Development and Application of a Metric on Semantic Nets

ROY RADA, HAFEDH MILI, ELLEN BICKNELL, AND MARIA BLETTNER

Abstract—**Motivated by the properties of spreading activation and conceptual distance, the authors propose a metric, called Distance, on the power set of nodes in a semantic net. Distance is the average minimum path length over all pairwise combinations of nodes between two subsets of nodes. Distance can be successfully used to assess the conceptual distance between sets of concepts when used on a semantic net of hierarchical relations. When other kinds of relationships, like "cause," are used, Distance must be amended but then can again be effective. The judgments of Distance significantly correlate with the distance judgments that people make and help us determine whether semantic net S_1 is better or worse than semantic net S_2. First a "conceptual distance" task is set, and people are asked to perform it. Then the same task is performed by Distance on S_1 and S_2. If Distance on S_1 performs more like people than Distance on S_2, the conclusion is that S_1 is better than S_2. Distance embedded in the methodology facilitates repeatable quantitative experiments.**

I. Introduction

OUR RESEARCH group focuses on developing better hierarchical knowledge bases [1] and, in the course of this work, requires a method for assessing the value of a knowledge base. Our application area is the retrieval of biomedical literature, and a natural problem which the knowledge base should help solve is the ranking of documents listed in response to a query. Our method for ranking documents assumes that both are represented as sets of nodes in a hierarchical knowledge base. The method has reliably helped us judge the success of our merging algorithms and might similarly help other people. The method uses a metric called Distance that is easy to manipulate mathematically and to interpret.

Some examples of the psychological and information science significance of Distance are given elsewhere [2], [3]. This paper focuses on the mathematical characteristics of Distance and presents new cases and interpretations. Experiments in which Distance is applied to pairs of concepts and to sets of concepts in a hierarchical knowledge base show the power of hierarchical relations in representing information about the conceptual distance between concepts.

In this paper a knowledge base will be viewed as a graph. In many problems dealing with discrete objects and binary relations, a graphical representation of the objects and the binary relations on them is a very convenient form of representation [4], [5] that often leads to a solution using algorithms from graph theory. Depending on the nature of the problem, the edges could represent physical links (as for communication networks), time duration (as for task planning), or abstract relationships (as for association [6] structures).

Many graph problems, such as the minimum cost spanning tree and traveling salesman problems, require minimizing the sum of weights of some edges given a set of constraints [7]. Distances between entire graphs have been investigated in the pattern recognition literature, where a graph represents a picture [8], [9]. Such distances are based on the number of transformations that would make one graph similar to another one. However, to the best of our knowledge, a distance as defined in this paper has not been explored in the graph theory literature or elsewhere.

Human information processing often involves comparing concepts. There are various ways of assessing the similarity of concepts depending on the representation adopted for knowledge. In featural representations, concepts are represented by sets of features. The similarity between two concepts a and b with feature sets A and B can be expressed as a weighted sum of functions of different set operations on A and B [10]. The theory of spreading activation applies to comparing concepts in a semantic net [11]. Another method consists of constructing network fragments for sought for objects and then matching these fragments against the network data base [12]. In yet another strategy the in-degree and out-degree of two nodes in a semantic net are compared in the course of deciding how similar the two nodes are [13]. Faceted thesauri are a kind of semantic net and are being used in indexing and retrieving software. To decide whether two nodes in such a faceted thesaurus are similar, software scientists have devised measures based on the number and type of overlapping facets that the two nodes have [14].

Manuscript received May 10, 1987; revised February 24, 1988 and August 29, 1988. This work was supported in part by the National Science Foundation under Grant ECS-84-06683.

R. Rada is with the Department of Computer Science, University of Liverpool, Liverpool L69 3BX, England.

H. Mili is with the Department of Electrical Engineering and Computer Science, George Washington University, Washington, DC 20052.

E. Bicknell is with the Mesh Section, National Laboratory of Medicine, Bethesda, MD 20894.

M. Blettner is with the Department of Statistics and Computational Mathematics, University of Liverpool, Liverpool L69 3BX, England.

IEEE Log Number 8824462.

0018-9472/89/0100-0017$01.00

Distance is based on a simplified version of spreading activation. One of the assumptions of the theory of spreading activation is that the semantic network is organized along the lines of semantic similarity. The more properties two concepts share in common, the more links there are between the concepts and the more closely related they are. In these terms, semantic relatedness is based on an aggregate of the interconnections between the concepts. This is different from semantic distance which is equal to the minimal path length between two concepts. Links may be assigned criteriality tags to indicate the importance (strength) of the link between the connected nodes [11]. Links between a concept and its *defining* features (versus *characteristic* features [15]) expectably have higher "criterialities." Because "is-a" relations [16] are based on similarity between defining features, we hypothesize that when only is-a relations are used in semantic nets, semantic relatedness and semantic distance are equivalent (we could use the latter as a measure of the former).

Distance is principally designed to work with hierarchical knowledge bases. Hierarchies, both of abstractions and of concrete entities, are commonplace in the world and important in intelligent behavior [17]. They can be useful in controlling search [18] and in learning about the world [19]. The types of hierarchies most explored in this work are the type embedded in thesauri as used in the information retrieval field. These thesauri have long histories of being maintained for the indexing and retrieving of documents, and since they are now typically parts of computer information systems, they lend themselves to experiments in which the computer uses the thesaurus to help searchers [20], [21].

In the next section, we will present the mathematical properties of Distance. After that we will describe 1) what the mathematical properties mean in terms of knowledge engineering, and 2) several experiments using Distance on a semantic net with hierarchical relations. The major points will be that

- Distance is one tool to use in comparing one semantic net against another,
- Distance is a metric on sets of nodes of a graph, and
- applications of Distance to nonhierarchical semantic nets reveal the need for amendments to Distance.

II. Methodology

In this section, we discuss the methodology followed in the design of Distance. The design of Distance was guided by two observations:

1) the behavior of conceptual distance resembles that of a metric, and
2) the conceptual distance between two nodes is often proportional to the number of edges separating the two nodes in the hierarchy.

In this section, we first discuss the extent to which conceptual distance satisfies the properties of a metric. A function $f(x, y)$ is a metric if the following properties are satisfied:

1) $f(x, x) = 0$, zero property,
2) $f(x, y) = f(y, x)$, symmetric property,
3) $f(x, y) \geq 0$, positive property, and
4) $f(x, y) + f(y, z) \geq f(x, z)$, triangular inequality.

Second, we explore the extent to which the shortest path length between two nodes can be used as a measure of conceptual distance, using spreading activation as a model. In particular, we show that under some conditions, the shortest path length between two nodes indicates the conceptual distance between the nodes—for two nodes Distance is simply the shortest path length between them. Finally, we extend the definition of Distance to handle concepts represented by sets of nodes, rather than single nodes. This extension is significant in the context of information retrieval systems where a document or a query is represented by more than one concept from a semantic net.

