the relative importance of domain relations is stable, objective, quantified, and reproducible. The sliding window calculation of the relational metric is conceptually simple and it captures both the frequency of co-occurrence among terms within multiple contexts of selectable size and also their proximities within each context. The method is also reasonable. It works in a way that is analogous to the way a human reader might evaluate relatedness among important words in a text. In fact, the method was designed to automate and objectify a process applied in a previous study to derivation of a domain model from an interview transcript (McGreevy, 1994). In that study, the important words were determined by frequency counts, just as in this study. These words were then highlighted in the text and the text was reviewed to evaluate the relationships among words. Words found in the context of a highlighted term seemed to be reasonably related to it, and those frequently occurring closer to it seemed to be more closely related. The size of a meaningful context varied, but often appeared to range from the sentence before to the sentence after a word. A sliding window that is one sentence wide captures a similar context. Thus, the relational metric method of this paper can be considered to be operationally effective because it automates and objectifies a process that one can usefully apply manually.

As a quantitative basis of relational networks, application of the proximity-weighted co-occurrence metric to domain text has some advantages over paired comparisons and sorting tasks. The metric is more contextualized and specific because the associative contexts in which individual relational weights are established occur during thoughtful exposition of ideas in the creation of the text. Further, the relational metric method finds multiple instances of relatedness between terms in the text, where each instance can have a different context and a different degree of relatedness, and the final relational metric value takes all of these instances into account. Thus, the relational metric is more contextualized and specific than a single judgment concerning the overall degree of relatedness between isolated words.

As noted in the method section, the sliding window method of calculating the relational metric values can produce asymmetric results, so that the relational metric value between terms A and B is different from the value between B and A. The asymmetry arises when an active probe term appears more than once within the same context window because the method of calculation has the effect of preventing their contexts from fully overlapping. This is demonstrated in the example given in figures 2(b) and 2(c). The asymmetry is greater for relations involving terms which are densely distributed in the text, such that multiple instances of the term often occur within the context window. For example, the distance in number of words between two instances of the word "flow" is less than 22 (the size of the context window) a total of 47 times within the analyzed text. As a result, the method of calculation produces a relational metric value for R(flow,old) which is 23 percent smaller than the value for R(old,flow). In general, relations in which "flow" is the probe term have lower values than relations in which "flow" is a term-in-context. In the worst case, the value for R(flow,young) is 38 percent smaller than the value for R(young,flow). Among the top 164 relations, the median difference between the relational metric value of R(flow,X) and that of R(X,flow), when calculated using the asymmetric method, is 17 percent.

The density of other words in the text is much lower than that of "flow" so the asymmetry is greatly reduced for relational metric values involving these terms. The next most densely occurring word (after "flow") is "component." Instances of that word are separated by less than 22 words (the window size) only 15 times, compared with 47 times for "flow." The relational metric values for R(component,X) are not systematically larger or smaller than those for R(X,component), for any word X, and the values differ by an average of 6 percent. Other probe terms besides "flow" and "component" occur with much less density, and the relational metric values between each of them and other terms typically differ by only a few percent. Further, the relational metric values for R(A,B) and R(B,A), for all words A and B which are mutually related, are highly correlated (r = 0.96). It is also important to note that while the asymmetry decreases the weight of one of the two relations between two nodes, it has no effect on the other. That is, while the value of R(A,B) is decreased to some extent, the value of R(B,A) remains uninfluenced and still represents the maximum degree of relatedness between the two nodes. Thus, while avoidance of overlapping contexts does indeed introduce some asymmetries, they are limited and the results produced do effectively represent proximity-weighted co-occurrence relationships.

While the method produces effective results, a symmetrical version of the method of calculating the relational metric values would eliminate the differences in mutual relational metric values between pairs of words. These differences are introduced by the relative density of probe terms and the fact that contexts of closely neighboring instances of the same probe term are not allowed to fully overlap. Since these differences might not represent useful or meaningful domain information, their elimination could improve the effectiveness of the results. It would also simplify the networks by eliminating the directionality of relations and reducing the amount of data to be processed. A symmetrical version of the method has now been implemented for future application. Unlike the asymmetric method, it allows the contexts of neighboring instances of the same probe term to overlap, so that R(A,B)=R(B,A), as illustrated in figures 14(a) and 14(b).

Comparison of the network models obtained with the asymmetrical method (fig. 6) and the symmetrical version of the method (fig. 14(c)) indicate that the networks obtained barely differ with respect to which nodes and relationships are included in the domain model. What differences exist are minor and peripheral. For example, typical differences between figure 14(c) and figure 6 are that figure 14(c) adds a link of low weight (0.27) between the word "group" and the number "1.5" while it omits the node "brown," whose node weight is low (0.034) and whose largest link weight is 0.26 (see Appendix). These slight differences do not significantly change the character of the domain representation because the most prominent nodes and relations are unchanged. In addition, the top 164 relational metric values computed by the asymmetric method and those computed by the symmetrical version of the method are correlated (r = 0.89). As a consequence, the essential features of the model of the domain remain virtually unchanged. While the symmetric version of the method differs slightly from the asymmetric method and will henceforth be preferred to it for the sake of simplicity, the latter method does indeed produce effective results which are consistent with results obtained using the new method.

The decision to use a context window size equal to the average sentence length was an attempt to define a standard, linguistically reasonable verbal context around each probe term in the analyzed text. The effect of varying the context window size on the relational metric values and on the resulting domain model is an area in need of further investigation. While some preliminary work has been done on this, the early versions of the software that were used to conduct this study made the task extremely arduous. More efficient versions of the software for computing the relational metric values have just been developed, and these will enable further investigation. From the work done so far, however, it is clear that a very small window size emphasizes lexical co-occurrences such as the relation between "data" and "set," and adjective-noun pairs, such as "historic flow." At the same time, small context windows de-emphasize conceptual co-occurrences, such as the relations between "flow" and "age," "flow" and "image," and "flow" and "data." It also appears that as the context window size increases, the lexical relations remain at relatively low metric values, while the globally important relations rise rapidly in magnitude. It also appears that the rate of increase in the metric value of a relation, as context window size increases, is a function of the global importance of the relation. More work must be done in this area, not only for a thorough sensitivity analysis of the context window size, but also because it may lead to new and useful (but computationally costly) ways to rank the important relations in the domain.

An alternative to using a context window that reaches from one average sentence length before an occurrence of a probe term to one average sentence length after it, is to use the particular sentence in which each probe term occurs as the context for each occurrence. As an additional alternative, one could use the sentence containing the occurrence of the probe term, as well as the sentence preceding and the one following, as the context. It would be valuable, in a future study, to compute the relational metric values using these alternative contexts, and to

compare the results with those obtained using the fixed-size context window.

One question about the relational metric which must be considered in more detail is the degree to which the frequency of occurrence of each node in a pair of related nodes influences the relational metric value between them. If the text were random, or if it included such a diverse collection of themes that it had no thematic coherence, then a strong correlation between the product occ(X)*occ(Y) and the relational metric co-occ(X,Y) would indicate that the relatedness was likely due to chance (Church, Gale, Hanks, and Hindle, 1991). Since the analyzed text is thematically coherent, the product of the frequencies of occurrence of words X and Y, occ(X)*occ(Y), is not a reasonable estimate of chance co-occurrence. Instead, the fact that the authors of the text refer to word X very frequently and also to word Y very frequently would suggest that these are important words within the theme, and that the words are in fact likely to be closely related semantically, that is, related by the coherence of the theme of the text. If, however, the relational metric were highly correlated with the product of frequencies of occurrence, it would suggest that the metric provides little if any relatedness information beyond that indicated by the frequencies of the individual words.

Figures 15 and 16 indicate that the relational metric captures relatedness between pairs of words that is largely independent of the frequencies of the individual words, especially for those pairs having higher relational metric values. Figure 15 shows the correlation between the product of normalized (i.e., each value divided by the maximum value) frequencies of occurrence and the normalized relational metric for the top 164 relationships used to produce the object-oriented analysis in this paper, as well as the correlation for all 9075 relationships in R-list (all those with nonzero relational metric values). The correlations are weak, as indicated by the correlation coefficient values of 0.577 and 0.678 respectively. Further, squaring these values indicates that the percentage of the variance of the relational metric values that is due to the product of the frequencies of occurrence is 33.2 percent for the top 164 relations and 46.0 percent for all 9075 relations. These numbers suggest that there might be a relationship between the number of top relations considered and the influence of the frequency product on the variance of the relational metric. Figure 16 confirms that there is a relationship, and shows how that influence varies. These graphs show that the variance of the relational metric, co-occ(X,Y), is increasingly independent of the product of the frequencies of occurrence of the related nodes, occ(X)*occ(Y), as the number of less important relations decreases. (Less important relations are those with lower relational metric values and therefore less prominence, so their rank order or "relation number" is higher.) Figure 15 clearly shows this graphically. As dots representing less important relations are erased from the bottom of the figure, the value of r shrinks. Comparing figure 16 with figure 4 shows that the relations with the highest metric values are those least influenced by the frequency product of the related nodes. Thus, the relational metric captures a largely independent aspect of relatedness, especially for the higher relational metric values.

**Domain model**– The effectiveness of the domain analysis method, based on the proximity-weighted co-occurrence metric, is indicated by evidence that the method captures the essence of the domain structure, to the extent that it is contained in the analyzed text, in an explicit, object-oriented model of the domain. Review of the top relationships identified by the method, and reading of the domain text to which the method was applied, indicates that the few relationships having the highest relational metric values seem to be the most important among the many possible relationships in the text. The method assigns a relational metric value of zero to all but 9075 of the 621,732 nonreflexive relations among the 789 unique terms in the analyzed domain text. Of these 9075 relations, only 1222 have relational metric values of 25 to 314, only 378 have values of 50 to 314, only 164 have values of 75 to 314 as shown in figure 4. As demonstrated in the results section (see the section "Network Models Based on the Most Prominent Relationships in the Text" and fig. 5), a domain model containing merely the top five relations captures the most prominent structure of the domain text, that is to say, the five relations and the nodes they relate capture core components of the meaning of the analyzed text. The results section also shows that domain models based on the top 40 or 164 relations capture the essence of the domain model and the core of the text's meaning in greater detail. The domain model based on the top 164 relations is thoroughly described in the results section and especially in the Appendix, and it is clear that this is a detailed model of the key components of the domain, and that it captures in some detail the relational structure of the meaning of the text, despite the fact that it is based on only 164/621,732 = 0.0264 percent of the total number of relations in the domain text. This evidence argues in favor of the efficacy of the results, and it is consistent with Simon's (1968) "empty world hypothesis," discussed in the Introduction of this paper. That is, "for a tolerable description of reality only a tiny fraction of all possible interactions needs to be taken into account" (pg. 221).

Additional evidence in support of the efficacy of the method is provided by the Pathfinder networks derived from R-list, figures 9–11. As discussed above, while the Pathfinder networks are not ideal for object-oriented analysis, they do capture a genuinely meaningful network

representation of the data in association matrices. Based on the relational metric data, the three different Pathfinder networks contain reasonable semantic associations which are in fact contained in the domain text. This attests to the semantic coherence of the relational metric values and the effectiveness of the relational metric method introduced in this paper.

The step of describing the key domain entities and relationships is one which requires judgment, and is therefore less objective and repeatable than the automated steps of the method. This step includes identification of domain classes/objects, and assignment of attributes, attribute values, and actions to classes/objects. Use of the object-oriented paradigm does, however, ensure that the structure into which the domain is fit is one that is widely accepted as appropriate for mapping domains of human endeavor to explicit models and then to software. The internal structures of classes/objects are well defined, and constrained to include a small set of components, including: attributes, attribute values, and actions. Some practitioners might add more components to the structure of objects, but few would eliminate attributes, attribute values, or actions. Thus, the essential form of the domain model is already widely accepted.

Because the method produces a short, prioritized list of the most prominent entities and relationships, the judgment of the analyst in describing the prominent objects, attributes, actions, and relations is tightly focused. Entities are described by the analyst in order of importance (as determined by their relatedness to other entities), and this order is the same for every analyst. In addition, relationships are also described in a prioritized order used by every analyst. Given that the importance metrics, and thus the orders of entities and relationships, are completely objective and independent of the analyst, the process of describing the entities and relations is well structured. Further, in describing each relationship, only one pair of entities is considered at a time, which focuses the attention of the analyst.

The utility and appropriateness of the contents of the particular descriptions is supported by the correspondence of the descriptions to a reading understanding of the material contained in the source domain document, as described in the results section. Since the source text describes what is important to the authors in a part of their domain, the diagram and metrics derived from that text according to the relational metric method represent explicit, quantitative, and objective information about that part of the domain. While it would be easy to read the short source text used in this study to obtain a similar view of the domain, the result of mere reading would not be explicit, objective, or quantitative. Further, the analytical method can more easily scale up to rapidly and effectively process and model the contents of many documents, and larger bodies of text. By providing the description of the entities and relationships (see the Appendix) and the additional transformation of networks such as figure 6 to a more clearly object-oriented network of figure 8, the method provides software designers with domain models that can map directly into object-oriented designs. A mere reading knowledge of any number of texts, without producing explicit, objective, and quantitative models, would not provide the same benefit.

**Further Evaluation of the Method**

Application of the method to a variety of domains would help to test its effectiveness. Currently, for example, work is underway to apply the method to incident reports in an aeronautical database. Application to multiple domains should show whether the method captures the various structural features of the domains, and the similarities and differences among the domains that are evident by other means or observations. For example, when the method is applied to domains in which presence in natural environments is particularly important, those relations which are important in presence should be prominent. That is, relations of physical adjacency should dominate, so topological relations among environmental classes/objects will likely have high relational metric values, as will relations between objects in the environment and the explorer. In addition, the associations of attributes, attribute values, and actions with prominent physical entities should be among the prominent relations. Further, representations (e.g., images and data) should be much less prominent than environmental entities. For example, an interview of field geologists was analyzed in a previous study (McGreevy, 1994) without benefit of the current method of deriving relational metrics. To further test the method, it could be applied to a transcript of the interview and the results could be compared with those obtained previously. Similarly, prominent relations in domains such as Earth Observing System studies, where relations of physical presence are much less salient, should be quite different in character. Prominent relations in EOS are likely to be internal to the environmental, representational, and data classes/objects, between representational and data objects, and between those and environmental classes/objects.

Whether the relational metric method is effective for deriving domain models from large bodies of text remains to be determined. As the size of the analyzed text increases, not only are logistical challenges multiplied, but the structure of the contained domain model might become incoherent or unwieldy. The coding of text, in particular, involves considerable overhead when the

number of unique words in the text is large. A comparison of domain models derived from coded and uncoded texts would help to indicate the specific contribution of coding to the final product. It may be that for some purposes, differentiating the noun form of a word like "flow" from the verb form, or the rock form from the molten form, is too costly. A related problem is the derivation of models from a large collection of small texts. Is it possible to derive a meaningful core of commonality, or would the multiple topics lead to a patchwork domain model? These issues remain to be addressed.

If the method is used to prepare for interviewing domain experts and/or field observations of their work, further evaluation of the method could be obtained by using key findings from analysis of domain text to develop hypotheses, and then testing them in interview questions or field observations. The results of the current study, as represented by the network of 164 relations among 53 nodes in figure 6 and described in the Appendix, provide a rich collection of material that could be used in preparation for field interviews of Abrams and his colleagues (the authors of the analyzed text), and for on-site observations of their domain activities. For more comprehensive preparation, it would be valuable to analyze a broader selection of domain material which is still focused on a coherent group of investigators. The EOS Interdisciplinary Studies volcanology group is one example of an appropriate scope for further object-oriented relational analysis. Material made available by the group on Internet via World Wide Web (at http://www.geo.mtu.edu/eos/) provides exactly the kind of information needed to conduct such an analysis.

If the method described in this paper is used at a later stage of domain analysis and testbed design, such as to analyze field interviews in preparation for the design and development of a domain-oriented testbed, further evaluation of the domain model could be achieved. First, the analyses would be provided to software implementers in order to gauge their contribution to design and implementation. Further, as the implementation evolves, domain experts would exercise early prototypes of the testbed. If the implemented system meets the needs of the domain experts, this would support the argument that the method of object-oriented relational analysis is indeed effective, that is, that it captures the essential elements of their domain model in a form that is useful for implementation of domain-oriented software.

### Domain Models are Models of Presence

The entities and relations identified as being prominent in a domain are the ones used to construct a model of the domain. Thus, the entities and relations with which the domain expert is persistently engaged in the domain itself are those which comprise the domain model. This suggests that every domain model is a model of presence. That is, the immersion of a domain expert in a domain is a persistent engagement, governed by the dictates of the domain, with entities which are related by logical and physical adjacencies or continuities (McGreevy, 1993). These relations are also called metonymic relations (McGreevy, 1994). The nature and character of the domain determine which entities and which adjacencies are important. The proportion and distribution of the strictly logical adjacencies relative to physical adjacencies vary from domain to domain. In every domain, the domain expert is logically present. That is, the entities of interest are logically related, the expert is persistently engaged with these entities, and transitions among them traverse logical adjacency relations. To the extent that the persistently engaged domain entities are also physically adjacent to one another, the domain expert is also attendant to relations which are fundamental to physical presence among the domain entities. In this case, attentional shifts among the prominent entities will tend to traverse physical adjacency relations, while attentional shifts among entities which are only logically related will tend to traverse topical, categorical, definitional, or other relations which are discontinuous or disparate physically, but are part of the connected logical fabric or logical topology of the domain.

In the EOS domain analyzed in this paper, strictly logical presence dominates. The network domain model shown in figures 6 and 8 and the object-oriented relational analysis report in the Appendix are based upon the entities and relationships with which the EOS experts are persistently engaged. The node weights represent the degree of engagement with each node, and the relational weights represent the degree of engagement with each relationship. The most persistent engagement is with those nodes and relationships having the largest weights. Thus, the part of the EOS domain represented in the analyzed text is one in which domain experts are persistently engaged with a number of concepts, as indicated by the networks in figures 6 and 8 and the relational metric values in table 6. The most important of these concepts include, in order of importance: age-related attributes of volcanic lava flows [R(flow,flow) = 14.02], the relationships of these volcanic lava flows with the colors in laboratory images [R(flow,image) = 8.83], the relationships among colors in these images [R(image,image) = 5.20], and the relationships of these images with principal components, a combination of reflectance data which are assigned to the colors red, green, or blue to produce false color images of lava flows [R(image,component) = 2.58]. As such, it is a domain dominated by strictly logical

adjacencies among entities which are not physically adjacent. (The physical adjacency of colors in the images do not constitute domain-defining inter-object relations.) While these conceptual emphases of the small source text can be readily understood by reading it, the fact that they are also explicitly quantified in the domain model based on the relational metric method lends credence to the method.

The notion that such domain models can be analyzed as models of presence is supported by the ability of the domain model to explicitly, objectively, and quantitatively represent the conceptual emphases with which the authors of the source text were persistently engaged. That is, object-oriented domain models can represent presence because they model persistent engagement. More evidence for this would be provided by domain models of field geology and planetary surface exploration, including Apollo mission lunar surface exploration and robotic rover missions. Models of these domains, objectively and quantitatively derived using the relational metric method, should indicate a measurably greater degree of persistent engagement among physically adjacent entities, which would indicate the greater role of physical presence in these domains. Such models could be used to improve the designs of virtual environment systems for planetary exploration, and aid in a theoretical understanding of presence. In general, the relational metric method of domain analysis that is introduced here has the potential to produce useful domain models to guide the designs of a variety of computer-based systems, and contribute to a better understanding of the analyzed domains.

## Conclusion

This paper describes a relational metric method of verbal data analysis which produces object-oriented domain models that are explicit, objective, and quantified. The method produces models of the relational structure of domains, as represented in domain-produced verbal data, by computational means. Relational metric values are based on proximity-weighted frequencies of co-occurrence between a small number of probe terms and the terms in their contexts. Object-oriented relational analysis of the resulting domain structure produces a model of the domain in a form that is useful for software implementers. Models from related or very different domains can be compared and contrasted, providing the ability to observe structural similarities and differences. This can lead to a better understanding of the domains themselves, and to more effective and less expensive domain systems. One important use of the relational metric method of domain analysis is to investigate the role of logical and physical presence in a variety of domains, which can support development of a theory of presence, and improve the design of virtual environment systems.

## References


Abbott, R.: Program Design by Informal English Descriptions. Communications of the ACM, vol. 26, no. 11, 1983, pp. 882–894.

Abrams, M.; Abbott, E.; and Kahle, A.: Combined Use of Visible, Reflected Infrared, and Thermal Infrared Images for Mapping Hawaiian Lava Flows. J. Geophys. Res., vol. 96, no. B1, 1991, pp. 475–484.

Asrar, G.; and Dokken, D. J. (eds.): EOS Reference Handbook. NASA, Washington, D.C., 1993.

Bates, R. L.; and Jackson, J. A. (eds.): Glossary of Geology (3rd ed.). American Geological Institute, Alexandria, Va., 1987.

Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, G. A.: WordNet: A Lexical Database Organized on Psycholinguistic Principles. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon (pp. 211–232). Lawrence Erlbaum, Hillsdale, N.J., 1991.

Booch, G.: Object-Oriented Design. Benjamin/ Cummings, Redwood City, Calif., 1991.

Boose, J. H.: A Survey of Knowledge Acquisition Techniques and Tools. Knowledge Acquisition, 1, 1989, pp. 3–37.

Botts, M. E.: The Roles and Requirements of Visualization in NASA's EOS Era. Information Systems Newsletter, 1993, pp. 22–25.

Botts, M. E.: The State of Scientific Visualization with Regard to the NASA EOS Mission to Planet Earth: A Report to NASA Headquarters (rev-B). Earth System Science Lab, University of Alabama at Huntsville, 1993.

Chen, P. P.-S.: The Entity-Relationship Model: Toward a Unified View of Data. ACM Trans. on Database Systems, vol. 1, no. 1, 1976, pp. 9–36.

Chen, Z.: Human Aspects in Object-Oriented Design: An Assessment and a Methodology. Behaviour and Information Technology, vol. 11, no. 5, 1992, pp. 256–261.

Church, K.; Gale, W.; Hanks, P.; and Hindle, D.: Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon (pp. 115–164). Lawrence Erlbaum, Hillsdale, N.J., 1991.

Coad, P.; and Yourdan, E.: Object-Oriented Analysis. Yourdan Press, Englewood Cliffs, N.J., 1991.

Cooke, N. J.; and Schvaneveldt, R. W.: Effects of Computer Programming Experience on Network Representations of Abstract Programming Concepts. International Journal of Man-Machine Studies, vol. 29, no. 4, 1988, pp. 407–427.

Dillon, T. S.; and Tan, P. L.: Object-Oriented Conceptual Modeling. Prentice Hall, New York, 1993.

Dunphy, D. C.: The Construction of Categories for Content Analysis Dictionaries. In P. J. Stone and et al. (eds.), The General Inquirer (pp. 134-168). MIT Press, Cambridge, Mass., 1966.

Ericsson, K. A.; and Simon, H. A.: Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, Mass., 1984.

Fetterman, D. M.: Ethnography: Step by Step. Sage, Newbury Park, Calif., 1989.

Fichman, R.; and Kemerer, C.: Object-Oriented and Conventional Analysis and Design Methodologies. Computer, vol. 25, no. 10, 1992, pp. 22–39.

Gillan, D. J.; and Breedin, S. D.: Designers' Models of the Human-Computer Interface. In CHI '90. ACM, 1990, pp. 391–398

Graham, I.: Object-Oriented Methods. Addison-Wesley, Wokingham, England, 1994.

Grefenstette, G.; and Hearst, M. A.: A Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results (Technical Report 92/10). Sequoia 2000 project, UC Berkeley, 1992.

Jacobson, D.: Reading Ethnography. State University of New York, New York, 1991.

Jorgensen, D. L.: Participant Observation: A Methodology for Human Studies. Sage, Newbury Park, Calif., 1989.

Kaindl, H.: Comparing Object-Oriented Analysis with Knowledge Acquisition. OOPS Messenger, vol. 5, no. 3, 1994, pp. 1–5.

Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage, Newbury Park, Calif., 1980.

Kruskal, J. B.; and Wish, M.: Multidimensional Scaling. Sage, Beverly Hills, Calif., 1978.

Laurini, R.; and Thompson, D.: Fundamentals of Spatial Information Systems. Academic Press, London, 1992.

McDonald, J. E.; and Schvaneveldt, R. W.: The Application of User Knowledge to Interface Design. In R. Guindon (ed.), Cognitive Science and Its Applications for Human-Computer Interaction (pp. 289–338). Lawrence Erlbaum, Hillsdale, N.J., 1988.

McGreevy, M. W.: The Presence of Field Geologists in Mars-Like Terrain. Presence, vol. 1, no. 4, 1993, pp. 375–403.

McGreevy, M. W.: An Ethnographic Object-Oriented Analysis of Explorer Presence in a Volcanic Terrain Environment. NASA TM-108823. Ames Research Center, Moffett Field, Calif., 1994.

Miller, G. A.: Nouns in WordNet: A Lexical Inheritance System. In G. A. Miller, et. al., Five Papers on WordNet (CSL-43). Cognitive Science Lab, Princeton University, 1990.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K.: Five Papers on WordNet (CSL-43). Cognitive Science Lab, Princeton University, 1990.

Monarchi, D.; and Puhr, G.: A Research Typology for Object-Oriented Analysis and Design. Communications of the ACM, vol. 35, no. 9, 1992, pp. 35–47.

NASA: Kilauea: Compiled Volcanology Data (CD-ROM No. USA_NASA_PLDS_VOLC_01-06), 1992.

Nilsson, N. J.: Principles of Artificial Intelligence. Tioga, Palo Alto, Calif., 1980.

Osgood, C. E.: The Representation Model and Relevant Research Methods. In I. de Sola Pool (ed.), Trends in Content Analysis (pp. 33–88). Urbana: University of Illinois Press, 1959.

Quillian, M. R.: Semantic Memory. In M. Minsky (ed.), Semantic Information Processing (pp. 216–270). MIT Press, Cambridge, Mass., 1968.

Roske-Hofstrand, R. J.; and Paap, K. R.: Cognitive Networks as a Guide to Menu Organization: An Application in the Automated Cockpit. Ergonomics, 29, 1986, pp. 1301–1311.

Schütze, H.: Word Space. In S. J. Hanson, J. D. Cowan, and C. L. Giles (eds.), Advances in Neural Information Processing Systems (pp. 895–902). Morgan Kaufman, San Mateo, Calif., 1993.

Schvaneveldt, R. W.; Durso, F. T.; and Dearholt, D. W.: Network Structures in Proximity Data. In G. Bower (ed.), The Psychology of Learning and Motivation: Advanced in Research and Theory (pp. 249–284). Academic Press, New York, 1989.

Schvaneveldt, R. W.; Durso, F. T.; Goldsmith, T. E.; Breen, T. J.; and Cooke, N. M.: Measuring the Structure of Expertise. Int. J. Man-Machine Studies, 23, 1985, pp. 699–728.

Shaw, M. L. G.; and Woodward, J. B.: Modeling Expert Knowledge. Knowledge Acquisition, 2, 1990, pp. 179–206.

Simon, H. A.: The Sciences of the Artificial. MIT Press, Cambridge, Mass., 1969.

Smadja, F.: Macrocoding the Lexicon with Co-occurrence Knowledge. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon (pp. 165–189). Lawrence Erlbaum, Hillsdale, N.J., 1991.

Thomson, J.: Conc: A Concordance Generator. Summer Institute of Linguistics, Dallas, Tex., 1992.

Tracz, W.; Coglianese, L.; and Young, P.: A Domain-Specific Software Architecture Engineering Process Outline. Software Engineering Notes, vol. 18, no. 2, 1993, pp. 40–49.

Weber, R. P.: Basic Content Analysis (2nd ed.). Sage, Newbury Park, Calif., 1990.

Zernik, U. (ed.): Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Lawrence Erlbaum, Hillsdale, N.J., 1991.

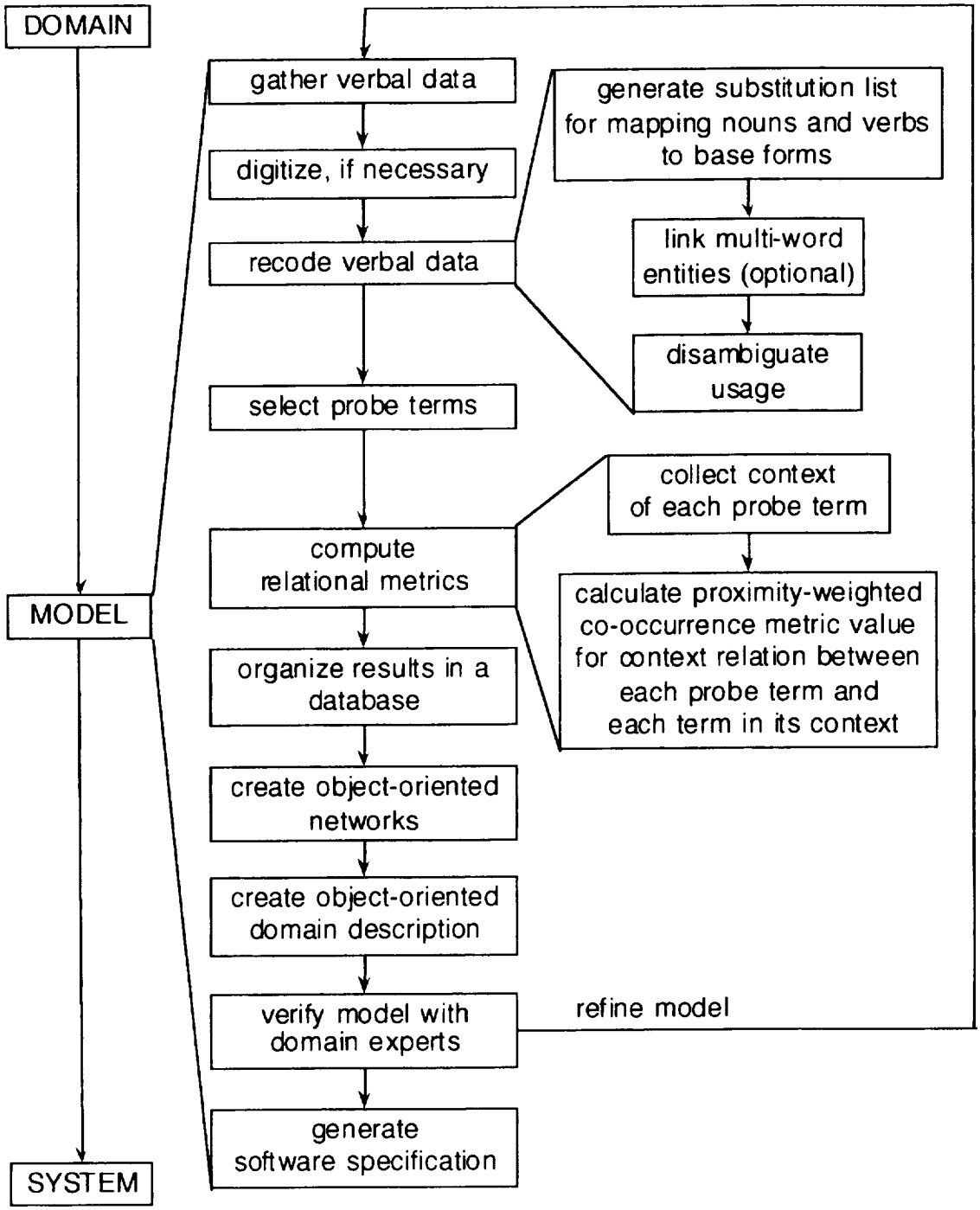



*Figure 1. Overview of a domain modeling process based on object-oriented verbal data analysis.*

| | <- - - - - - - - - - context of PT - - - - - - - - - -> | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | w1 | w2 | w3 | w4 | w5 | w6 | **PT** | w7 | w8 | w9 | w10 | w11 | w12 |
| a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | |
| b | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | |
| c | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | |
| d | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | |
| e | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | |
| f | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| g | | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| h | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0 | 1 | 2 | 3 | 4 | 5 | - | 5 | 4 | 3 | 2 | 1 | 0 |

*Figure 2(a). Example calculation of proximity-weighted co-occurrence relational metric values for 12 words in the context of one occurrence of one active probe term (PT). See text for explanation.*

| | <- - - - - - - - - - context of PT2 - - - - - - - - -> | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | **age** | <wd> | flow | <wd> | flow | color | | | | |
| | w0 | w1 | w2 | w3 | w4 | **PT2** | w6 | pt1a | w7 | pt1b | w9 | w10 | w11 | w12 | w13 |
| a | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | | | |
| b | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | | |
| c | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | |
| d | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | |
| e | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | |
| f | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | |
| g | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | |
| h | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| i | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| j | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 2 | 3 | 4 | 5 | - | 5 | **4** | 3 | **2** | 1 | 0 | 0 | 0 | 0 |

*Figure 2(b). Relational metric value between two probe terms, pt1 (e.g., "flow") and PT2 (e.g., "age"), in which PT2 is the active probe term (in uppercase bold type). The terms pt1a and pt1b are two instances of the same probe term, pt1. The relational metric value of pt1 in the context of PT2 is 4 + 2 = 6. Compare this figure with figure 2(c).*

| | <- - - - - - - context of PT1a -> | | | | | | | <- - context of PT1b - - - - - -> | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | age | <wd> | **flow** | <wd> | **flow** | color | | | | |
| | w1 | w2 | w3 | w4 | pt2 | w6 | **PT1a** | w7 | **PT1b** | w9 | w10 | w11 | w12 | w13 | w14 |
| a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | |
| b | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | | |
| c | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | | |
| d | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | | |
| e | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | |
| f | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | |
| g | | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | |
| h | | | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| i | | | | | | | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| j | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0 | 1 | 2 | 3 | **4** | 5 | - | 6 | - | 5 | 4 | 3 | 2 | 1 | 0 |

*Figure 2(c). Relational metric value between two probe terms, PT1 (e.g., "flow") and pt2 (e.g., "age") , in which PT1 is the active probe term (in uppercase bold type). PT1a and PT1b are two instances of the same probe term, PT1. Note that context of an active probe term never reaches beyond an instance of the same active probe term. The relational metric value of pt2 in the context of PT1 is 4. Compare this figure with figure 2(b).*

Abrams, M.; Abbott, E.; and Kahle, A.: Combined Use of Visible, Reflected Infrared, and Thermal Infrared Images for Mapping Hawaiian Lava Fows. J. Geophys. Res., vol. 96, no. B1, 1991, pp. 475–484.

The weathering of Hawaiian basalts is accompanied by chemical and physical changes of the surfaces. These changes have been mapped using remote sensing data from the visible and reflected infrared and thermal infrared wavelength region. They are related to the physical breakdown of surface chill coats, the development and erosion of silica coatings, the oxidation of mafic minerals, and the development of vegetation cover. These effects show systematic behavior with age and can be mapped using the image data and related to relative ages of pahoehoe and aa flows. The thermal data are sensitive to silica rind development and fine structure of the scene; the reflectance data show the degree of oxidation and differentiate vegetation from aa and cinders. Together, data from the two wavelength regions show more than either separately. The combined data potentially provide a powerful tool for mapping basalt flows in arid to semiarid volcanic environments.