A. Conceptual Distance is a Metric

Much human information processing involves concept matching. Category membership and similarity are two important aspects of concept matching. The more similar two concepts are, the smaller the conceptual distance between them. Conceptual distance is a decreasing function of similarity. When concepts are represented by points in a multidimensional space, conceptual distance can conveniently be measured by the geometric distance between the points representing the concepts at hand and, as such, satisfies the properties of a metric.

Some information retrieval systems use vector descriptions for documents and queries [22]. Each dimension corresponds to an elementary concept known to the system. The coordinate of a vector along a dimension attests to the relative importance of the corresponding elementary concept for the document (or query) at hand. In such systems, conceptual distance is measured by the geometric distance between the corresponding concepts [23]. In such systems, a query retrieves the documents whose vectors are "closest" to the vector representing the query. Such a retrieval strategy is prohibitively costly for large document collections. Accordingly, researchers have explored the extent to which the search for close documents can be reduced by organizing documents into a hierarchy of classes, where the elements of each class are within a prescribed conceptual distance from each other [22]. An incoming query is compared to "exemplar" documents, one from each top-level class. If the conceptual distance is higher than a prescribed value (depending on the user's input and the "radius" of the class), it is concluded that no document from that class is close enough to the query, and the whole class is disregarded. For this reasoning to hold, it is essential that conceptual distance satisfy the properties of a metric, especially the triangle inequality. In their work on memory-based reasoning, Stanfill and Waltz have

advocated a similar strategy and stressed the importance of metric properties [24].

Some cognitive scientists have challenged the applicability of metric properties to conceptual similarity measures. On the symmetry property, Tversky noted that there are instances where similarity appears to be asymmetric [10]. We believe that the asymmetry in those instances does not derive from the asymmetry of similarity (as a feature comparison process) but from the existence of another asymmetric relationship between two concepts. An instance of such a relationship is the instance–class relationship. For example, if people say that robins are more similar to birds than birds are similar to robins, we suspect that the people are making a fuzzy category-membership decision, rather than a similarity assessment. In a related argument, Ortony noted an extreme asymmetry in metaphorical statements [25]. Most students may consider the statement, "Lectures are like sleeping pills," to be more true than the statement, "Sleeping pills are like lectures." The asymmetry is due to the different roles played by the concepts in a metaphor. We build metaphors such as "A is like B," if the characteristic features of A match the defining features of B [15]. Therefore, although "is like" may sound like "is similar to," it is implicit in metaphors that the feature comparison process is selective. The authors believe that if similarity is limited to an (unconstrained) feature comparison process, it is symmetric.

Tversky [10] argued that the triangle inequality translates to what we will refer to as the "reverse triangle inequality." Given three concepts A, B, and C, the reverse triangle inequality says that the similarity of A to C is greater than the sum of the similarity of A to B and the similarity of B to C. Tversky showed that the reverse triangle inequality can be violated. We have two objections to Tversky's argument. First, it is not always the case that the triangle inequality for conceptual distance translates into the reverse triangle inequality for similarity. If the similarity between two concepts x and y is given by $S(x, y) = [1 + D(x, y)]^{-1}$, where $D(x, y)$ is a metric of conceptual distance between x and y,[1] then $S(x, y)$ does not satisfy the reverse triangle inequality.[2] Second, one of the examples used by Tversky to illustrate the violation of the reverse triangle inequality also illustrates an inconsistent use of similarity:

> ...although Jamaica is very similar to Cuba (due to its geographical characteristics) and Cuba is very similar to Russia (politically), Jamaica is not at all similar to Russia.... [15].

The problem with this example is that the similarity between Jamaica and Cuba is based on the geographical characteristics and ignores the political differences, and conversely, for the Cuba–Russia similarity. If we account for both defining properties, Jamaica will not be very similar to Cuba, nor will Cuba be very similar to Russia.

In summary, treating conceptual distance as a metric is consistent with the practical view of concepts as points in a multidimensional space. Exceptions to the symmetry and triangle inequality seem to result from a broad, if not inconsistent, use of similarity. Our work shows the viability of treating conceptual distance as a metric.

B. Shortest Path Lengths in is-a Hierarchies

In this section, we discuss the extent to which shortest path lengths in is-a hierarchies can be used to measure conceptual distance. In particular, we show that, in the context of Quillian's model of semantic memory [26], shortest path lengths are not sufficient to measure the conceptual distance between concepts. However, when the paths are restricted to is-a links, the shortest path length does measure conceptual distance. We also discuss how well the metric properties are supported by spreading activation [11].

In Quillian's model of semantic memory, concepts are represented by nodes and relationships by links. Links are labeled by the name of the relationship and are assigned "criteriality tags" that attest to the importance of the link. In computer implementations, criteriality tags are numerical values that represent the degree of association of the two concepts (such as how often that link is traversed) and the nature of the association. The association is positive if the existence of that link indicates some sort of similarity between the end nodes and negative otherwise. For example, *superordinate* links (the term used for is-a) have a positive association, while "is-not-a" links have a negative association.

Roughly speaking, spreading activation [11] prescribes that to compare two concepts, the paths that separate the two nodes and that satisfy the constraints defined by the semantics of the relations and the context are considered for evaluation. These paths are traced by propagating two "activation tags" from the nodes corresponding to the concepts, one tag originating from each node. When two activation tags "meet" at one node, the paths from the originating nodes to that node are concatenated to form a path between the originating nodes. For each path, positive criteriality tags contribute to "positive evidence" (for similarity), and negative criteriality tags contribute to "negative evidence." When positive evidence exceeds some predetermined threshold, the comparison is concluded successfully. On the contrary, if negative evidence falls below some negative threshold, it is concluded that the concepts are not similar.

In Quillian's model superordinate (is-a) links are assigned high criteriality tags. If spreading activation is used across is-a links only, short paths will significantly con-

[1] Prior to our work on Distance, we developed a similarity (Relevance) measure based on the same formula, where $D(x, y)$ was the shortest path length between x and y in an is-a hierarchy. Relevance simulated well people's assessments of similarity, but its results were not as easily interpretable as Distance's.

[2] Let a, b, and c be three concepts such that $D(a, b) = D(b, c) = D(a, c) = d$. While $D(a, c)$ is less than $D(a, b) + D(b, c)$—i.e., the triangle inequality is true—it is not true that $S(a, c)$ $(= [1 + d]^{-1})$ is bigger than $S(a, b) + S(b, c)$ $(= 2 \times [1 + d]^{-1})$—i.e., the reverse triangle inequality is not true.

tribute to positive evidence of similarity, and the correspondence between semantic distance (shortest path length) and semantic relatedness (conceptual distance) will be strong. We hypothesize that such correspondence is strong enough for the length of is-a paths to be used as a measure of semantic relatedness.