*Figure 3. Reference and abstract of the text that was analyzed via the proximity-weighted co-occurrence relational metric.*

Table 1. The 40 most important relations in the domain sample, and the nodes involved in those relations. These are the first 40 of 9075 records in R-list, the tabular results database describing the relational structure of the domain sample. The database is sorted in decreasing order of the relational metric, co-occ(PT,TIC). The 40 records shown here have the 40 highest values of co-occ(PT, TIC).

| PT[a] | occ(PT)[b] | TIC[c] | occ(TIC)[d] | co-occ(PT,TIC)[e] value | co-occ(PT,TIC)[e] max – value + 1 |
|---|---|---|---|---|---|
| old | 24 | flow | 81 | max = 314 | 1 |
| aa | 19 | flow | 81 | 299 | 16 |
| young | 13 | flow | 81 | 296 | 19 |
| pahoehoe | 19 | flow | 81 | 287 | 28 |
| age | 32 | flow | 81 | 271 | 44 |
| flow | 81 | old | 24 | 241 | 74 |
| flow | 81 | age | 32 | 238 | 77 |
| image | 41 | flow | 81 | 224 | 91 |
| flow | 81 | pahoehoe | 19 | 222 | 93 |
| color | 35 | flow | 81 | 221 | 94 |
| tims | 21 | data | 44 | 220 | 95 |
| data | 44 | flow | 81 | 217 | 98 |
| flow | 81 | aa | 19 | 211 | 104 |
| color | 35 | image | 41 | 211 | 104 |
| image | 41 | color | 35 | 208 | 107 |
| blue | 18 | flow | 81 | 205 | 110 |
| group | 14 | flow | 81 | 194 | 121 |
| blue | 18 | green | 22 | 193 | 122 |
| green | 22 | flow | 81 | 192 | 123 |
| green | 22 | blue | 18 | 190 | 125 |
| flow | 81 | data | 44 | 186 | 129 |
| flow | 81 | young | 13 | 185 | 130 |
| flow | 81 | image | 41 | 184 | 131 |
| ns_oo1 | 13 | data | 44 | 183 | 132 |
| data | 44 | tims | 21 | 181 | 134 |
| flow | 81 | color | 35 | 172 | 143 |
| reflectance | 26 | band | 17 | 169 | 146 |
| flow | 81 | group | 14 | 167 | 148 |
| use_verb | 18 | data | 44 | 166 | 149 |
| flow | 81 | green | 22 | 165 | 150 |
| flow | 81 | blue | 18 | 161 | 154 |
| image | 41 | data | 44 | 153 | 162 |
| data | 44 | use_verb | 18 | 152 | 163 |
| dark | 11 | green | 22 | 152 | 163 |
| old | 24 | year | 12 | 148 | 167 |
| year | 12 | old | 24 | 146 | 169 |
| band | 17 | reflectance | 26 | 146 | 169 |
| data | 44 | ns_oo1 | 13 | 145 | 170 |
| green | 22 | component | 27 | 141 | 174 |
| relative | 16 | age | 32 | 140 | 175 |

[a]Probe term.

[b]PT's frequency of occurrence within the body of the text (highest is 81).

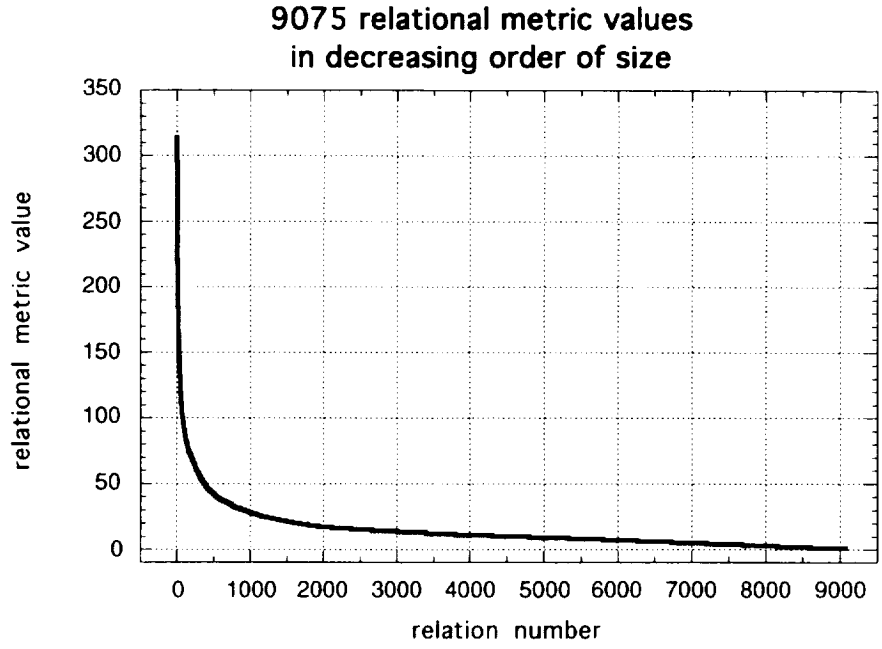
[c]Term-in-context which was found in the context of PT within the text.

[d]TIC's frequency of occurrence within the body of the text (highest is 81).

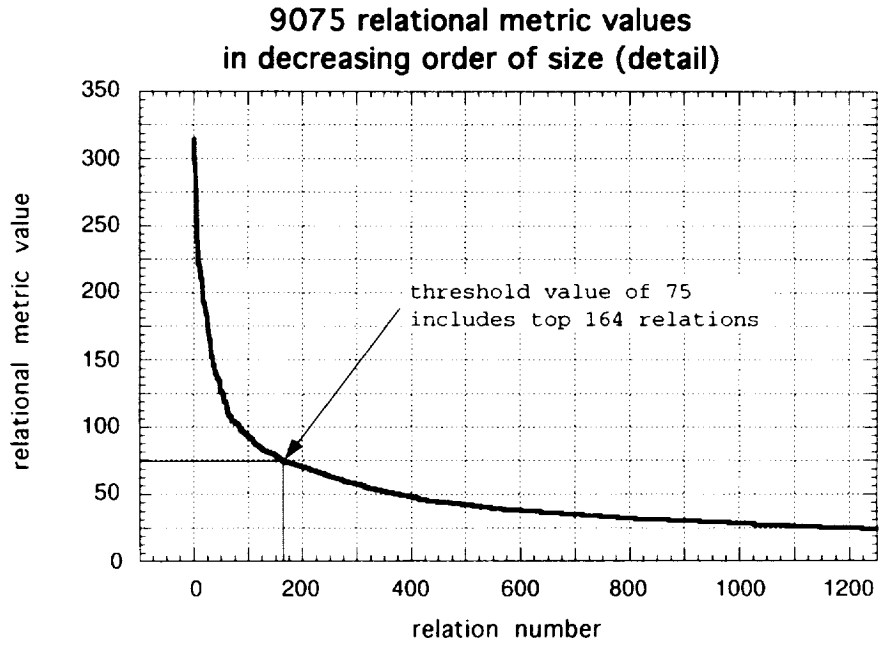
[e]Proximity-weighted co-occurrence value (the relational metric) between PT and TIC.

Table 2. The 50 probe terms, each with its frequency of occurrence in the domain sample. As will be seen later, not all of these terms are included as probe terms among the terms participating in the top 164 relationships (see fig. 6 and table 5). Instead, some of these terms are included only in the contexts of other terms. These are shown in parentheses. Further, some of these terms are not included at all among the top 164 relationships. These are shown in square brackets. The terms "change", "show", and "combine" are verbs but have no "_verb" tag because in the analyzed text they have no noun usages from which they must be distinguished.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| flow | 81 | use_verb | 18 | [spectral] | 13 | (brown) | 9 |
| image | 44 | band | 17 | thermal | 13 | [channel] | 9 |
| data | 41 | iron | 17 | (we) | 13 | difference | 9 |
| color | 35 | show | 17 | young | 13 | ferric | 9 |
| age | 32 | infrared | 16 | [surface] | 12 | [oxidation] | 9 |
| component | 27 | relative | 16 | year | 12 | plate | 9 |
| reflectance | 26 | group | 14 | dark | 11 | [silica] | 9 |
| old | 24 | red | 14 | emittance | 11 | study | 9 |
| green | 22 | area | 13 | ka | 11 | weathering | 9 |
| tims | 21 | [change] | 13 | [map_verb] | 11 | [unit] | 8 |
| aa | 19 | field | 13 | (spectrum) | 11 | visible | 8 |
| pahoehoe | 19 | micron | 13 | combine | 10 | | |
| blue | 18 | ns_oo1 | 13 | [vegetation] | 10 | | |




(a)




(b)

*Figure 4. Relational metric values for the 9075 proximity-weighted co-occurrence relations identified in the analyzed text. (a) 9075 relational metric values in decreasing order of size, (b) 9075 relational metric values in decreasing order of size (detail).*

Table 3. Association matrix extracted from the 789 × 789 R-matrix implicit in the R-list data base. The cells in the matrix represent relations among the 21 nodes, and numbers in the cells are the relational metric values of the relations. Only the top 40 relations are shown. This association matrix corresponds to the list of relations in table 1 and the network in figure 5.

| TIC / PT | frequency | flow | data | image | color | age | component | reflectance | old | green | TIMS | aa | pahoehoe | blue | use_verb | band | relative | group | NS-001 | young | year | dark |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| frequency | | 81 | 44 | 41 | 35 | 32 | 27 | 26 | 24 | 22 | 21 | 19 | 19 | 18 | 18 | 17 | 16 | 14 | 13 | 13 | 12 | 11 |
| flow | 81 | | 186 | 184 | 172 | 238 | | | 241 | 165 | | 211 | 222 | 161 | | | | 167 | | 185 | | |
| data | 44 | 217 | | | | | | | | | 181 | | | | 152 | | | | 145 | | | |
| image | 41 | 224 | 153 | | 208 | | | | | | | | | | | | | | | | | |
| color | 35 | 221 | | 211 | | | | | | | | | | | | | | | | | | |
| age | 32 | 271 | | | | | | | | | | | | | | | | | | | | |
| component | 27 | | | | | | | | | | | | | | | | | | | | | |
| reflectance | 26 | | | | | | | | | | | | | | | 169 | | | | | | |
| old | 24 | 314 | | | | | | | | | | | | | | | | | | | 148 | |
| green | 22 | 192 | | | | | 141 | | | | | | | 190 | | | | | | | | |
| TIMS | 21 | | 220 | | | | | | | | | | | | | | | | | | | |
| aa | 19 | 299 | | | | | | | | | | | | | | | | | | | | |
| pahoehoe | 19 | 287 | | | | | | | | | | | | | | | | | | | | |
| blue | 18 | 205 | | | | | | | | 193 | | | | | | | | | | | | |
| use_verb | 18 | | 166 | | | | | | | | | | | | | | | | | | | |
| band | 17 | | | | | | | 146 | | | | | | | | | | | | | | |
| relative | 16 | | | | | 140 | | | | | | | | | | | | | | | | |
| group | 14 | 194 | | | | | | | | | | | | | | | | | | | | |
| NS-001 | 13 | | 183 | | | | | | | | | | | | | | | | | | | |
| young | 13 | 296 | | | | | | | | | | | | | | | | | | | | |
| year | 12 | | | | | | | | 146 | | | | | | | | | | | | | |
| dark | 11 | | | | | | | | | 152 | | | | | | | | | | | | |

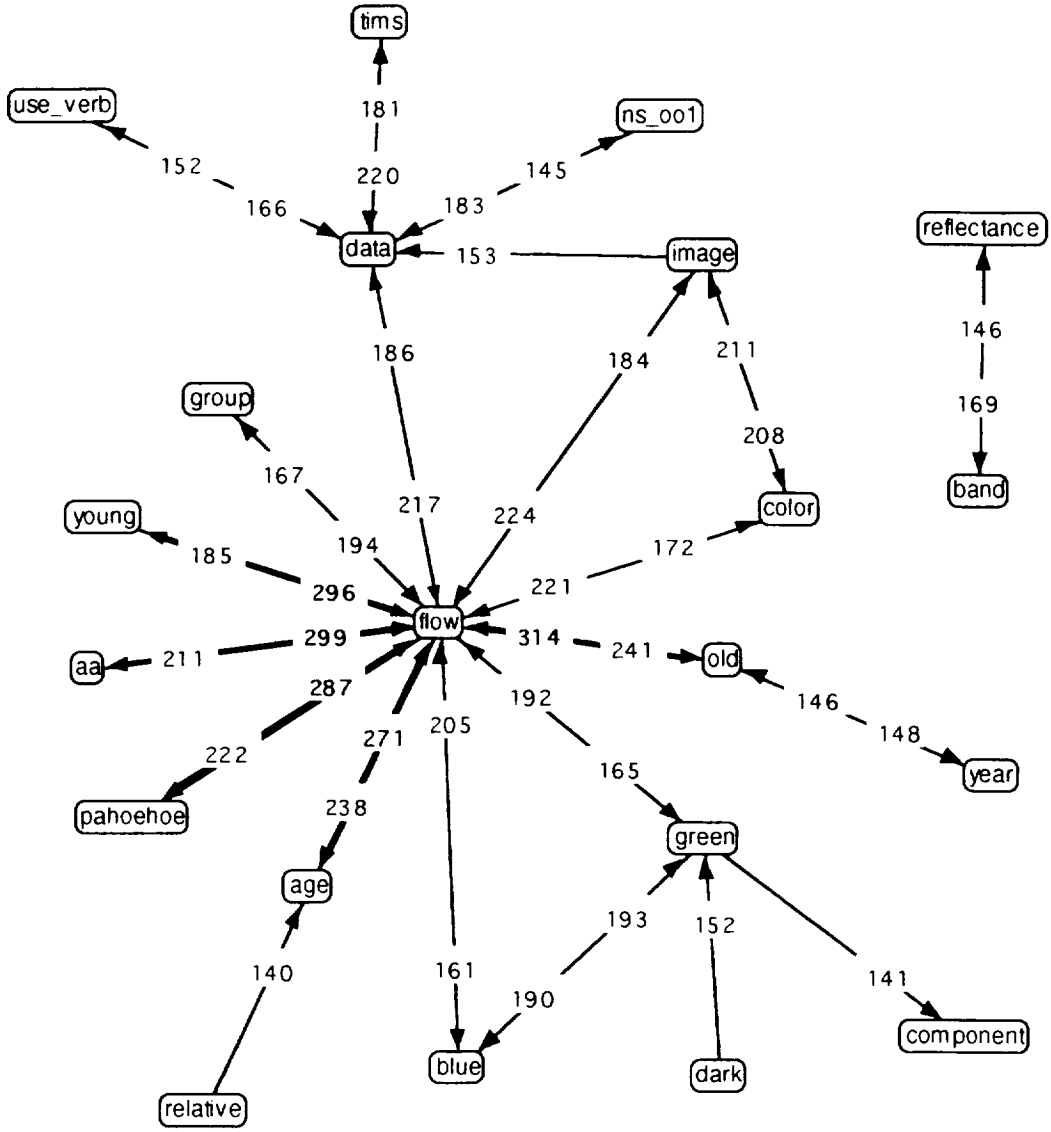



*Figure 5. Network based on the 40 most important relations in the domain sample. The top five relations are shown with bold arcs. This network corresponds to the list of relations in table 1 and the association matrix in table 3.*

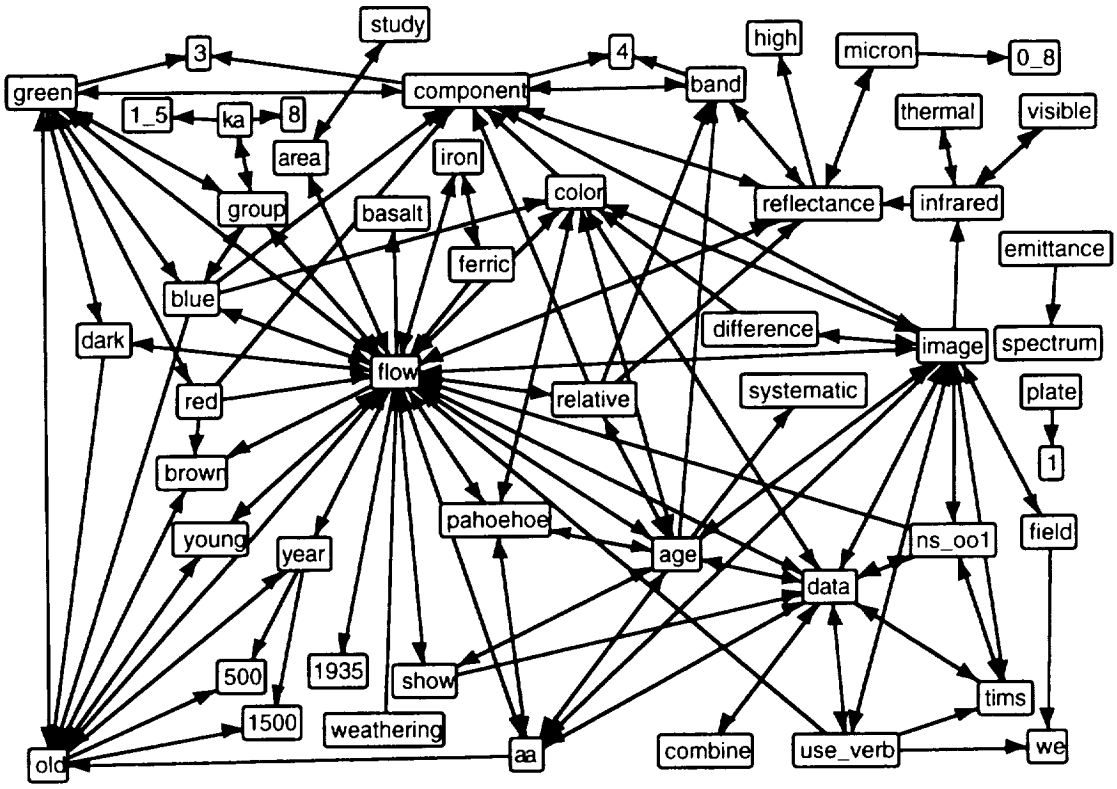



*Figure 6. Network domain model based on the threshold method. The network includes the 164 top relations, which associate 53 nodes. Of the 53 nodes, 41 are probe terms. In the Appendix, the terms (nodes) and relationships (arcs) of this network are described, and the relative importance of each is quantified.*