It is not necessary that a link and its inverse have the same criteriality tag. Consider the is-a relation between "robin" and "bird." Such a relation may be more important for the concept robin than it is for bird. This asymmetry is considered as a fundamental property of semantic constructs in KL-ONE [27]. However, for the purposes of spreading activation, it does not really matter whether a path is made of is-a links or inverse is-a links, or a mix of both; at the end, its contribution to positive evidence will be the same.

Based on the above observations, we define the conceptual distance between two concepts represented by nodes in an is-a semantic net as follows.

Definition 1: Let A and B be two concepts represented by the nodes a and b, respectively, in an is-a semantic net. A measure of the conceptual distance between A and B is given by

$$\text{Distance}(A, B) = \text{minimum number of edges separating } a \text{ and } b.$$

Henceforth, we will use interchangeably in the expression of Distance concepts or nodes representing those concepts. Clearly, Distance satisfies 1) the zero property, 2) the positive property, and 3) the symmetry property. The triangular inequality is between on the fact that by concatenating a shortest path between A and B to a shortest path between B and C, we get a path between A and C whose length is bigger or equal to the minimum path length between A and C. Thus we have the following result.

Theorem 1: Distance is a metric.

We later show that Distance, as defined above, well simulates people's assessments of conceptual distance.

C. Distance Between Sets of Nodes

Traditionally, documents and queries are represented by a combination of concepts from a predetermined set of concepts, called the indexing vocabulary. When the indexing vocabulary is a hierarchical semantic net, documents, and queries are represented by sets of nodes from the hierarchy. To extend Distance to handle sets of nodes, we relate concepts as sets of nodes to semantic constructs in Quillian's semantic net model [26], and then we use spreading activation to guide the design of the Distance algorithm. First, we study a special kind of concept in Quillian's model, call it a compound concept, and study how spreading activation would operate on such concepts. Then we map documents and queries to compound concepts and translate spreading activation properties into desirable properties for Distance. Finally, we give a complete definition of Distance with a number of important mathematical properties.

1) Compound Concepts in Quillian's Model: In Quillian's model, concepts can be combined to define more complex concepts. In the sequel, concepts that are expressed by an English word will be called *elementary* concepts. A concept such as "the old red house" is represented as a combination of the elementary concepts "old," "red," and "house." Roughly speaking, Quillian's model prescribes that the old red house be represented by a node that has *conjunctive* links to instances of the three concepts "old," "red," and "house,"[3] to say that "the old red house" is at the same time a house *and* a red object and an old object. Conjunctive links can be thought of as links labeled "and" from the node "the old red house" to each one of the three nodes "house," "old," and "red."

Similarly, concepts that have several meanings, such as "plant," are represented by a node that has *disjunctive* links to nodes, each of which is an instance of one of the alternate meanings. To take an example used by Quillian [26], the word "plant" may mean 1) a physical plant, that is a building used for manufacturing processes, 2) plant as a living organism, and 3) the verb "to plant." We will refer to these three meanings by plant_1, plant_2, and plant_3 respectively. Accordingly, the concept plant is represented by a node labeled "plant" that has three links labeled "or" to plant_1, plant_2 and plant_3. This says that plant is plant_1 or plant_2 or plant_3.

Now assume that the concept "plant" is compared to the concept "flower." The two concepts are definitely similar, because in one interpretation (plant_2), "flower" is-a "plant." In general, a disjunctive compound concept matches all concepts that match one of its alternate interpretations. In terms of conceptual distance, the conceptual distance between a disjunctive concept and another concept is the minimum conceptual distance between the disjunctive concept's alternatives and that other concept. We shall later refer to this property as the disjunctive minimum.

For the case of a conjunctive concept, it is important that all the elementary concepts be considered. For example, "the old red car" is not close to "the old red house," although both concepts share the elementary concepts old and red. Conversely, "the old pink mansion" is conceptually close to the "old red house" because "old" equals "old," "pink" is close to "red," and "mansion" is close to "house."

2) Documents and Queries: In information retrieval systems, documents (articles, books, records, etc.) are often characterized by a set of index terms chosen from a hierarchical semantic net. When maximum specificity is sought in the indexing procedure, the index terms representing a document often represent significantly distinct concepts. In this case, removing an index term would

[3]This description is not exact, as it does not handle all the subtleties involved. However, it will suffice for our purposes, and it does not violate any of the basic assumptions of the model.

adversely affect the precision of indexing. The concept reflected by a document is best described by ANDing the concepts represented by its index terms. As such, documents are similar to Quillian's conjunctive concepts.

For queries, the way index terms are combined is explicit. For instance, in many operational information retrieval systems natural language queries are coded into Boolean queries by a trained librarian. A Boolean query is a parenthesized logical expression composed of index terms (atoms) and the logical operators, $\vee$, $\wedge$, and $\neg$. A query consisting of a single term retrieves the documents that have that term in their index. When many terms are used, the operators stand for the corresponding set operations between the sets of documents that would be retrieved by the corresponding single term queries [22].

Using the Quine–McCluskey algorithm [28], a query (or any Boolean expression for that matter) can be converted into minimal disjunctive normal form. As such, a query can be seen as a disjunction of conjunctive compound concepts, except for the fact that conjunctions may contain NOTed terms. Negated terms are difficult to interpret in the context of the semantic net representation. If X is a node in a semantic net, what is $\neg X$? One way to address negations of concepts in semantic nets is through exceptions. Using Touretzky's [29] formalism, if X is-not-a Y and Z is-a X, then Z is-not-a Y; thus $\neg Y$ includes anything that is under X. We regard $\neg Y$ as the set of nodes that are farthest in the semantic net from Y. Then, the conceptual distance between X and $\neg Y$ is the conceptual distance between X and that set. Admittedly, there may be contexts in which the interpretation of negation is inappropriate.

3) Distance on Sets of Nodes: In this section, we use the behavior of spreading activation on compound concepts to guide the extension of Distance to handle sets of nodes. The disjunctive minimum rule translates into the identity:

$$\text{Distance}(C_1 \vee \cdots \vee C_k, C) = \min_{i=1,\cdots,k} \text{Distance}(C_i, C) \quad (1)$$

where C_i, for $i = 1, \cdots, k$ and C represent concepts (compound or elementary). When C itself is a disjunctive concept, $\text{Distance}(C_i, C)$ above is in turn computed as a minimum over the component concepts of C.