Table 4. Example of calculation of N, the estimate of node importance used to determine the node order in the Appendix. N is a measure of the overall relatedness of each term. It is equal to the sum of the normalized relational weights for all relations involving the term, divided by the largest value of N. Key: PT = probe term, TIC = term in context, "Norm'd co-occ" is the normalized co-occurrance metric. That is, the value of the metric is divided by the maximum metric value of the 9075 relations in R-list: co-occ(old, flow) = 314.

| PT or TIC = "flow" | | | PT or TIC = "image" | | |
|---|---|---|---|---|---|
| PT | TIC | norm'd co-occ | PT | TIC | norm'd co-occ |
| old | flow | 1.000 | image | flow | 0.713 |
| aa | flow | 0.952 | color | image | 0.672 |
| young | flow | 0.943 | image | color | 0.662 |
| pahoehoe | flow | 0.914 | flow | image | 0.586 |
| age | flow | 0.863 | image | data | 0.487 |
| flow | old | 0.768 | image | tims | 0.443 |
| flow | age | 0.758 | tims | image | 0.433 |
| image | flow | 0.713 | data | image | 0.401 |
| flow | pahoehoe | 0.707 | image | infrared | 0.325 |
| color | flow | 0.704 | image | use_verb | 0.309 |
| data | flow | 0.691 | image | field | 0.306 |
| flow | aa | 0.672 | use_verb | image | 0.299 |
| blue | flow | 0.653 | image | component | 0.290 |
| group | flow | 0.618 | image | age | 0.274 |
| green | flow | 0.611 | aa | image | 0.274 |
| flow | data | 0.592 | field | image | 0.274 |
| flow | young | 0.589 | ns_oo1 | image | 0.268 |
| flow | image | 0.586 | image | aa | 0.264 |
| flow | color | 0.548 | age | image | 0.261 |
| flow | group | 0.532 | image | difference | 0.255 |
| flow | green | 0.525 | difference | image | 0.255 |
| flow | blue | 0.513 | image | ns_oo1 | 0.245 |
| show | flow | 0.433 | component | image | 0.239 |
| relative | flow | 0.430 | Total relatedness of "image" to others: | | 8.535 |
| dark | flow | 0.398 | | | |
| flow | show | 0.354 | | | |
| flow | relative | 0.338 | Example calculations of N(node) used in Appendix, from total relatedness: | | |
| reflectance | flow | 0.331 | | | |
| year | flow | 0.328 | | | |
| flow | year | 0.322 | N(flow) = 22.253 / 22.253 = 1.000 | | |
| ns_oo1 | flow | 0.322 | N(image) = 8.535 / 22.253 = 0.384 | | |
| area | flow | 0.315 | etc. | | |
| flow | reflectance | 0.303 | | | |
| flow | area | 0.299 | | | |
| flow | dark | 0.290 | | | |
| iron | flow | 0.283 | | | |
| red | flow | 0.271 | | | |
| flow | iron | 0.268 | | | |
| flow | 1935 | 0.261 | | | |
| use_verb | flow | 0.258 | | | |
| weathering | flow | 0.258 | | | |
| ferric | flow | 0.255 | | | |
| flow | brown | 0.245 | | | |
| flow | basalt | 0.239 | | | |
| Total relatedness of "flow" to others: | | 22.253 | | | |

Table 5. List of 164 top relationships, sorted by frequency of occurrence of probe terms (PT) then by the normalized co-occurrence of PT and TIC (term-in-context): norm'd co-occ(PT,TIC). This list serves as the starting point for creation of the object-oriented relational analysis report (Appendix).

| PT | TIC | norm'd co-occ(PT,TIC) |
|---|---|---|
| flow | old | 0.77 |
| flow | age | 0.76 |
| flow | pahoehoe | 0.71 |
| flow | aa | 0.67 |
| flow | data | 0.59 |
| flow | young | 0.59 |
| flow | image | 0.59 |
| flow | color | 0.55 |
| flow | group | 0.53 |
| flow | green | 0.53 |
| flow | blue | 0.51 |
| flow | show | 0.35 |
| flow | relative | 0.34 |
| flow | year | 0.32 |
| flow | reflectance | 0.30 |
| flow | area | 0.30 |
| flow | dark | 0.29 |
| flow | iron | 0.27 |
| flow | 1935 | 0.26 |
| flow | brown | 0.25 |
| flow | basalt | 0.24 |
| image | flow | 0.71 |
| image | color | 0.66 |
| image | data | 0.49 |
| image | tims | 0.44 |
| image | infrared | 0.32 |
| image | use_verb | 0.31 |
| image | field | 0.31 |
| image | component | 0.29 |
| image | age | 0.27 |
| image | aa | 0.26 |
| image | difference | 0.25 |
| image | ns_001 | 0.25 |
| data | flow | 0.69 |
| data | tims | 0.58 |
| data | use_verb | 0.48 |
| data | ns_001 | 0.46 |
| data | image | 0.40 |
| data | age | 0.36 |
| data | combine | 0.32 |
| data | color | 0.29 |
| data | aa | 0.26 |
| color | flow | 0.70 |
| color | image | 0.67 |
| color | age | 0.38 |
| color | component | 0.33 |
| color | data | 0.30 |
| color | pahoehoe | 0.27 |
| age | flow | 0.86 |
| age | color | 0.43 |
| age | relative | 0.41 |
| age | data | 0.34 |
| age | image | 0.26 |
| age | systematic | 0.25 |
| age | aa | 0.25 |
| age | band | 0.25 |
| age | show | 0.25 |
| age | pahoehoe | 0.24 |
| component | reflectance | 0.43 |
| component | band | 0.40 |
| component | green | 0.32 |
| component | 3 | 0.32 |
| component | 4 | 0.31 |
| component | image | 0.24 |
| reflectance | band | 0.54 |
| reflectance | component | 0.38 |
| reflectance | flow | 0.33 |
| reflectance | micron | 0.28 |
| reflectance | high | 0.27 |
| old | flow | 1.00 |
| old | year | 0.47 |
| old | green | 0.39 |
| old | 500 | 0.29 |
| old | 1500 | 0.27 |
| old | brown | 0.26 |
| old | young | 0.25 |
| green | flow | 0.61 |
| green | blue | 0.61 |
| green | component | 0.45 |
| green | old | 0.39 |
| green | dark | 0.34 |
| green | group | 0.29 |
| green | red | 0.24 |
| green | 3 | 0.24 |
| tims | data | 0.70 |
| tims | image | 0.43 |
| tims | ns_001 | 0.33 |
| aa | flow | 0.95 |
| aa | data | 0.32 |
| aa | pahoehoe | 0.31 |
| aa | old | 0.28 |
| aa | image | 0.27 |
| aa | age | 0.26 |

| | | |
|---|---|---|
| pahoehoe | flow | 0.91 |
| pahoehoe | aa | 0.31 |
| pahoehoe | color | 0.30 |
| pahoehoe | age | 0.26 |
| blue | flow | 0.65 |
| blue | green | 0.61 |
| blue | component | 0.33 |
| blue | color | 0.29 |
| blue | old | 0.28 |
| blue | group | 0.28 |
| use_verb | data | 0.53 |
| use_verb | image | 0.30 |
| use_verb | we | 0.26 |
| use_verb | flow | 0.26 |
| use_verb | tims | 0.25 |
| band | reflectance | 0.46 |
| band | component | 0.36 |
| band | 4 | 0.26 |
| iron | ferric | 0.31 |
| iron | flow | 0.28 |
| show | flow | 0.43 |
| show | data | 0.25 |
| show | age | 0.25 |
| infrared | thermal | 0.31 |
| infrared | visible | 0.30 |
| infrared | reflectance | 0.24 |
| relative | age | 0.45 |
| relative | flow | 0.43 |
| relative | band | 0.33 |
| relative | reflectance | 0.29 |
| relative | component | 0.25 |
| group | flow | 0.62 |
| group | ka | 0.33 |
| group | green | 0.30 |
| group | blue | 0.29 |
| red | component | 0.38 |
| red | flow | 0.27 |
| red | brown | 0.24 |
| area | study | 0.38 |
| area | flow | 0.32 |
| field | image | 0.27 |
| field | we | 0.25 |
| micron | reflectance | 0.35 |
| micron | 0_8 | 0.26 |
| ns_oo1 | data | 0.58 |
| ns_oo1 | tims | 0.34 |
| ns_oo1 | flow | 0.32 |
| ns_oo1 | image | 0.27 |
| thermal | infrared | 0.44 |
| young | flow | 0.94 |
| young | old | 0.26 |
| year | old | 0.46 |
| year | flow | 0.33 |
| year | 500 | 0.29 |
| year | 1500 | 0.27 |
| dark | green | 0.48 |
| dark | flow | 0.40 |
| dark | old | 0.26 |
| emittance | spectrum | 0.25 |
| ka | group | 0.32 |
| ka | 1_5 | 0.30 |
| ka | 8 | 0.26 |
| combine | data | 0.37 |
| difference | image | 0.25 |
| difference | color | 0.25 |
| ferric | iron | 0.40 |
| ferric | flow | 0.25 |
| plate | 1 | 0.27 |
| study | area | 0.34 |
| weathering | flow | 0.26 |
| visible | infrared | 0.44 |

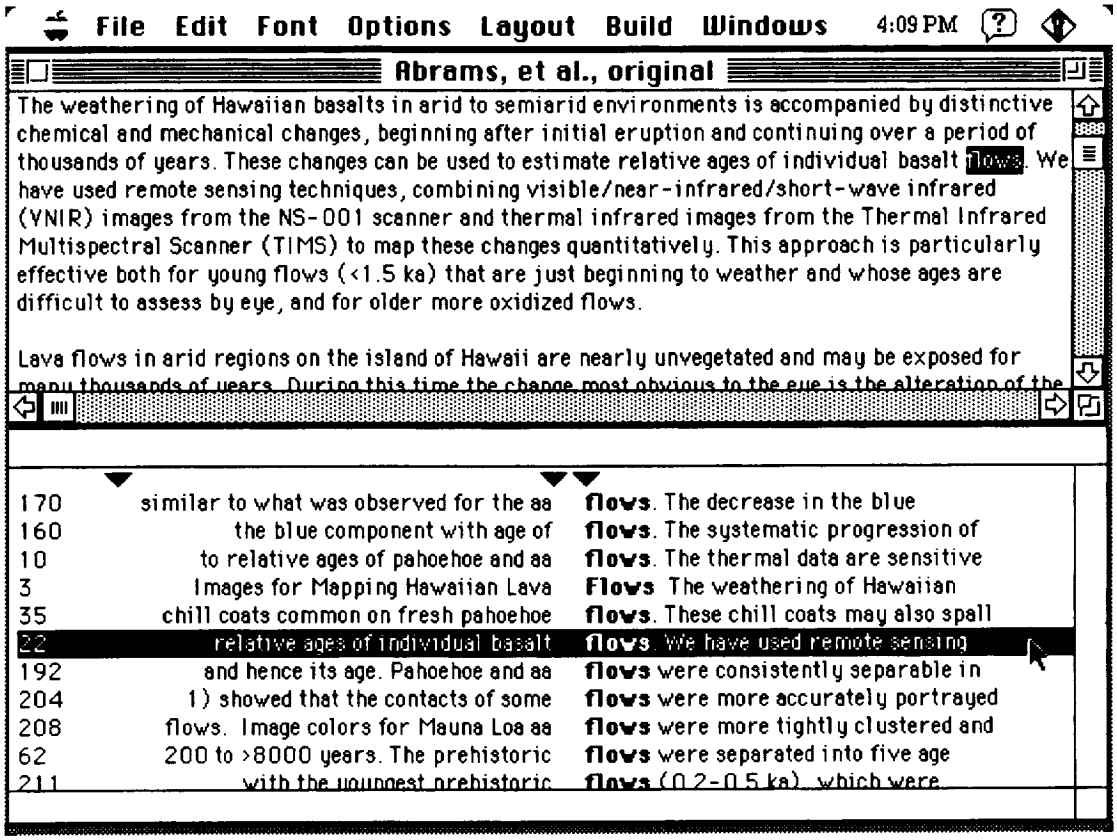



*Figure 7. Example of use of an interactive concordance (Thomson, 1992) showing a computer screen image as it appears while being used to search the original text for the node term, "flow." The window at the bottom shows some of the contexts around the term "flow" (the rest are available by scrolling) while the window at the top contains the full text context for any line selected in the bottom window. (In practice, the windows are made much larger on the computer screen, so as to display more of the contexts.)*

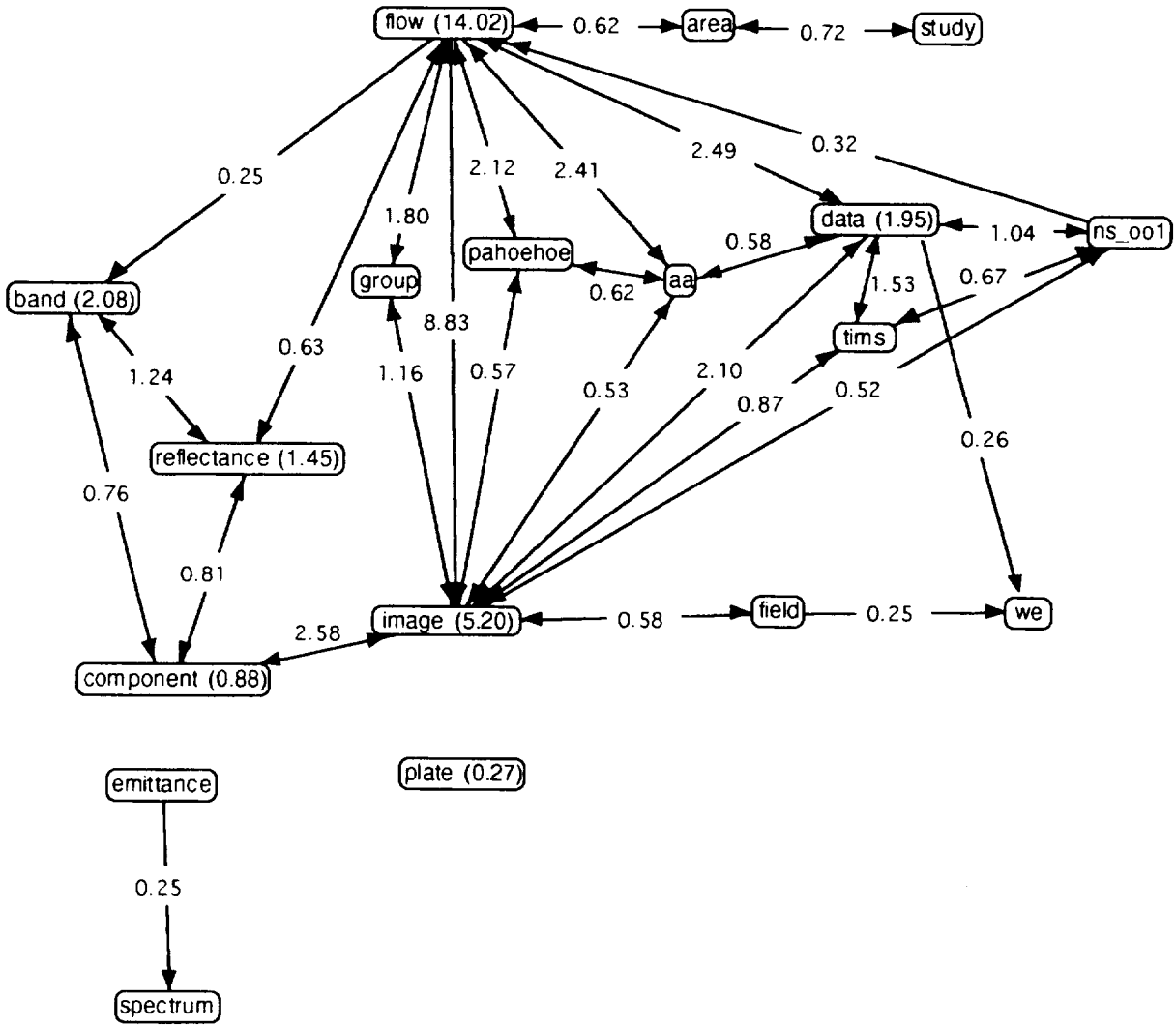



*Figure 8. Top 164 relationships, combined into inter-class relations (shown as one-or two-way arcs), and node weights indicating the internal complexity of classes/objects based on the sum of intra-class/object relational weights. Nodes without weights have no internal structural relations among the top 164 domain relations. Thus, for example, the internal structure of the class/object "flow" is the most elaborated, and "image" has the second most elaborated internal structure. Furthermore, the largest sum of inter-class/object relational weights is between "flow" and "image" indicating their close association in this domain sample. The 164 relationships and their mapping to class/object relationships are shown in table 6.*

Table 6. Top 164 relationships mapped to class/object relations, listed in alphabetical order by class/object, and by co-occ when the PT class/object equals the TIC class/object. Items not in parentheses are classes/objects, while items in parentheses are internals of classes/objects. These relationships are the basis of figure 8.

| PT | TIC | norm'd co-occ(PT,TIC) |
|---|---|---|
| aa | data | 0.32 |
| aa | flow | 0.95 |
| aa | flow(old) | 0.28 |
| aa | flow(age) | 0.26 |
| aa | image | 0.27 |
| aa | pahoehoe | 0.31 |
| area | flow | 0.32 |
| area | study | 0.38 |
| band(thermal) | band(infrared) | 0.44 |
| band(visible) | band(infrared) | 0.44 |
| band(relative) | band | 0.33 |
| band(infrared) | band(thermal) | 0.31 |
| band(infrared) | band(visible) | 0.30 |
| band | band(4) | 0.26 |
| band | component | 0.36 |
| band | reflectance | 0.46 |
| band(infrared) | reflectance | 0.24 |
| component | band | 0.40 |
| component | component(3) | 0.32 |
| component | component(4) | 0.31 |
| component(relative) | component | 0.25 |
| component | image(green) | 0.32 |
| component | image | 0.24 |
| component | reflectance | 0.43 |
| data | aa | 0.26 |
| data(use_verb) | data | 0.53 |
| data | data(use_verb) | 0.48 |
| data(combine) | data | 0.37 |
| data | data(combine) | 0.32 |
| data(show) | data | 0.25 |
| data | flow | 0.69 |
| data | flow(age) | 0.36 |
| data(use_verb) | flow | 0.26 |
| data | image | 0.40 |
| data(use_verb) | image | 0.30 |
| data | image(color) | 0.29 |
| data | ns_001 | 0.46 |
| data | tims | 0.58 |
| data(use_verb) | tims | 0.25 |
| data(use_verb) | we | 0.26 |
| emittance | spectrum | 0.25 |
| field | image | 0.27 |
| field | we | 0.25 |
| flow | aa | 0.67 |
| flow(age) | aa | 0.25 |
| flow | area | 0.30 |
| flow(age) | band | 0.25 |
| flow | data | 0.59 |
| flow(age) | data | 0.34 |
| flow(age) | data(show) | 0.25 |
| flow(old) | flow | 1.00 |
| flow(young) | flow | 0.94 |
| flow(age) | flow | 0.86 |
| flow | flow(old) | 0.77 |
| flow | flow(age) | 0.76 |
| flow | flow(young) | 0.59 |
| flow(old) | flow(year) | 0.47 |
| flow(year) | flow(old) | 0.46 |
| flow(relative) | flow(age) | 0.45 |
| flow(relative) | flow | 0.43 |
| flow(show) | flow | 0.43 |
| flow(age) | flow(relative) | 0.41 |
| flow(dark) | flow | 0.40 |
| flow(ferric) | flow(iron) | 0.40 |
| flow | flow(show) | 0.35 |
| flow | flow(relative) | 0.34 |
| flow(year) | flow | 0.33 |
| flow | flow(year) | 0.32 |
| flow(iron) | flow(ferric) | 0.31 |
| flow(ka) | flow(1_5) | 0.30 |
| flow(old) | flow(500) | 0.29 |
| flow(year) | flow(500) | 0.29 |
| flow(iron) | flow | 0.28 |
| flow | flow(iron) | 0.27 |
| flow(old) | flow(1500) | 0.27 |
| flow(year) | flow(1500) | 0.27 |
| flow | flow(1935) | 0.26 |
| flow(ka) | flow(8) | 0.26 |
| flow(weathering) | flow | 0.26 |
| flow(young) | flow(old) | 0.26 |
| flow(ferric) | flow | 0.25 |
| flow(old) | flow(young) | 0.25 |
| flow(show) | flow(age) | 0.25 |
| flow | flow(basalt) | 0.24 |
| flow | group | 0.53 |
| flow(ka) | group | 0.32 |
| flow | image | 0.59 |
| flow | image(color) | 0.55 |
| flow | image(green) | 0.53 |
| flow | image(blue) | 0.51 |
| flow(age) | image(color) | 0.43 |
| flow(old) | image(green) | 0.39 |
| flow | image(dark) | 0.29 |
| flow(age) | image | 0.26 |
| flow(old) | image(brown) | 0.26 |
| flow | image(brown) | 0.25 |
| flow(age) | image(systematic) | 0.25 |
| flow | pahoehoe | 0.71 |
| flow(age) | pahoehoe | 0.24 |
| flow | reflectance | 0.30 |
| group | flow | 0.62 |
| group | flow(ka) | 0.33 |
| group | image(green) | 0.30 |
| group | image(blue) | 0.29 |

| | | |
|---|---|---|
| image | aa | 0.26 |
| image(green) | component | 0.45 |
| image(red) | component | 0.38 |
| image(blue) | component | 0.33 |
| image(color) | component | 0.33 |
| image | component | 0.29 |
| image(green) | component(3) | 0.24 |
| image | data | 0.49 |
| image | data(infrared) | 0.32 |
| image(color) | data | 0.30 |
| image | field | 0.31 |
| image | flow | 0.71 |
| image(color) | flow | 0.70 |
| image(blue) | flow | 0.65 |
| image(green) | flow | 0.61 |
| image(green) | flow(old) | 0.39 |
| image(color) | flow(age) | 0.38 |
| image(blue) | flow(old) | 0.28 |
| image | flow(age) | 0.27 |
| image(red) | flow | 0.27 |
| image(dark) | flow(old) | 0.26 |
| image(green) | group | 0.29 |
| image(blue) | group | 0.28 |
| image(color) | image | 0.67 |
| image | image(color) | 0.66 |
| image(blue) | image(green) | 0.61 |
| image(green) | image(blue) | 0.61 |
| image(dark) | image(green) | 0.48 |
| image(green) | image(dark) | 0.34 |
| image | image(use_verb) | 0.31 |
| image(blue) | image(color) | 0.29 |
| image | image(difference) | 0.25 |
| image(difference) | image | 0.25 |
| image(difference) | image(color) | 0.25 |
| image(green) | image(red) | 0.24 |
| image(red) | image(brown) | 0.24 |
| image | ns_oo1 | 0.25 |
| image(color) | pahoehoe | 0.27 |
| image | tims | 0.44 |
| ns_oo1 | data | 0.58 |
| ns_oo1 | flow | 0.32 |
| ns_oo1 | image | 0.27 |
| ns_oo1 | tims | 0.34 |
| pahoehoe | aa | 0.31 |
| pahoehoe | flow | 0.91 |
| pahoehoe | flow(age) | 0.26 |
| pahoehoe | image(color) | 0.30 |
| plate | plate(1) | 0.27 |
| reflectance | band | 0.54 |
| reflectance | component | 0.38 |
| reflectance | flow | 0.33 |
| reflectance(micron) | reflectance | 0.35 |
| reflectance(relative) | reflectance | 0.29 |
| reflectance | reflectance(micron) | 0.28 |
| reflectance | reflectance(high) | 0.27 |
| reflectance(micron) | reflectance(0_8) | 0.26 |
| study | area | 0.34 |
| tims | data | 0.70 |
| tims | image | 0.43 |
| tims | ns_oo1 | 0.33 |

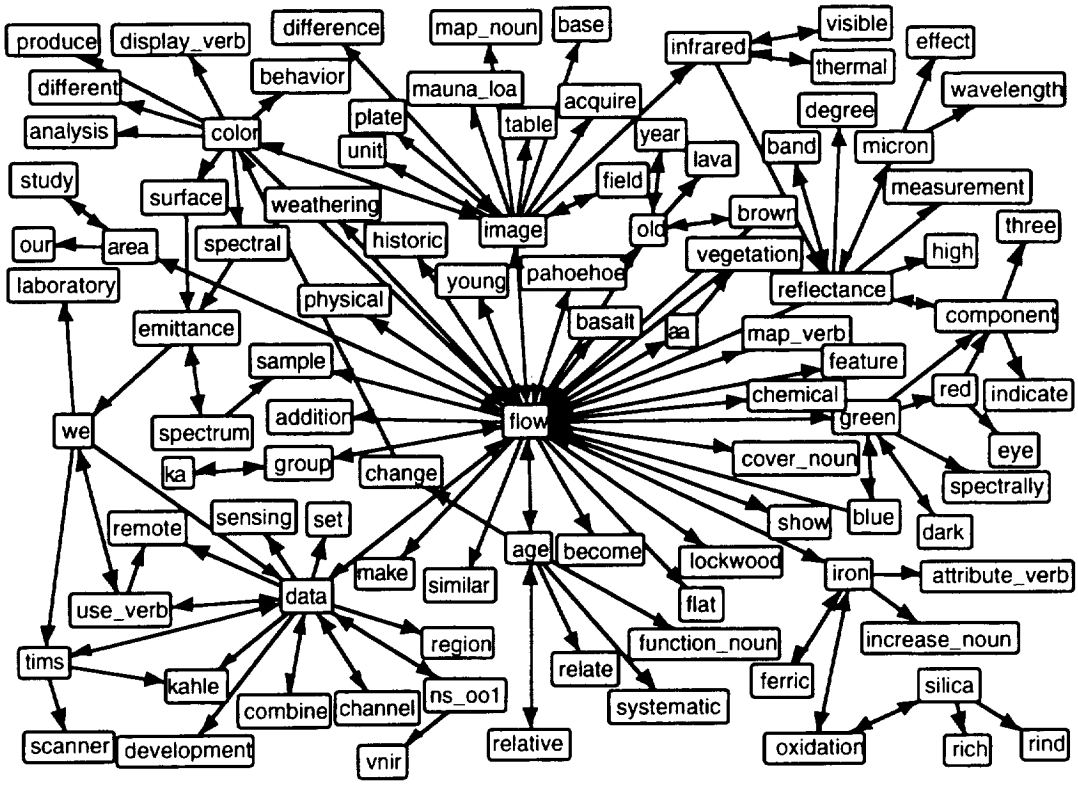



*Figure 9(a). Pathfinder network with 155 links among 100 classes/objects, attributes, attribute values, and actions.*

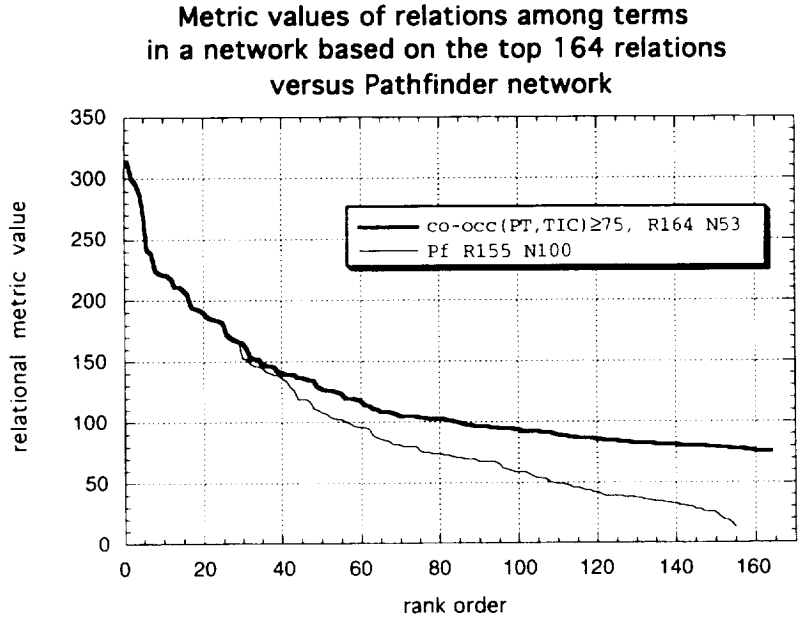



*Figure 9(b). Graph showing that the network in figure 9(a) does not use the relations with the highest metric values.*

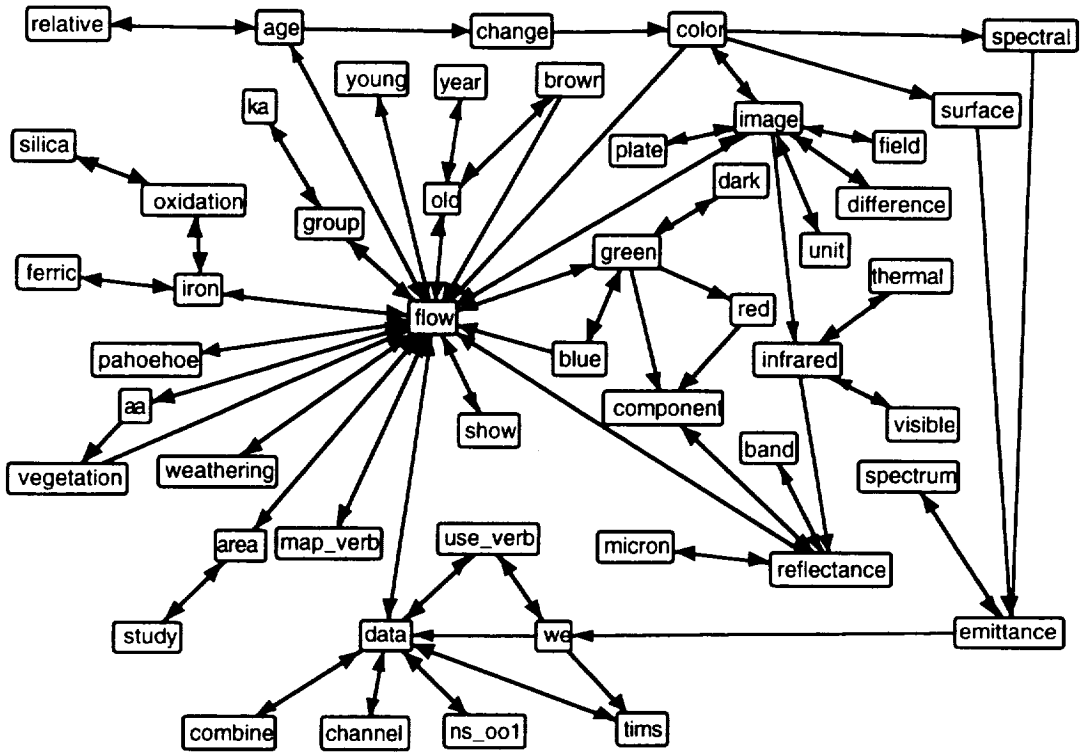



*Figure 10(a). Pathfinder network with 102 links among 50 probe terms.*

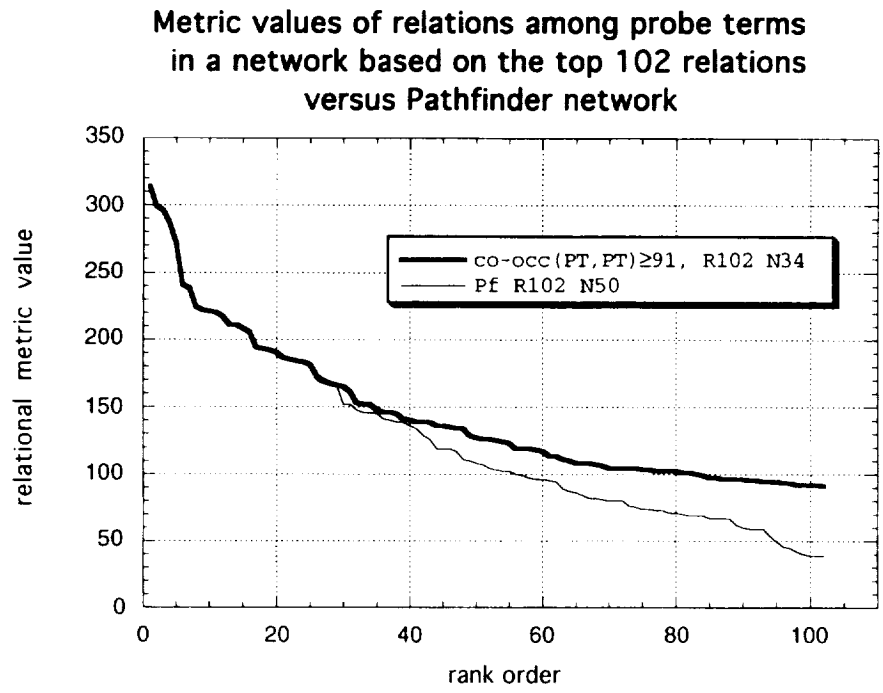



*Figure 10(b). Graph showing that the network in figure 10(a) does not use the relations with the highest metric values.*