When conjunctive concepts are compared, we must take into account the conceptual distances among elementary concepts. In the previous example of "the old red house" and "the old pink mansion," the two concepts are similar because pink is similar to red and mansion is similar to house: we compare values of comparable features, and we do not pay attention to the fact that "pink" is far from "house" and "mansion" is far from "red." This corresponds to a feature comparison process as advocated by Tversky [10] and Smith *et al.* [30]. It is not, however, clear how spreading activation handles feature comparisons; the method has been criticized for not handling features properly [30]. If a clear mapping exists between the elementary concepts of concept X and the elementary concepts of concept Y, Distance need only be applied to pairs of corresponding elementary concepts. (In fact, we later see an experiment where such an approach provided a good measure of conceptual distance.) However, our experience with documents and queries shows that such a mapping is not readily obtainable. Accordingly, we define Distance between conjunctive concepts as

$$\text{Distance}(X_1 \wedge \cdots \wedge X_k, Y_1 \wedge \cdots \wedge Y_m) = \frac{1}{\text{km}} \sum_{i=1}^{k} \sum_{j=1}^{m} \text{Distance}(X_i, Y_j) \quad (2)$$

where the X_i and Y_j are elementary concepts. Notice that we choose to divide the double sum by the product km. This normalization has been used to reduce the bias of number of elementary concepts; without it, concepts with more elementary concepts tend to be further apart. It is also consistent with some of the processing assumptions of spreading activation [11]. Roughly speaking, when a node A is activated and B is adjacent to A, B is subsequently activated by an "amount" inversely related to the number of nodes adjacent to A, and proportional to the strength of the link between A and B.[4] Thus the more elementary concepts a compound concept has, the less (relatively) a path through an elementary concept will account for similarity.

Finally, for some of our applications we need to define Distance between a concept and the "null" concept—an empty set of conjunctive concepts. In our efforts to evaluate semantic nets we have also developed an algorithm called Indexer for automatic indexing of document titles into terms of a semantic net [31]. The performance of Indexer is compared to that of expert human indexers by checking the distance between the human-produced and the computer-produced sets of index terms. Our automatic indexer would at times fail to produce any terms to index a document. However, we had not defined Distance over the empty set initially. Attempts to analyze the experimental data in which these documents were treated as missing values were unsatisfactory. Accordingly, we decided to extend Distance to the case where one of the concepts is the empty set. For indexing purposes, returning the empty set constitutes the worst possible answer. For instance, a document that is not indexed cannot be retrieved at all. Thus we define Distance between a concept X and the "empty set" as the maximum $\text{Distance}(X, Y)$ where Y is any conjunctive compound concept built over all subsets of the set of nodes in the semantic net. This extension of Distance both well captures the importance of the empty set and, as will be seen later, is simple to compute. The final interpretation of the experiments were radically different when we were able to handle the empty set versus when we were not able to handle the empty set. Based on the above considerations,

[4]Activation spreads like electric current in a network of resistors, where the smaller the resistance, the higher the criteriality tag. This last rule is equivalent to Kirschoff's law for nodes in electrical networks.

and on the definition of Distance between concepts represented by single nodes, we define Distance between conjunctive concepts as follows.

Definition 2: Let V be the set of nodes of an is-a semantic net, and let V_1, V_2, and V_3 be three subsets of V, each representing the compound concept consisting of a conjunction of its elements. We have

$$\text{Distance}(V_1, V_2) \equiv \begin{cases} 0, & \text{if } V_1 = V_2 \\ \dfrac{1}{|V_1||V_2|} \displaystyle\sum_{u \in V_1} \sum_{v \in V_2} \text{Distance}(u, v), & \\ & \text{if } V_1 \neq V_2, V_2 \neq \phi, V_2 \neq \phi \end{cases}$$

and

$$\text{Distance}(V_1, V_2) \equiv \begin{cases} \max\limits_{U \supseteq V} \{\text{Distance}(V_1, U)\}, & \\ \quad \text{if } V_2 = \phi \text{ and } V_1 \neq \phi & \\ \max\limits_{U \supseteq V} \{\text{Distance}(U, V_2)\}, & \\ \quad \text{if } V_1 = \phi \text{ and } V_2 \neq \phi & \end{cases}$$

where Distance(u, v) is the shortest path length between nodes u and v (as in definition 1). Let v^{-1} be the set

$$\{ w \in V | \text{Distance}(v, w) = \max_{u,V} \text{Distance}(u, v) \};$$

then

$$\text{Distance}(u, \neg v) = \text{Distance}(\{u\}, v^{-1}). \quad (3)$$

Notice that when $V_1 = \{v_1\}$ and $V_2 = \{v_2\}$ (V_1 and V_2 are singletons), Distance(V_1, V_2) reduces to Distance(v_1, v_2), as in Definition 1. Using (1) above, Distance readily generalizes to arbitrary compound concepts (any combinations of AND's and OR's), provided that these concepts are expressed in disjunctive minimal form.

Notice that (2) does not yield zero for identical compound concepts, and the zero property has to be imposed in Definition 2. This is another problem related to the fact that Distance is computed indiscriminantly between all pairs of concepts. What is needed is a reference value of Distance that attests to the minimum conceptual distance between concepts, that is the conceptual distance between identical concepts. We choose the value zero because it is the smallest value attainable by Distance, and because it happens to correspond to a mathematical property of metrics (zero property). However, we later see that a zero value (or any other fixed value for that matter) leads to some undesirable behavior of Distance (see Section IV-C).

The computation of Distance(U, Φ) according to the above definition would be prohibitively costly. Were we to generate all the subsets of V ($2^{|V|}$ of them), the time to compute Distance(U, Φ) would be exponential. Theorem 2 allows us to compute Distance(U, Φ) in less than the $O(n^3)$ required for the common "all-pairs shortest path" algorithm [32]:

Theorem 2: Let U be a nonempty subset of V, then

$$\text{Distance}(U, \phi) = \max_{v \in V} \text{Distance}(\{v\}, U)$$

The proof of this theorem is given in the Appendix.

Theorem 3: Distance is a metric on the sets of concepts (single and compound) defined on a semantic net.

The proof of this theorem is given in the Appendix.

The mathematical properties of Distance allow us to answer certain questions straightforwardly. For instance, under what conditions are two sets (V_1 and V_2) of nodes closer to or further from a third set V_3? If $V_3 = \{V_1 \cup \{v\}\}$, then we can determine whether $D(V_3, V_2)$ is greater than, equal to, or less than $D(V_1, V_2)$ by determining whether $D(\{v\}, V_2)$ is greater than, equal to, or less than $D(V_1, V_2)$, respectively.

III. Experimental Results

In this section, we study a series of experiments where we use Distance to measure the conceptual distance between concepts. We compute Distance on an is-a hierarchy between concepts represented by single nodes, and concepts represented by sets of nodes. In either case, Distance proves to be a valuable tool to

1) simulate human assessments of conceptual distance and
2) evaluate some cognitive aspects of our semantic nets.