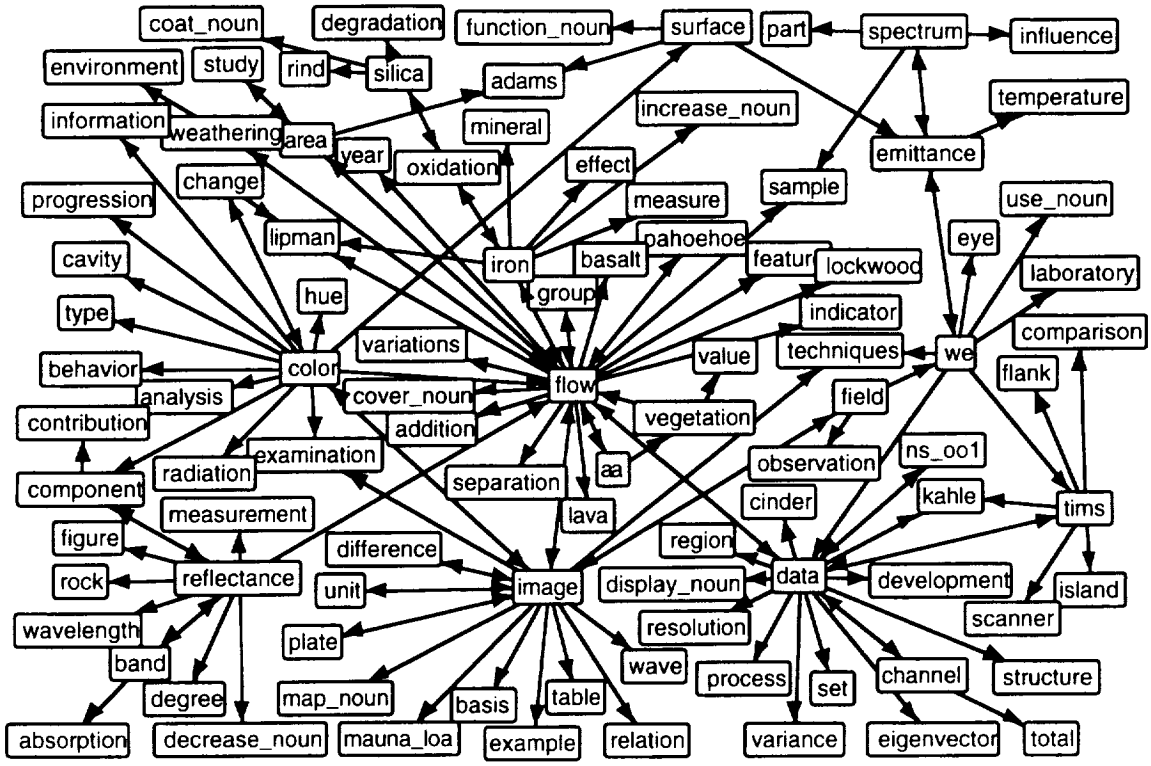



*Figure 11(a). Pathfinder network with 137 links among 100 classes/objects.*

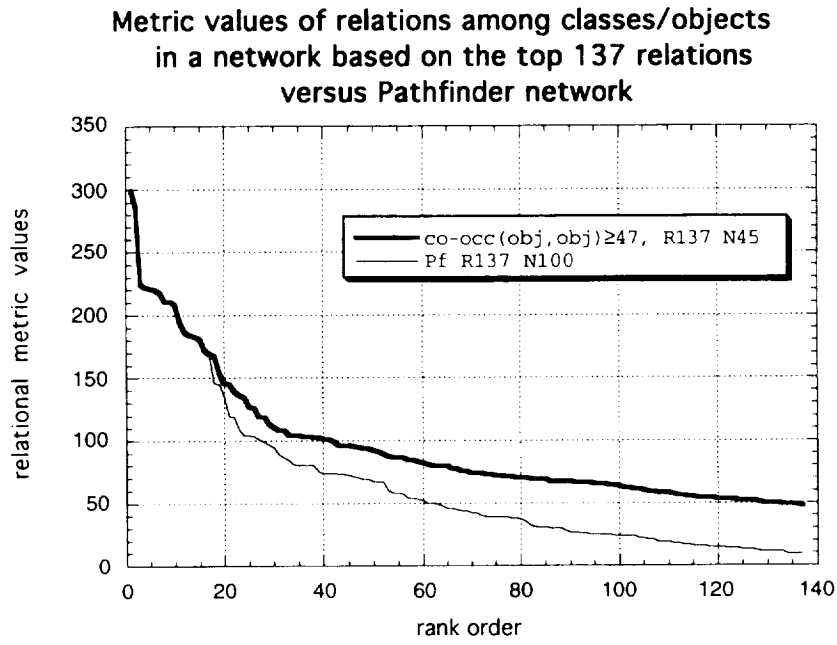
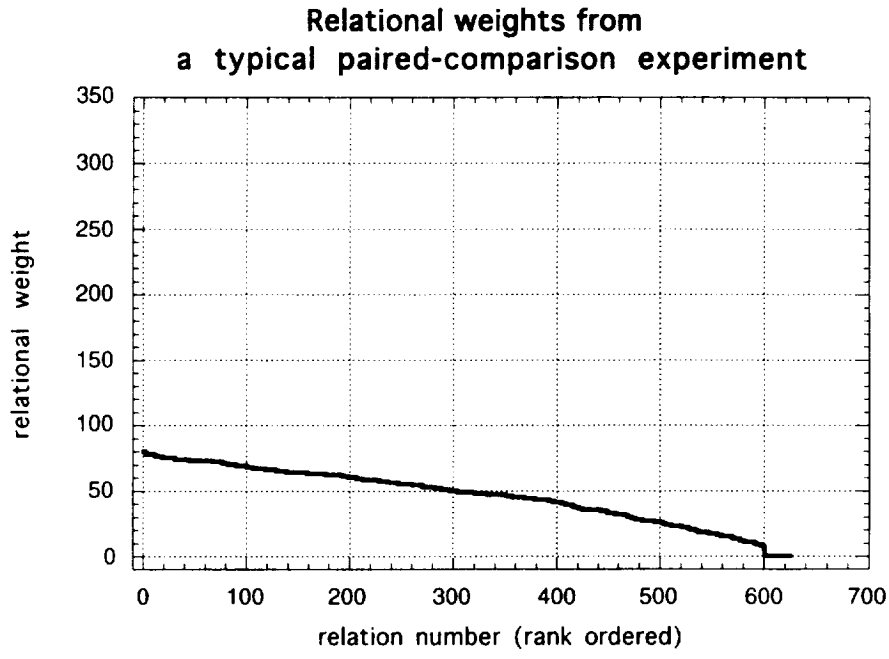
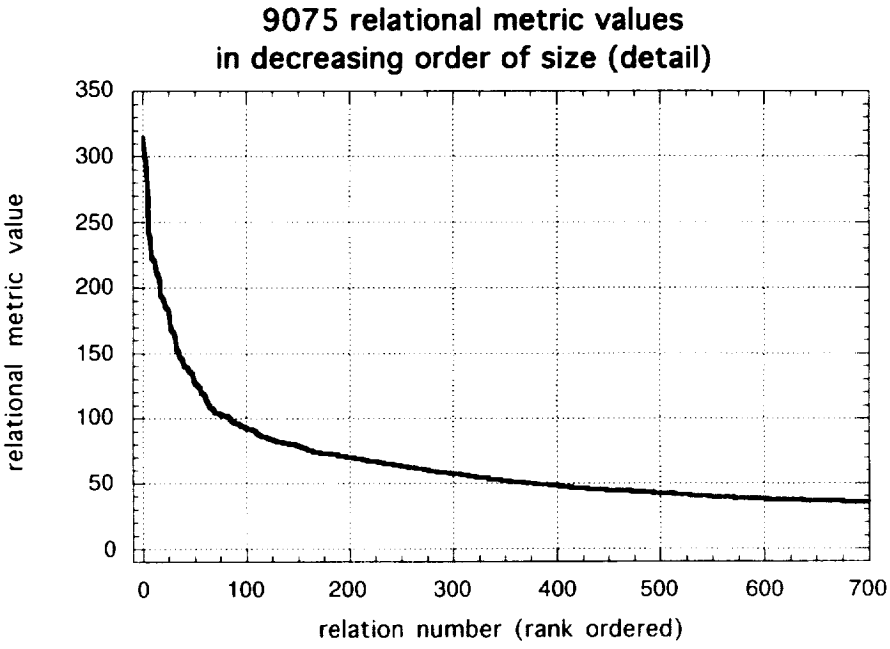



*Figure 11(b). Graph showing that the network in figure 11(a) does not use relations with the highest metric values.*




(a)




(b)

*Figure 12. Comparison of relational data from a typical paired comparison experiment (Schvaneveldt, Durso, and Dearholt, 1989) and relational metric values derived from verbal data based on proximity-weighted co-occurrence. (a) Relational weights from a typical paired-comparison experiment, (b) 9075 relational metric values in decreasing order of size (detail).*

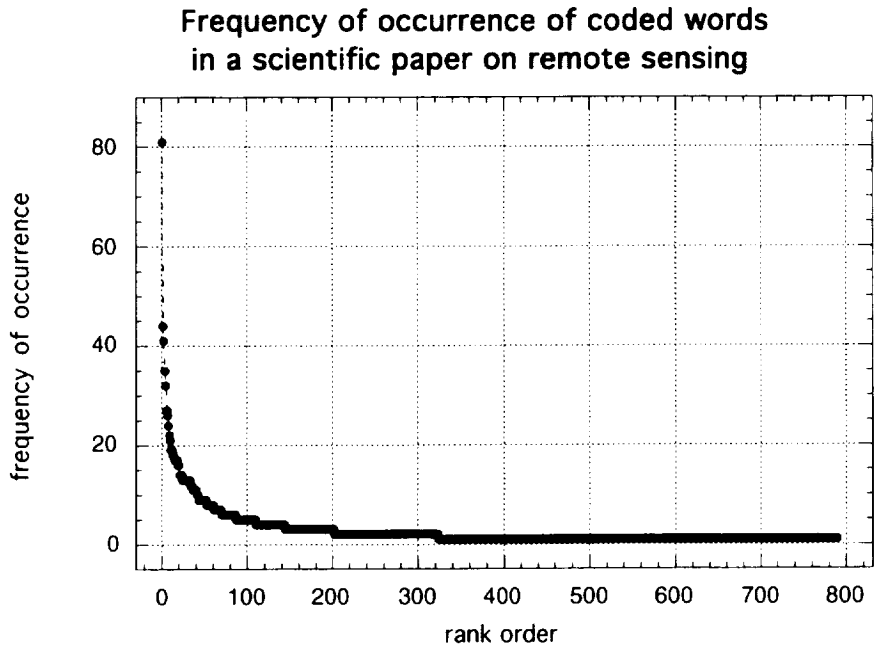



*Figure 13. Frequency of occurrence of coded words in the original source text, a scientific paper on remote sensing. The probe terms, listed in table 2, were selected from among those words with a frequency of occurrence greater than or equal to eight.*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | <- - - - - - - - - - context of PT1a - - - - - - - - -> | | | | | | | | | | |
| | | | | | | <- - - - - - - - - - context of PT1b - - - - - - - - -> | | | | | | | | |
| | | | | | age | <wd> | **flow** | <wd> | **flow** | color | | | | | |
| | w1 | w2 | w3 | w4 | pt2 | w6 | **PT1a** | w7 | **PT1b** | w9 | w10 | w11 | w12 | w13 | w14 |
| PT1a | 0 | 1 | 2 | 3 | **4** | 5 | - | 5 | - | 3 | 2 | 1 | 0 | 0 | 0 |
| PT1b | 0 | 0 | 0 | 1 | **2** | 3 | - | 5 | - | 5 | 4 | 3 | 2 | 1 | 0 |
| sum | 0 | 1 | 2 | 4 | **6** | 8 | - | 10 | - | 8 | 6 | 4 | 2 | 1 | 0 |

*Figure 14(a). Illustration of the symmetric method of computing the relational metric value between two words (e.g., "flow" and "age") when two instances of a probe term (e.g., "flow") are in close proximity. To find the relational metric value for R(flow,age), that is, "age" in the context of the probe term (PT) "flow," the metric values for each instance of the probe term "flow" relative to "age" are summed, in this case resulting in a value of 6. This equals the metric value of "flow" in the context of the probe term "age," as shown in figures 14(b) and 2(b), illustrating that this method is, indeed, symmetric. The context window here is six words, including the probe term, to the left and right of the probe term. Active probe term: **PT1a** and **PT1b**. Terms in context: w2-4, pt2, w6, w7, w9-13.*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <- - - - - - - - - - context of PT2 - - - - - - - - -> | | | | | | | | | | | | | | |
| | | | | | | **age** | <wd> | flow | <wd> | flow | color | | | | |
| | w0 | w1 | w2 | w3 | w4 | **PT2** | w6 | pt1a | w7 | pt1b | w9 | w10 | w11 | w12 | w13 |
| PT2 | 1 | 2 | 3 | 4 | 5 | - | 5 | **4** | 3 | **2** | 1 | 0 | 0 | 0 | 0 |

*Figure 14(b). Illustration of the symmetric method of computing the relational metric value between two words (e.g., "flow" and "age") when two instances of a term (e.g., "flow") are in close proximity to the probe term (e.g., "age"). To find the relational metric value for R(age,flow), that is, "flow" in the context of the probe term "age," the metric values for each instance of "flow" relative to the probe term "age" are summed, in this case resulting in a value of 6. Active probe term: **PT2**. Terms in context: w0-4, w6, pt1a, w7, pt1b, w9.*

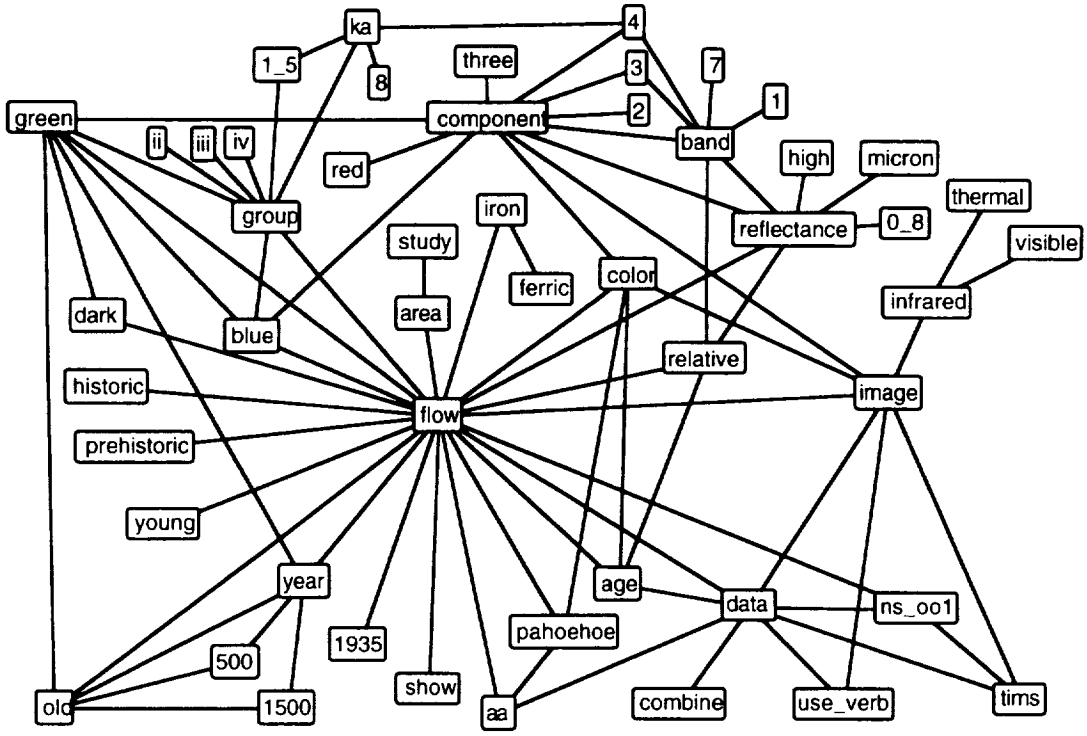
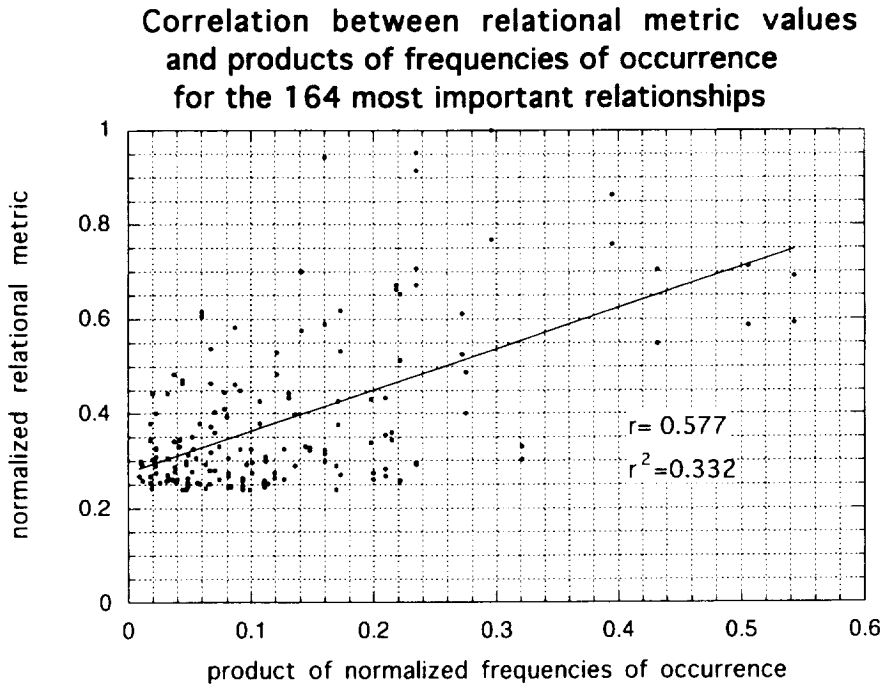
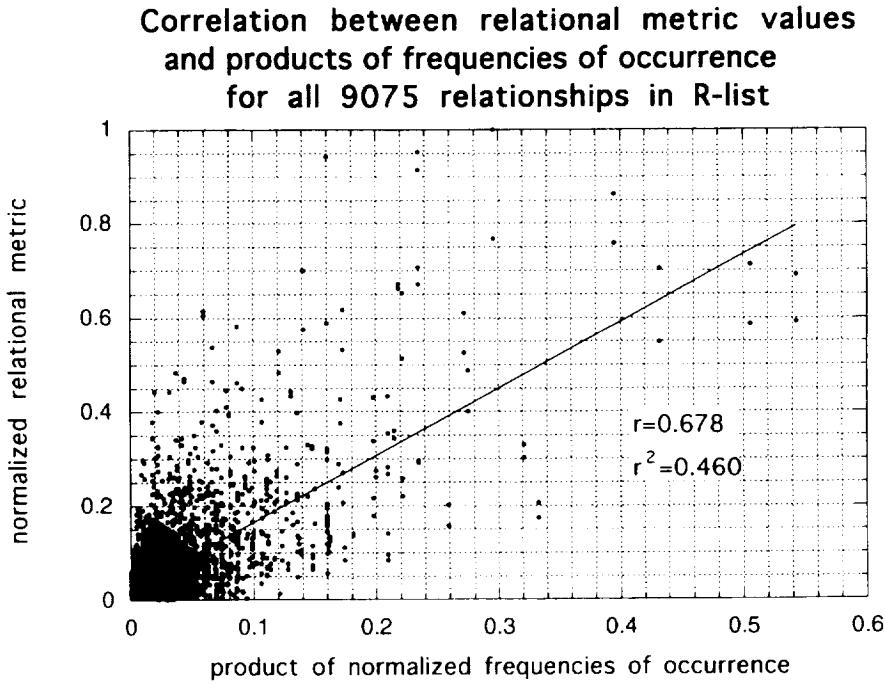



*Figure 14(c). Network based on symmetric method of calculating relational metric values. Note that arrowheads on relations are not needed since the relations have the same values in both directions. This network is very similar to figure 6, which is based on the asymmetric method.*

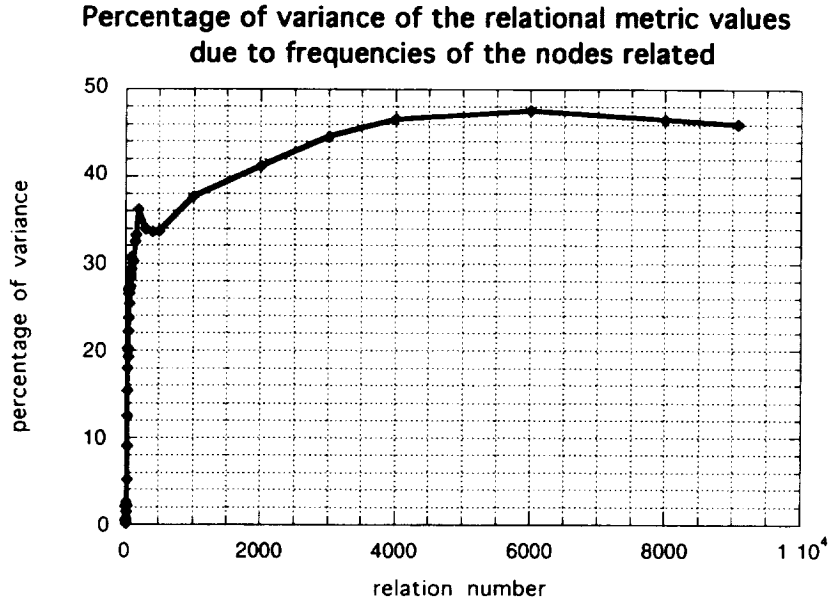



(a)

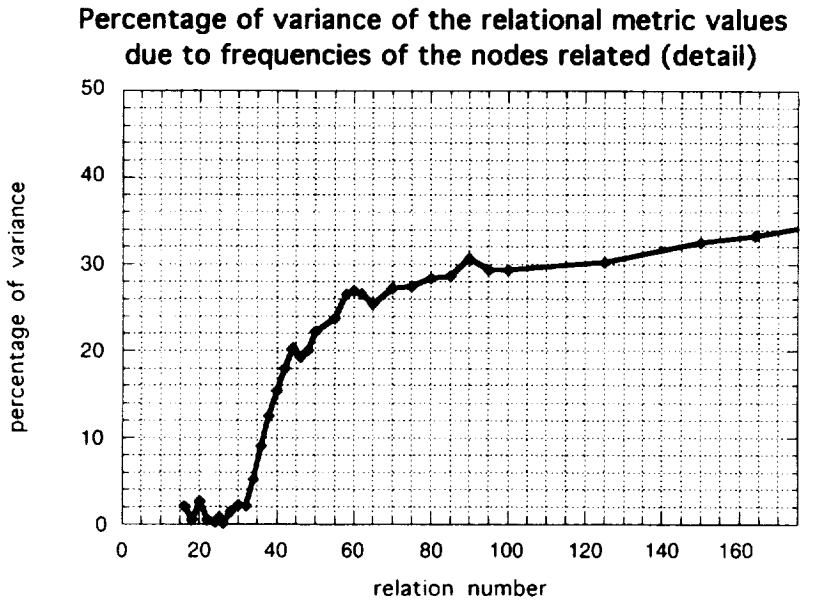



(b)

*Figure 15. Correlations between normalized relational metric values and products of normalized frequencies of occurrence for the top 164 relationships and for all 9075 relationships in R-list. (a) Correlation between relational metric values and products of frequencies of occurrence for the 164 most important relationships, (b) correlation between relational metric values and products of frequencies of occurrence for all 9075 relationships in R-list.*




(a)




(b)

*Figure 16. Influence of the frequencies of occurrence of related words on the relational metric values between them, as indicated by the percentage of variance of the relational metric values due to the products of the frequencies of occurrence of the nodes related. The influence of the frequencies of occurrence is lowest for the most prominent relations. Percentage of variance increases rapidly as the number of top relations in the domain model increases to about 60, then increases more gradually as the number of relations approaches 6000. "Relation number" is the rank order of each relation based on its relational metric value. The most prominent relations have the lowest relation numbers. Compare this figure with figure 4. (a) Percentage of variance of the relational metric values due to frequencies of the nodes related, (b) percentage of variance of the relational metric values due to frequencies of the nodes related (detail).*

# Appendix

Object-oriented relational analysis report, which describes the nodes and relationships in the object-oriented network shown in figures 6 and 8. The report consists of a list of nodes (entities), each with a sublist of the nodes in its context. Nodes are listed in order of estimated importance. N is an estimate of the node's relational importance. (See table 4.) R is the normalized proximity-weighted co-occurrence metric. The term "internals," as used in this appendix, refers to attributes, attribute values, or methods of a class/object.

FORMAT:

**node** (N=x.xxx): definition ([**node**] => included in figure 6 as term-in-context only, not as a probe term)

| related node | R | description of relationship |
|---|---|---|
| xxx | 0.xx | ... |

*xxx => appears in two places; intra- or inter-node weight determined by this interpretation*

*{xxx} => appears in two places; intra- and inter-node weight NOT determined by this interpretation*

**flow** (N=1.000): A lateral, surficial outpouring of molten lava from a vent or a fissure; also, the solidified body of rock that is so formed; synonymous with lava flow. (Bates and Jackson, 1987)

Relations between "flow" and its attributes and attribute values:

Age-related attributes and attribute values:

| | | |
|---|---|---|
| age | 0.76 | attribute of flow; The fundamental issue concerns measuring the ages of flows using remote sensing techniques. |
| old | 0.77 | attribute value of relative age, which is an attribute of flow; |
| young | 0.59 | attribute value of relative age, which is an attribute of flow |
| *relative* | *0.34* | The relative reflectance of flows in different spectral bands is systematically related to the relative age of flows. |
| year | 0.32 | age of a flow is measured in years |
| 1935 | 0.26 | historic lava flow from Mauna Loa ; The 1935 flow is adjacent to the 1843 and 1899 flows, and these are the three historic lava flows within the study area. |

Material-related attribute values

| | | |
|---|---|---|
| iron | 0.27 | one of the major constituents of basaltic lava flows; Iron is oxidized by weathering of the flow. |
| basalt | 0.24 | type of rock which comprises the studied lava flows |

Color-related attribute and attribute value

| | | |
|---|---|---|
| *{color}* | *{0.55}* | attribute of flow |
| *brown* | *0.25* | true surface color of certain flows |

Methods:

| | | |
|---|---|---|
| show | 0.35 | Flows, flow data, and/or flow data images show spectral features and other effects of chemical and physical processes, especially systematic effects of weathering with flow age. |

Relations between "flow" and other objects and their internals:

| | | |
|---|---|---|
| pahoehoe | 0.71 | kind of flow distinguished by its ropey texture |
| aa | 0.67 | kind of flow distinguished by its blocky texture |
| data | 0.59 | collection of remote sensing measurements related to flow |
| image | 0.59 | Flows are represented in images. |
| *color* | *0.55* | attribute of image; Image colors vary with flow age. |
| green | 0.53 | attribute value of image color that represents selected flow data |
| blue | 0.51 | attribute value of image color that represents selected flow data |
| dark | 0.29 | quality of colors representing flows |
| *{brown}* | *{0.25}* | color of certain flows in false color images |
| group | 0.53 | age-based collection of flows |
| area | 0.30 | region containing flows |
| reflectance | 0.30 | ratio of energy reflected by a flow to that incident upon it, which varies over the surface of a flow |
| *{relative}* | *{0.34}* | The relative reflectance of flows in different spectral bands is systematically related to the relative age of flows. |

**image** (N=0.384): two dimensional array of pixels, each assigned a color which represents a numeric data value;

Relations between "image" and its attributes and attribute values:

| | | |
|---|---|---|
| color | 0.66 | attribute of flow images used to judge relative age of flows |
| *difference* | 0.25 | Some components used to create images are the difference between two bands. Color differences in images indicates different flow ages. |

Methods:

| | | |
|---|---|---|
| use_verb | 0.31 | action applied by the researchers to image data, image processing methods, and systems for gathering image data |

Relations between "image" and other objects and their internals:

| | | |
|---|---|---|
| flow | 0.71 | Flows are represented in images. |
| age | 0.27 | Relative flow age can be determined from systematic color variations in multispectral images. |
| data | 0.49 | images are created from remote sensing data |
| infrared | 0.32 | some of the data used in images is from in the infrared part of the spectrum |
| TIMS | 0.44 | system for acquiring image data; kind of image |
| field | 0.31 | TIMS images are checked in the field. |
| component | 0.29 | Principal components are assigned to red, green, and blue to produce false color images. |
| aa | 0.26 | Aa flows of different ages are more readily differentiated in NS-001 images than in TIMS images. |
| (band) | | |
| *[difference]* | *0.25* | Some components used to create images are the difference between two bands. Color differences in images indicates different flow ages. |
| NS-001 | 0.25 | system for acquiring image data; kind of image |

**data** (N=0.374): a collection of measurements;

Methods:

| | | |
|---|---|---|
| use-verb | 0.48 | Data are used to assess relative ages of flows, and to map flows. Methods are used to process, especially to combine, data. Tools are used to gather data, measure effects, and analyze samples. |
| combine | 0.32 | TIMS thermal infrared data are combined with NS-001 visible/near infrared/short wavelength infrared data. |

Relations between "data" and other objects and their internals:

| | | |
|---|---|---|
| flow | 0.69 | TIMS and NS-001 scanned lava flows from aircraft to gather remote sensing data. |
| age | 0.36 | The data consist of spectral measurements that are assigned colors to indicate flow age. |
| TIMS | 0.58 | system for acquiring thermal infrared data; kind of data |
| NS-001 | 0.46 | system for acquiring visible/near infrared/short wavelength infrared data; kind of data |
| image | 0.40 | representation created from data |
| color | 0.29 | Data values are represented by image colors. |
| aa | 0.26 | TIMS and NS-001 data are used to determine the relative ages of different aa flows and to distinguish aa from pahoehoe. |

**age** (N=0.293): years since eruption of a lava flow;

Relations between "age," its object "flow," other internals of "flow," and objects which inherit from "flow":

| | | |
|---|---|---|
| flow | 0.86 | Age is an attribute of flow. |
| relative | 0.41 | Relative age can be indicated by image color differences. |
| aa | 0.25 | Image colors indicate the relative ages of aa flows. |
| pahoehoe | 0.24 | Image colors indicate the relative ages of pahoehoe flows. |

Relations between "age" and objects besides "flow" and their internals:

| | | |
|---|---|---|
| data | 0.34 | TIMS and NS-001 data is processed to determine flow age. |
| *show* | *0.25* | Data values and image colors show systematic variations with flow age. |
| image | 0.26 | False color images created from the data indicate different flow ages with different colors. |
| color | 0.43 | Age can be indicated by image color. |
| *{show}* | *0.25* | Data values and image colors show systematic variations with flow age. |
| systematic | 0.25 | There are systematic weathering effects which show up as a systematic progression of image colors with increasing flow age. |
| band | 0.25 | Reflectance in one band versus another varies with flow age. |

**green** (N=0.261): one of the primary colors in the RGB triad;

Relations between "green" and other internals of its object "image":

| | | |
|---|---|---|
| (image) | | |
| blue | 0.61 | In some NS-001 images, pahoehoe flows range in color from blue for younger ones to blue-green to green-blue to green for older flows, which does not provide for the best differentiation. |
| dark | 0.34 | Dark green is the color of pahoehoe flows which are >4000 years old in certain NS-001 and TIMS images. |
| red | 0.24 | Red and green are two of the three primary colors in RGB color triad. |

Relations between "green" and objects besides "image" and their internals:

| | | |
|---|---|---|
| flow | 0.61 | Flows in various age groups are green, dark green, blue-green, or green-blue in certain false color images. |
| old | 0.39 | In NS-001 images, with principal components assigned to primary image colors, aa flows 500-1500 years old are light blue-green; those 1500-4000 years old are green-yellow; and those >4000 years old are dark green. |
| component | 0.45 | A numbered component is associated with a primary color such as green. |
| 3 | 0.24 | Channel 3 and component 3 are associated with the color green. |
| group | 0.29 | Different shades of green are associated with different flow age groups. |

**old** (N=0.254): aged;

Relations between "old," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 1.00 | In this domain sample, the most strongly related entities are "old" and "flow". The most fundamental idea in the analyzed text is that of old flows. |
| year | 0.47 | Flows are X years old, where X is a variable. |
| 500 | 0.29 | Flows 200-500 and 500-1500 years old were differentiated in the images. |
| 1500 | 0.27 | Flows 500-1500 and 1500-4000 years old were differentiated in the images. |
| young | 0.25 | Young and old are the two major categories of relative age. |

Relations between "old" and objects besides "flow" and their internals:

| | | |
|---|---|---|
| (image) | | |
| green | 0.39 | Old pahoehoe flows are green in the processed NS-001 and TIMS images. |
| brown | 0.26 | Older aa flows appear uniformly brown in TIMS images, while younger aa is well differentiated in a range from reddish-browns to blue-browns in NS-001 images. |

**color** (N=0.243): a characteristic of images and flows which visually distinguishes one part from another;

Relations between "color" and its object "image":

| | | |
|---|---|---|
| image | 0.67 | Color is the most important attribute of images in this domain sample. |

Relations between "color" and objects besides "image" and their internals:

| | | |
|---|---|---|
| flow | 0.70 | Flows of various ages are represented in images by various assigned colors. Flows have natural color, which varies slightly with age. |
| age | 0.38 | Image color varies with flow age. |
| component | 0.33 | A color may be associated with a component. |
| data | 0.30 | Data values are represented in images by colors. |
| pahoehoe | 0.27 | Aa and pahoehoe flows of different ages are displayed in different colors. |

**component** (N=0.216): a combination of image data from different bands, produced by principal components analysis, that contains a certain percentage of the total statistical variance in the image; A component is assigned a number, and selected ones are assigned the color red, green, or blue.

Relations between "component" and its internals:

| | | |
|---|---|---|
| (component number) | | |
| 3 | 0.32 | number assigned to a component |
| 4 | 0.31 | number assigned to a component |

Relations between "component" and other objects and their internals:

| | | |
|---|---|---|
| reflectance | 0.43 | Reflectance data from several bands are processed to identify principal components. |
| band | 0.40 | Data from several bands are processed to identify principal components. |
| image | 0.24 | Principal components are derived from image data, and are used to create derived images. |
| green | 0.32 | Green is one of the RGB triad of colors assigned to principal components. |

**aa** (N=0.187): A Hawaiian term for lava flows typified by a rough, jagged, spinose, clinkery surface. (Bates and Jackson, 1987)

Relations between "aa," its superclass "flow," the internals of "flow," and other subclasses of "flow":

| | | |
|---|---|---|
| flow | 0.95 | generalization of aa |
| old | 0.28 | NS-001 data distinguishes old aa flows from one another. |
| age | 0.26 | The relative ages of old aa flows are best differentiated in NS-001 images. |
| pahoehoe | 0.31 | With age, pahoehoe undergoes physical and chemical changes that differ from those undergone by aa. |

Relations between "aa" and other objects:

| | | |
|---|---|---|
| data | 0.32 | Remote sensing data can be used to differentiate between young and old aa, and to distinguish aa from pahoehoe |
| image | 0.27 | The appearance of aa flows in multispectral images differs from that of pahoehoe flows. |

**reflectance** (N=0.174): The ratio of energy reflected by a body to that incident upon it. (Bates and Jackson, 1987) A radiometric quantity, varying over the surface of a flow, which can be measured in different spectral bands to produce spectral data.