All of our experiments use human subjects as standard references against whom Distance performance is measured. The next section describes in order:

1) the information retrieval context in our experiments,
2) the reliability of human observers,
3) the results of applying Distance to pair of nodes, and
4) the results of using Distance in document retrieval experiments.

A. Experimental Data

The National Library of Medicine maintains one of the world's largest bibliographic retrieval systems, called Medline [33]. Medline contains bibliographic information for over five million articles from over 3000 biomedical periodicals. In addition to the usual bibliographic information (such as author, title, journal, and date of publication), each article is also represented by a set of terms from a semantic net called Mesh (see Fig. 1). Over 2000 queries are addressed to Medline each day from sites around the world. These queries are often encoded as Boolean expressions over Mesh terms.

Mesh is a hierarchical semantic net of over 15000 terms [34]. The 15000 terms are placed into a nine-level hierarchy that includes high-level nodes such as "anatomy," "organism," and "disease" (see Fig. 2). The hierarchy is based on "broader-than" relationships, where the broader terms are higher in the tree. The broader-than relationship is very similar to the is-a relationship [35], but also in-

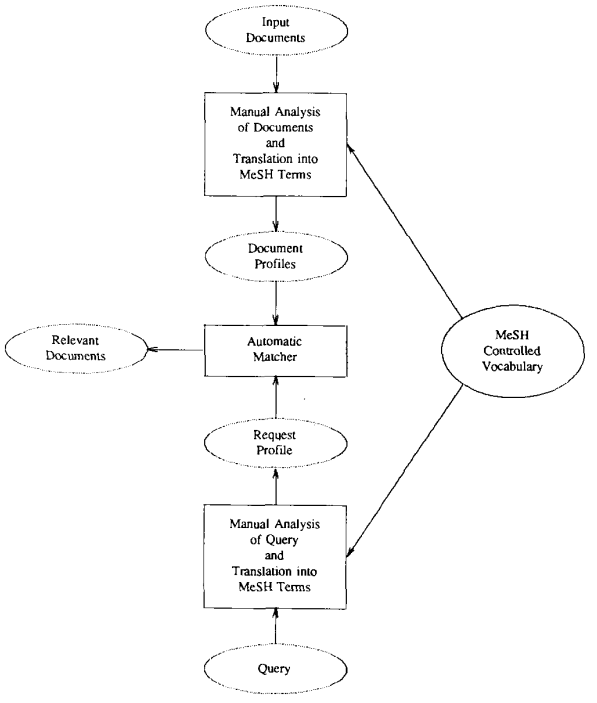

Fig. 1. Medline with emphasis on manual Mesh encoding.

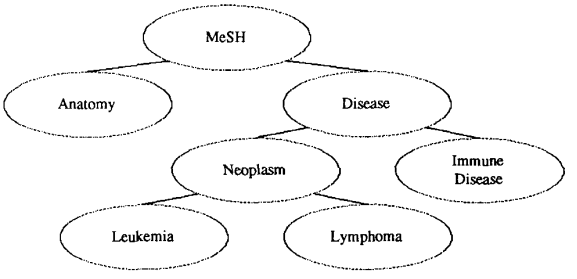

Fig. 2. Small portion of Mesh tree. Overall tree is nine levels deep and includes over 15000 terms.

cludes other broader-than relations, such as "part-of." One of our assumptions is that these other types of broader-than relationship exhibit the same effect on conceptual distance measurements as the is-a relationship does.

B. Observer Reliability

Ranking documents by order of relevance to a query is based on an assessment of the strength of the relationships between each document and the query. Psychological studies have shown that human judges are prone to biases when confronted with a similar task [36]. Therefore, before comparing the performance of Distance to people, we studied the performance of people.

Twenty-two students from a computer science course at George Washington University were given a query about computers and medicine and seven titles from the Medline response to "computers and medicine." Each student was asked to rank the titles at two different times separated by 24 h. For each student, the Spearman ρ correlation coefficient[5] between the two different rankings was computed. The (arithmetic) average of the 22 coefficients was 0.63. This suggests high intraobserver reliability. A Kendall's concordance [37] performed on the 22 first-time rankings of the students gave a value of 0.19. At the 0.05 confidence level, we can reject the null hypothesis that the student ranks were independent.

For a population of six titles and ten queries about "rheumatoid arthritis and knee prosthesis," rankings by two physicians were compared. To analyze the data, we grouped articles with queries in a way that allowed the 60 article–query pairs to be treated as one population for ranking. Then Spearman's correlation coefficient was used to compare two different rankings. The null hypothesis that the two rankings are independent was rejected.

C. Shortest Path Lengths in Mesh

In this section, we test the extent to which the minimum path length between two nodes provides a good measure of conceptual distance between the corresponding concepts. First, we challenge one of the basic assumptions of our model, that is the symmetry of conceptual distance. We mentioned earlier (Section II-A) that people's assessments of conceptual similarity (or distance) may not always correspond to a feature comparison process as prescribed by Tversky [10]. When the concepts to be compared are in an is-a relationship, for example, people might use solely that relation as a basis for their assessments. We also saw that the asymmetry of is-a links does not affect the similarity between concepts as determined by spreading activation [11] because a path of is-a links would be transversed in both directions equally.

We performed a set of experiments to determine whether the conceptual distance between two terms in a broader-than relationship is symmetric. We took a set of pairs of terms from Mesh. The terms constituting a pair in our experiments were such that one was broader-than the other. We asked students to assign numbers to each pair $(\text{term}_1, \text{term}_2)$, in answer to the question, "How close is term_1 to term_2?" The numbers assigned correspond to the conceptual distance between the two. There seemed to be no correlation between the numbers assigned, and which of the terms was broader-than the other. Although the experiment was not conclusive because of the underlying assumptions we made about the matching process (for example, that search in memory would always proceed

[5]Given k entities $e_1, \cdots, e_k$, the Spearman correlation coefficient between two rankings $r_1, \cdots, r_k$ and $r_1', \cdots, r_k'$ is given by

$$\rho = 1 - 6 \times \frac{\sum_{i=1}^{k} (r_i' - r_i)^2}{k(k^2-1)}.$$

The coefficient is 1 for identical rankings, 0 for unrelated rankings, and -1 for inversely related rankings.

from term_1), it suggested that were there to be a difference, it would probably be a small one.

The 1986 edition of Mesh had an inadequate coverage of information science related topics. Accordingly, we initiated a study of how to make the information science part of Mesh better [38]. One resource was the Association of Computing Machinery's hierarchical semantic net for computer scence, called the computing reviews classification structure (CRCS) [39]. The information science section of Mesh had about 200 terms, while CRCS had about 1000 terms in a four-level hierarchy.

Our merging algorithm first determined the similarities between Mesh and CRCS and then exploited the differences. In the simple case where a term t_1 existed in CRCS and not in Mesh but t_1 had a parent t_2 in CRCS which equaled a term t_3 in Mesh, we added t_1 to Mesh as a child of t_3. This algorithm had several other capabilities so that terms in CRCS could also accurately become parents of terms in Mesh. To test whether the merger had created a better semantic net, a sequence of experiments were performed in which human evaluations of distances between terms were compared to those of Distance on Mesh versus Distance on Mesh + CRCS.

Twelve pairs of terms that were both in Mesh and Mesh + CRCS were given to ten computer science students at George Washington University. The students were asked to assign a number between one and five to each pair of terms to indicate what they thought was the conceptual distance between the components of the pair. The 12 pairs of terms were then ranked in increasing order of their distance. Similarly, shortest path lengths in Mesh and Mesh + CRCS were computed for each pair of terms, and ranks were computed from these distances for the two semantic nets.

The average Spearman's correlation coefficient of ten students shows that their rankings significantly agree at the 0.01 level of confidence ($\alpha = 0.01$). Now comparing the average of the students' rankings against Mesh and against Mesh + CRCS, we get

$$\rho_{\text{avg. Stud, Mesh}} = 0.17$$

$$\rho_{\text{avg. Stud, (Mesh + CRCS)}} = 0.52.$$

As a descriptive statistic we can accept that these two correlation coefficients are significantly different. The augmentations provide a better correlation between people and the semantic net. We have done a similar experiment with four physicians who ranked the terms; there again the results clearly show the augmented semantic net as more accurately representing the cognitive distances that people hold true.

In our own subjective evaluations the merged Mesh + CRCS was better than Mesh alone. Distance has helped us systematically document the difference in functionality between Mesh + CRCS and Mesh alone. In general, we have found that Distance is a useful tool for the evaluation of term–term distances in hierarchical semantic nets, and that with a good hierarchical semantic net the rankings determined by Distance roughly correspond to those which people perceive. We have done a host of other experiments with term–term distances in hierarchical semantic nets and with variants on Distance and always concluded the same thing, namely, that this approach to validating semantic net merging strategies has merit.

D. Distance Applied to Documents and Queries

Given a document D characterized by a set of index terms $D = \{t_{D,1}, t_{D,2}, \cdots, t_{D,n}\}$ and a query Q coded into an ANDed set of index terms $Q = \{t_{Q,1}, t_{Q,2}, \cdots, t_{Q,m}\}$, we hypothesized that the distance between D and Q gives a measure of the conceptual distance of the document to the query. In the experiments reported below, we computed Distance between a query and a number of documents. We then ranked the documents with Distance, assuming that the greater the distance between the query and a document, the less relevant the document was to the query. Similarly, we asked people to rank a set of documents with respect to a given query and then compared their ranks to those produced by Distance.

For ten different queries and six articles the averages of two physicians' evaluations were compared to those produced by Distance on Mesh. The agreement between the computer and the people was significant at the 0.05 level. To show that this ranking by the computer depended on more than the exact matches among terms of the query and document, the experiments were repeated but now with path lengths constrained. If only exact matches between terms in the query and document descriptions were used, then there was a negative correlation between the people's and comuter's rankings, proving that Distance was sensitive to the structure of Mesh.

Two scientists compared each of 52 documents against the query "lipids and encephalitogenic basic proteins." The 52 documents were retrieved from Medline by a search with the term lipids in it. Each document was represented by all the Mesh terms stored in Medline for that document (typically, ten terms per document). The ranking of each scientist and the ranking of Distance was statistically significantly correlated. The correlation between the rankings of the two scientists was also significant at the 0.05 level. The same methodology was applied to the queries "suicide and substance dependence," "liver diseases and peritoneoscopy," "shock and endorphins," and "biocompatible materials and dental implementation," and the same results attained. That is, the human judges agreed with each other and with Distance in the ranking of documents to query. These and other experiments support the claim that Distance on Mesh sets a baseline for performance that is not disconnected from the decisions of people regarding the conceptual similarity between sets of terms.

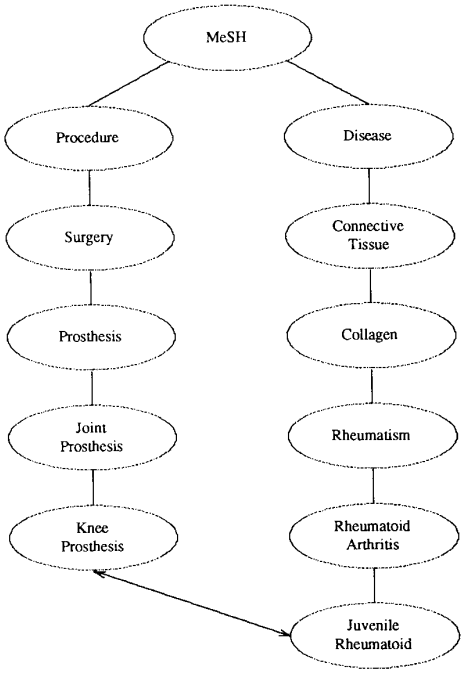

Fig. 3. Deep hierarchy of Mesh is evident here. Added edge between "juvenile rheumatoid arthritis" and "knee prosthesis" misled Distance.

NA: Keratoconus
AT: Cornea, conical.
ET: Hereditary; associated with Down syndrome, atopic dermatitis, Marfan syndrome, retinitis pigmentosa, aniridia, vernal catarrh, Alpert syndrome, Ehlers-Danlos syndrome.
SX: Blurred vision uncorrected by glasses.
SG: More frequent in females; onset at puberty; myopia; astigmatism; possibly more advanced in one eye, eventually bilateral.
LB: Ophthalmoscopy: progressive bulging of cornea; apex of cone usually slightly below center of cornea; corneal protrusion recognized by viewing eye from side; sometimes pulsation of corneal conus synchronous with arterial pulse. Increased intraocular tension; clefts in Descemet membrane. Retinoscopy: distortion of light reflex; distortion of appearance of nerve head, vessels of fundus. Keratoscopy: distortion of corneal light reflex.
CR: Prognosis: astigmatism progressing for years, then becoming stationary; possible corneal perforation.
PA: Opacity at apex of cone; line of gray, yellow, or olive-green pigment forming incomplete ring.

Fig. 4. Example of CMIT disease description. Disease is keratoconus. Fields in CMIT mean: NA is name, AT is alternate terms, ET is etiology, SX is symptoms, SG is signs, LB is laboratory findings, CR is course, PA is pathology.

IV. Directions

A. Distance Applied to Other Relations

In studying a document set about juvenile rheumatoid arthritis (JRA), we applied Distance to Mesh alone and to Mesh augmented with other relationships [40]. The augmented Mesh contained additional edges such as the one connecting JRA to knee prothesis (see Fig. 3) and to granuloma. A knee prothesis can be used in the surgical treatment of a knee destroyed by JRA, and a granuloma is a pathological finding in JRA. We hypothesized that the addition of edges or relationships from another knowledge base to Mesh would augment the ability of Distance to reflect the decisions of people about cognitive distance. Distance applied to this augmented Mesh failed to simulate people.

These results may appear to be predictable because Distance was designed to work for is-a links. However, the property of is-a links that seemed to matter most was that they acted like highly critical links [11], and thus, links of is-a paths provided a good measure of conceptual Distance. In this experiment, we hypothesized that because relations such as "cause" or "treat" are important and should have high criteriality, shortest paths along those relations should also indicate conceptual distance. Instead, it became clear that broader-than relationships had a peculiar significance in adjudicating distance.

When Distance was modified so as to cross only such nonhierarchical relationships when both the query and the document made clear that they were "about" such relationships, then this enhanced Distance again correlated well with the ranking decisions of people. Each nonhierarchical relation, such as "cause" or "treat," had to be handled in a distinct way [40]. The ramifications of such enhancements to metric aspects of Distance have only been partially explored.

B. Distance Applied to Featural Models

In one set of experiments, we tried to merge current medical information and terminology (CMIT) with Mesh [41]. CMIT is a system for naming and describing diseases that is produced by the American Medical Association. CMIT describes approximately 3600 diseases in a structured format with eight attributes. The eight attributes are alternate terms, etiology, symptoms, signs, laboratory findings, radiologic findings, course, and pathology. Within each attribute, the description of a disease consists of a series of noun phrases, usually separated from one another by a semicolon. These noun phrases are a terse form of natural language (see Fig. 4). First CMIT diseases were "parsed" by lexical matching into Mesh. Only those CMIT terms or phrases that occurred specifically as Mesh terms were retained. This method had the advantage that all the resulting terms for characterizing diseases from CMIT were embedded in the hierarchy of Mesh.

CMIT has many diseases that Mesh does not. One of our goals was to develop an algorithm that would insert these CMIT diseases into the appropriate place in Mesh and thus improve Mesh. The strategy was first to locate all the diseases that existed in both Mesh and CMIT and to use those as references in Mesh against which conceptual distance of CMIT diseases could be assessed. We identified a number of such diseases and for each disease we added to Mesh the edges that connected the disease to its attributes, thus producing Mesh + CMIT.

The first attempt at computing conceptual distance was to treat each disease as a compound concept where the attributes were ANDed. To compute the conceptual distance between two such diseases, Distance was applied between the sets of attributes, the same way that Distance was applied in our earlier document retrieval

experiments. To test the validity of this approach, we determined the extent to which the shortest path length between two disease names was correlated to the Distance between their sets of attributes. For example, "myopia" and "hyperopia" existed in both Mesh and CMIT. Their distance in Mesh is two. We applied Distance to their sets of attributes, and although we did not expect this Distance to yield exactly two, we expected that, by taking a number of such diseases, there would be a correlation between the ranking on disease names and the ranking on disease attributes.

Ten eye diseases that had the same names in CMIT and Mesh were first used. Distance was applied to all pairs of the disease names and then to all pairs of descriptions of these eye diseases. From the two sets of scores we derived rankings and then checked the degree of correlation between the rankings based on Mesh disease names alone versus on disease descriptions. The correlation coefficient was 0.06—there was lack of agreement between the rankings.

Examination of the Mesh + CMIT descriptions and the path lengths between terms suggests that too many distances were being calculated that were not meaningful. Each attribute of a disease should be treated separately. For instance, the etiology feature for a disease is meaningfully compared to the etiology feature of other diseases but not to the laboratory findings feature. Taking advantage of the breakdowns in CMIT requires treating the different attributes of a disease differently. The distance between the etiology and laboratory findings features is less important than the distance between two etiologies or the distance between two laboratory findings. This is similar to our earlier example about "the old red house" and "the old pink mansion," where Distance should be applied selectively, i.e., as a feature comparison process, rather than applying it to all pairs of elementary concepts. Unlike the case for documents where a mapping of index terms along semantically distinct dimensions does not exist (and for which Distance performed well), a mapping not only exists in this case but may well be essential to the success of Distance, i.e., a breakdown or decomposition of the Mesh + CMIT descriptions into their natural parts, like etiologies and laboratory findings, might allow for distances that more closely related to the distances among Mesh disease names.

The etiology components of the diseases were next isolated. The hypothesis was that on the etiologies the diseases would have distances from one another that more closely corresponded to those distances that existed between the names alone. The same steps as used for assessing the correlation between Mesh disease names and Mesh + CMIT descriptions were now used to assess the correlation between the rankings on etiologies and the rankings on disease names. The degree of correlation was significant. Incidentally, these experiments proved that some of the Mesh diseases are hierarchically organized according to etiologies.

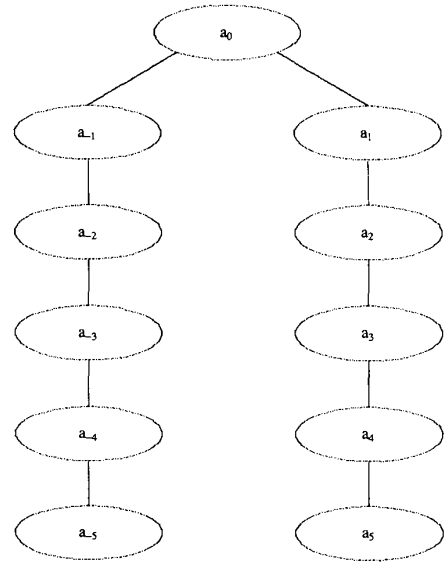

Fig. 5. Sample graph G_5 used to illustrate discontinuity in Distance.

C. Distance Near Zero

Distance fails in some ways to capture the intuitive notions of what it means for one concept to be close to another. Consider the query (knee prothesis AND rheumatoid arthritis). According to Distance on Mesh a document indexed (joint prosthesis AND rheumatoid arthritis) is as close to the query as a document indexed under disease[6] (see Fig. 3). Distance would also say that a document indexed under disease is closer to the query than a documented indexed (juvenile rheumatoid arthritis AND knee prothesis)!

Consider the case of graph G_5 with a root and two linear branches of five nodes each (see Fig. 5),

$$\text{Distance}(\{a_0, a_5, a_5\}, \{a_0, a_i, a_i\})$$

gets larger as i goes from one to four, but when $i = 5$, there is a drastic drop in the value of Distance to zero. G_5 can be generalized to G_n in which a_k and a_{-k} are k edges from a_0, and the leaves are a_n and a_{-n}. The pathology true for G_5 is also true for G_n. The Distance between $\{a_0, a_{-n}, a_n\}$ and $\{a_0, a_{-i}, a_i\}$ steadily grows as i grows but drops to zero when i reaches n.

The problem with the above cases is not so much the fact that there is a discontinuity near zero—after all, we imposed the zero property by definition. More damaging is that Distance values may increase as the conceptual distance seems to decrease. The basic problem seems to be that Distance between sets of nodes treats equally all pairs of nodes. This becomes an issue for extreme cases as the ones presented above. This situation is not unrelated to the

[6]Let the reader be assured that Medline indexing manuals do not allow a document to be simply indexed under "Disease."

problem we had originally had with the featural model. We have considered and found weaknesses in several metric alternatives to Distance, such as the distance between centroids for each set of nodes, and several nonmetric alternatives, such as the path length between the closest nodes in two sets.

V. Conclusion

Our research group has been for the past two years developing methods of merging semantic nets, such as Mesh and CMIT [1]. As is typical for such machine learning experiments, issues of representation and reasoning are as critical as those of learning. Our hypothesis is that better semantic nets result from the mergers, but to evaluate "betterness" we need a way of reasoning with a semantic net. This leads to the development of a measure of conceptual distance between sets of nodes in a semantic net. By transforming documents and queries into sets of nodes, we can do information retrieval experiments that test the value of our semantic net.

In our search for an evaluation tool, we realize that certain cognitively meaningful and mathematically convenient properties are desirable. Although the relationships in our semantic nets are directed, e.g., broader-than and narrower-than, our early experiments suggest that for the purposes of conceptual distance, these relationships can be treated as undirected. Accordingly, the measure of distance over sets of nodes has the property of symmetry. Furthermore, we are interested in knowing under what conditions one document is far from or close to another document. For this purpose a property like the triangle inequality is useful. The work in memory-based reasoning [24] is one example where metric properties are important. Our measure of conceptual distance, called Distance, satisfies the properties of a metric. On the surface it is surprisingly simple—just the average of the path lengths between pairs of nodes. However, it has proven remarkably powerful and flexible.

In our efforts to evaluate semantic nets, we have also developed an algorithm called Indexer for automatic indexing of document titles into Mesh [31]. In one set of experiments we added thousands of synonyms to the main terms of Mesh and tested the main-terms-plus-synonyms semantic net with Indexer. We compared the performance of Indexer against human indexers by counting the number of hits and misses. To our surprise the synonyms did not increase the hits any more than they increased the misses. Then we refined the measure of performance by applying Distance. When Distance measured the distance between the human and machine results, the synonyms proved to be helpful. In other words, the synonyms usually led Indexer to be closer to the human indexing. The measure of absolute hits and misses was unrealistically demanding of Indexer. In artificial intelligence experiments it might be expected that the machine gives answers close, but not identical, to those that humans would, and a way to measure this closeness is important.

In looking at the distance between sets of nodes, we have had to deal with different kinds of relationships. It seems that the broader-than and narrower-than relations can be treated in basically the same way, but other relationships, like cause, merit different handling. For queries and documents we argue that nonhierarchical relationships should only be traversed when both query and document specify that they are about that relationship. The integrity of the node-to-node path lengths in hierarchical semantic nets are the key to the success or failure of Distance as a measure of conceptual distance. In our evaluations, these path lengths have shown themselves to be cognitively meaningful. Some have argued that spreading activation does not spread across more than one link [42], [43], but they ignore link labels, while we have shown the importance of distinguishing hierarchical from nonhierarchical links.

Distance might be implemented in an information retrieval system which is based on the indexing of documents and queries into terms from a semantic net (see Fig. 1). Often in such systems a query retrieves more documents than the user wants, and the documents are arbitrarily ordered as they appear on the computer screen. A measure like Distance might be applied to help rank the documents to the query and allow the querist to pay most attention to those documents which are most like to be conceptually close to the query.

There are a host of specific questions about the cognitive realism of Distance that we have not adddressed. For instance, to the extent that the semantic net is a tangled hierarchy and has levels, should terms at different levels be treated differently [17]? Relative to Distance, "part-of" links seem to obey the same properties as is-a links, but how would a metric on causal links look? We do not claim that the brain is making Distance-like calculations in the course of determining cognitive similarity. Nor do we argue that the tangled hierarchy and Distance are adequate for other cognitive tasks [44]. Cognition probably does not rely on measurements that satisfy the properties of a metric. A metric has, however, many attractive features because of its mathematical and semantic tractability. We claim that the better the semantic net on which Distance operates, the more the conceptual similarity decisions of Distance match the conceptual similarity decisions of people. We have been surprised at how powerful a simple algorithm like Distance can be in evaluating hierarchical semantic nets.

Acknowledgment

Donald Bamber of the Navy Personnel Research and Development Center provided the important examples of the weaknesses of Distance as two sets approach one

another in the graph. Referees for this TRANSACTIONS provided guidance on substantial revisions of this paper.

APPENDIX

Theorem 2: Let U be a nonempty subset of V, then

$$\text{Distance}(U,\phi) = \max_{v \in V} \text{Distance}(\{v\},U).$$

Proof of Theorem 2: To prove Theorem 2, we first show the following lemmas.

Lemma 1: Let V_1, V_2, and V_3 be three nonempty sets of V, such that V_1 and V_2 are disjoint. If $\text{Distance}(V_1,V_3) = \text{Distance}(V_2,V_3)$, then

$$\text{Distance}(V_1 \cup V_2, V_3) = \text{Distance}(V_1,V_3). \tag{4}$$

In general,

$$\text{Distance}(V_1 \cup V_2, V_3) \geq \min_{i=1,2} \{\text{Distance}(V_i,V_3)\} \tag{5a}$$

$$\text{Distance}(V_1 \cup V_2, V_3) \leq \max_{i=1,2} \{\text{Distance}(V_i,V_3)\}. \tag{5b}$$

Proof of Lemma 1: Let $V_1 = \{v_1^1,\cdots,v_1^k\}$, $V_2 = \{v_2^1,\cdots,v_2^q\}$, and $V_3 = \{v_3^1,\cdots,v_3^m\}$,

$$\begin{aligned}
&\text{Distance}(V_1 \cup V_2, V_3) \\
&\quad = \frac{1}{(k+q)m} \sum_{v_i \in V_1 \cup V_2} \sum_{v_j \in V_3} d(v_i,v_j) \\
&\quad = \frac{1}{(k+q)m} \\
&\qquad \cdot \left\{ \sum_{v_i \in V_1} \sum_{v_j \in V_3} d(v_i,v_j) + \sum_{v_i \in V_2} \sum_{v_j \in