Relations between "reflectance" and its attributes and attribute values:

| | | |
|---|---|---|
| micron | 0.28 | unit of measure applied to wavelengths of radiant energy, including reflectance |
| high | 0.27 | Older aa flows have higher reflectance toward the red and infrared part of the spectrum. Vegetation has higher reflectance in other bands. |

Relations between "reflectance" and other objects:

| | | |
|---|---|---|
| band | 0.54 | range of electromagnetic spectrum in which reflectance or other radiometric quantity is be measured |
| component | 0.38 | Reflectance data from several bands are processed to identify principal components. |
| flow | 0.33 | Reflectance varies over the surface of a flow. |

**blue** (N=0.173): one of the primary colors in the RGB triad;

Relations between "blue" and other internals of its object "image":

| | | |
|---|---|---|
| (image) | | |
| green | 0.61 | Blue and green represent flows over a range of ages. Variants on blue and green include blue-green and green-blue. |
| color | 0.29 | Blue is one of the primary colors in the RGB color triad. In NS-001 images, young pahoehoe flows were an indistinguishable blue colored units. |

Relations between "blue" and objects besides "image" and their internals:

| | | |
|---|---|---|
| flow | 0.64 | Blue and variants of blue are typical flow colors in the processed images. |
| old | 0.28 | In processed NS-001 images, aa flows of 200-500 years old are blue-brown, and those of 500-1500 years old are light blue-green. |
| component | 0.33 | Blue is assigned to one of the principal components. |
| group | 0.28 | Flow age groups were associated with blue and variants of blue, green and its variants, and reds including brown. |

**pahoehoe** (N=0.148): A Hawaiian term for a type of basaltic lava flow typified by a smooth, billowy, or ropy surface. (Bates and Jackson, 1987)

Relations between "pahoehoe," its superclass "flow," internals of "flow," and other subclasses of "flow":

| | | |
|---|---|---|
| flow | 0.91 | generalization of pahoehoe |
| age | 0.26 | Relative ages of young pahoehoe flows are more readily seen in TIMS images. |
| aa | 0.31 | With age, aa undergoes physical and chemical changes that differ from those undergone by pahoehoe. |

Relations between "pahoehoe" and the internals of other objects:

| | | |
|---|---|---|
| (image) | | |
| color | 0.30 | Aa and pahoehoe flows of different ages are displayed in different colors. |

**TIMS** (N=0.138): Thermal Infrared Multispectral Scanner, a system flown on board a NASA C-130B aircraft that collects thermal infrared data in six spectral channels between 8.2 and 11.7 microns.

Relations between "TIMS" and other objects:

| | | |
|---|---|---|
| data | 0.70 | data are acquired by TIMS |
| image | 0.43 | image data are acquired by TIMS |
| NS-001 | 0.33 | Like TIMS, NS-001 is a system for acquiring data. Their spectral bands/channels differ (see their definitions). |

**group** (N=0.133): age-based collection of flows;

Relations between "group" and other objects and their internals:

| | | |
|---|---|---|
| flow | 0.62 | Flows are grouped according to age in five age groups:<br>I: 200-500 years old (0.2-0.5 ka)<br>II: 500-1500 years old (0.5-1.5 ka)<br>III: 1500-4000 years old (1.5-4 ka)<br>IV: 4000-8000 years old (4-8 ka)<br>V: >8000 years old (>8 ka) |
| ka | 0.33 | thousands of years before the present; A range of ages defines a group (see preceding table). |
| (image) | | |
| green | 0.30 | Variants of green are associated with several flow age groups. |
| blue | 0.29 | Variants of blue are associated with several flow age groups. |

**band** (N=0.117): A frequency or wavelength interval . (Bates and Jackson, 1987) Spectral data acquisition systems such as TIMS and NS-001 have sensors which are sensitive to different bands.

Relations between "band" and its attributes and attribute values:

| | | |
|---|---|---|
| (band number) | | |
| 4 | 0.26 | number assigned to a band of TIMS and a different band of NS-001 |

Relations between "band" and other objects:

| | | |
|---|---|---|
| reflectance | 0.46 | radiometric quantity measured in several spectral bands |
| component | 0.36 | Data from several bands are processed to identify principal components. |

**NS-001** (N=0.115): A multispectral scanner developed by NASA as a Thematic Mapper Simulator. It has been flown on board a NASA C-130B aircraft to collect multispectral data in the visible, and short and long wavelength infrared regions corresponding to the seven Landsat-4 and -5 Thematic Mapper bands. In addition, NS-001 has an eighth band in the short wavelength infrared between 1.13 and 1.35 microns.

Relations between "NS-001" and other objects:

| | | |
|---|---|---|
| data | 0.58 | Multispectral remote sensing data is acquired by NS-001. |
| TIMS | 0.34 | like NS-001, a system for acquiring data; Their spectral bands/channels differ (see their definitions). |
| flow | 0.32 | NS-001 data were acquired from Hawaiian lava flows. |
| image | 0.27 | Old pahoehoe flows appear greener in NS-001 images, given the processing used in this particular study. |

**relative** (N=0.112): comparative; relating each to the other;

Relations between "relative," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.43 | Relative ages of flows are displayed in color images created using TIMS and NS-001 data. |
| age | 0.45 | Relative ages of flows are displayed in color images created using TIMS and NS-001 data. |

Relations between "relative age" and objects besides "flow":

| | | |
|---|---|---|
| band | 0.33 | Reflectance in one band relative to another indicates relative flow age. |
| reflectance | 0.29 | Reflectance in one band relative to another indicates relative flow age. |

Relations between "relative" and its other object "component":

| | | |
|---|---|---|
| component | 0.25 | Relative contributions from each of three components determines image color. |

**use (verb)** (N=0.107): bring into action or service;

Relations between "use (verb)" and the users:

| | | |
|---|---|---|
| we | 0.26 | We (the researchers) used remote sensing techniques, combining TIMS and NS-001 data, to measure weather effects in order to determine relative ages of lava flows. We also used scanning electron microscopy to analyze field samples in checking our results. |

Relations between "use (verb)" and the objects being used (that is, objects in which "use (verb)" is a method):

| | | |
|---|---|---|
| data | 0.53 | The researchers used data to assess and map the relative ages of flows. |
| image | 0.30 | Researchers used data displayed as images. |
| flow | 0.26 | The relative ages of flows can be determined using TIMS and NS-001 data. |
| TIMS | 0.25 | TIMS can be combined with NS-001 data using remote sensing techniques. TIMS data can be used to map young pahoehoe flows. |

**year** (N=0.096): unit of flow age;

Relations between "year," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.33 | Age of a flow is measured in years. |
| old | 0.46 | an attribute value of age; Year is a unit of age. |
| 500 | 0.29 | Flows in group I were 200-500 years old, and those in group II were 500-1500 years old. |
| 1500 | 0.27 | Flows in group II were 500-1500 years old, and those in group III were 1500-4000 years old. |

**infrared** (N=0.093): part of the emittance spectrum of the lava which constitutes the flow; includes reflected infrared and thermal infrared;

Relations between "infrared," its object "band," and the internals of "band":

| | | |
|---|---|---|
| (band) | (note: "infrared," "thermal," and "visible" imply different upper and lower bounds of bands within the spectrum) | |
| thermal | 0.31 | TIMS gathers data in thermal infrared bands. |
| visible | 0.30 | NS-001 gathers data in visible, near-infrared, and short-wave infrared bands.. |

Relations between "infrared" and objects besides "band":

| | | |
|---|---|---|
| reflectance | 0.24 | NS-001 measures infrared reflectance, while TIMS measures thermal infrared. |

**young** (N=0.092): relatively recently erupted; not aged;

Relations between "young," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.94 | Young flows are relatively unweathered. Young aa flows are differentiated by age using both NS-001 and TIMS data, but young pahoehoe flows are only readily differentiated using TIMS data. |
| old | 0.26 | Old flows differ from young flows because of physical and chemical changes due to weathering. |

**dark** (N=0.080): absence of light; not light in color;

Relations between "dark" and other internals of its object "image":

| | | |
|---|---|---|
| green | 0.48 | Dark green is the color of the oldest pahoehoe flows in both NS-001 and stretched TIMS images. The youngest aa flows are dark blue-green in stretched TIMS images. |

Relations between "dark" and objects besides "image" and their internals:

| | | |
|---|---|---|
| flow | 0.40 | (see "green" above) Recent flows are relatively unweathered and are dark to the eye. |
| old | 0.26 | Dark green is the color of the oldest pahoehoe flows in both NS-001 and stretched TIMS images. |

**show** (N=0.069): exhibit, demonstrate, make evident;

Relations between "show," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.43 | Flows, flow data, and flow images show systematic effects with flow age. |
| age | 0.25 | Various measurements show systematic changes with flow age. |

Relations between "show" and its other object "data":

| | | |
|---|---|---|
| data | 0.25 | Flow data show systematic effects with flow age. Reflectance data shows the degree of oxidation. |

**area** (N=0.060): region of volcanic terrain;

Relations between "area" and other objects:

| | | |
|---|---|---|
| study | 0.38 | A study is conducted in a particular area. |
| flow | 0.32 | An area of interest can contain many lava flows. |

**iron** (N=0.057): metallic element which is one of two major components of basaltic lava flows (the other is magnesium)

Relations between "iron," its object "flow, and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.28 | Weathered flows have higher ferric iron content. |
| ferric | 0.31 | Ferric iron is a weathering product that forms on the surface of flows. The ferric iron content in a field sample of a flow can be measured using wet chemical analyses. Ferric iron content increases systematically with flow age. |

**ka** (N=0.054): unit of measure of time; thousands of years before the present;

Relations between "ka" and internals of its object "flow":

| (flow / age) | | |
|---|---|---|
| 1.5 | 0.30 | Groups II and III have 1.5 ka as lower and upper boundaries, respectively. |
| 8 | 0.26 | Groups IV and V have 8 ka as lower and upper boundaries, respectively. |

Relations between "ka" and objects besides "flow":

| | | |
|---|---|---|
| group | 0.32 | The age boundaries of a flow group are measured in ka, that is, thousands of years before the present. |

**red** (N=0.051): one of the primary colors in the RGB triad;

Relations between "red" and other objects and their internals:

| | | |
|---|---|---|
| component | 0.38 | Red, one of the three primary colors in the RGB triad, is associated with one of the principal components. |
| flow | 0.27 | Flows appear more red to the eye as they age, due to the oxidation of iron. In combined TIMS and NS-001 images, older flows become more rusty or redder with age. |
| (image) | | |
| brown | 0.24 | In TIMS images, the youngest prehistoric aa flows (0.2-1.5 ka) are reddish-brown. |

**ferric** (N=0.043): designating or of iron with a valence of three;

Relations between "ferric," its object "flow," and other internals of "flow":

| | | |
|---|---|---|
| flow | 0.25 | Ferric iron is a weathering product of basaltic lava flows. |
| iron | 0.40 | (same as above) |

**micron** (N=0.040): a unit of length; one thousandth of a millimeter or millionth of a meter;

Relations between "micron," its object "reflectance," and other internals of "reflectance":

| | | |
|---|---|---|
| reflectance | 0.35 | The wavelength of reflectance is measured in microns. |
| 0.8 | 0.26 | Increasingly older flows develop higher reflectance at 0.8 microns. |

**field** (N=0.037): the site, or pertaining to the site, of the lava flows;

Relations between "field" and other objects:

| | | |
|---|---|---|
| image | 0.27 | Remote sensing images were checked against field images and observations. |
| we | 0.25 | We (the researchers) studied the study area in the field to check findings from remote sensing data. |

**difference** (N=0.034): element of dissimilarity;

Relations between "difference," its object "image," and other internals of "image":

| | | |
|---|---|---|
| image | 0.25 | Differences in image colors correspond to differences in flow ages. |
| color | 0.25 | (same as above) |

**[brown]** (N=0.034): a color which is a combination of red with some green and/or blue;

Relations between "brown" and internals of its object "image":

| | | |
|---|---|---|
| (image) | | |
| red | 0.24 | Older aa flows appear uniformly brown in TIMS images, while younger aa is well differentiated in a range from reddish-browns to blue-browns in NS-001 images. |

Relations between "brown" and other objects and their internals:

| | | |
|---|---|---|
| flow | 0.25 | Older aa flows appear uniformly brown in TIMS images, while younger aa is well differentiated in a range from reddish-browns to blue-browns in NS-001 images. |
| old | 0.26 | (same as above) |

**thermal** (N=0.034): having to do with heat;

Relations between "thermal" and internals of its object "band":

| (band) | | (note: "infrared," "thermal," and "visible" imply different upper and lower bounds of bands within the spectrum) |
|---|---|---|
| infrared | 0.44 | TIMS gathers data in thermal infrared bands. |

**visible** (N=0.033): part of the emittance spectrum of the lava which constitutes the flow;

Relations between "visible" and internals of its object "band":

| (band) | | (note: "infrared," "thermal," and "visible" imply different upper and lower bounds of bands within the spectrum) |
|---|---|---|
| infrared | 0.44 | NS-001 gathers data in visible, near-infrared, and short-wave infrared bands. |

**study** (N=0.032): a systematic investigation;

Relation between "study" and other objects:

| area | 0.34 | A study is conducted in a particular area. |
|---|---|---|

**combine** (N=0.031): integration using Karhunen-Loeve transformations, that is, principal components analysis (PCA);

Relations between "combine" and its object "data":

| data | 0.37 | The researchers combined data from TIMS and NS-001 because TIMS cannot differentiate among old aa flows, but NS-001 can, and NS-001 cannot differentiate young pahoehoe flows, but TIMS can. (They each can differentiate among young aa flows and among old pahoehoe flows.) |
|---|---|---|

**[we]** (N=0.023): the researchers themselves;

Relations between "we" and other objects and their internals:

| (data; methods; equipment) | | |
|---|---|---|
| use_verb | 0.26 | We (the researchers) used remote sensing techniques, combining TIMS and NS-001 data, to measure weather effects in order to determine relative ages of lava flows. We also used scanning electron microscopy to analyze field samples in checking our results. |
| field | 0.25 | We (the researchers) studied the study area in the field to check findings from remote sensing data. |

**plate** (N=0.012): full page illustration;

Relations between "plate" and its internals:

| (number) | | |
|---|---|---|
| 1 | 0.27 | Plate 1 is a geologic map of the Mauna Loa test site. |

**[high]** (N=0.012): great in intensity;

Relations between "high" and its object "reflectance":

| reflectance | 0.27 | Older aa flows have higher reflectance toward the red and infrared part of the spectrum. Vegetation has higher reflectance in other bands. |
|---|---|---|

**weathering** (N=0.012): physical and chemical effects of weather on rock surfaces;

Relations between "weathering" and its object "flow":

| flow | 0.26 | There are systematic weathering effects which show up as a systematic progression of image colors with increasing flow age. |
|---|---|---|

**[spectrum]** (N=0.011): electromagnetic spectrum, the entire range of wavelengths or frequencies of electromagnetic radiation;

Relations between "spectrum" and its subclasses:

| | | |
|---|---|---|
| emittance | 0.25 | Data from the reflectance and emittance part of the spectrum were combined. |

**[systematic]** (N=0.011): orderly;

Relations between "systematic" and internals of its object "flow":

| | | |
|---|---|---|
| (flow) | | |
| age | 0.25 | Flows, flow data, and/or flow data images show spectral features and other effects of physical processes, especially systematic effects with flow age. |

**emittance** (N=0.011): ratio of emitted radiant flux per unit area of a substance to that of a blackbody radiator of the same temperature; The radiance of a surface is a function of both its temperature and spectral emittance. Emittance is related to the composition of a surface.

Relations between "emittance" and its superclass, "spectrum":

| | | |
|---|---|---|
| spectrum | 0.25 | Data from the reflectance and emittance part of the spectrum were combined. |

**[basalt]** (N=0.011): a volcanic rock composed largely of iron, magnesium, and calcium;

Relations between "basalt" and its object, "flow":

| | | |
|---|---|---|
| flow | 0.24 | The studied lava flows consist of basalt. |

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. AGENCY USE ONLY *(Leave blank)* | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | July 1995 | Technical Memorandum |

**4. TITLE AND SUBTITLE**

A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain

**5. FUNDING NUMBERS**

199-06-30

**6. AUTHOR(S)**

Michael W. McGreevy

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Ames Research Center
Moffett Field, CA 94035-1000

**8. PERFORMING ORGANIZATION REPORT NUMBER**

A-950075

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

National Aeronautics and Space Administration
Washington, DC 20546-0001

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

NASA TM-110358

**11. SUPPLEMENTARY NOTES**

Point of Contact: Michael W. McGreevy, Ames Research Center, MS 262-2, Moffett Field, CA 94035-100; (415) 604-5784

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**

Unclassified-Unlimited
Subject Category – 66

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

An objective and quantitative method has been developed for deriving models of complex and specialized spheres of activity (domains) from domain-generated verbal data. The method was developed for analysis of interview transcripts, incident reports, and other text documents whose original source is people who are knowledgeable about, and participate in, the domain in question. To test the method, it is applied here to a report describing a remote sensing project within the scope of the Earth Observing System (EOS). The method has the potential to improve the designs of domain-related computer systems and software by quickly providing developers with explicit and objective models of the domain in a form which is useful for design. Results of the analysis include a network model of the domain, and an object-oriented relational analysis report which describes the nodes and relationships in the network model. Other products include a database of relationships in the domain, and an interactive concordance. The analysis method utilizes a newly developed relational metric, a proximity-weighted frequency of co-occurrence. The metric is applied to relations between the most frequently occurring terms (words or multiword entities) in the domain text, and the terms found within the contexts of these terms. Contextual scope is selectable. Because of the discriminating power of the metric, data reduction from the association matrix to the network is simple. In addition to their value for design, the models produced by the method are also useful for understanding the domains themselves. They can, for example, be interpreted as models of presence in the domain.

| 14. SUBJECT TERMS | 15. NUMBER OF PAGES |
|---|---|
| Object-oriented domain analysis, Knowledge acquisition from text, Virtual reality and presence | 61 |
| | 16. PRICE CODE A04 |

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | | |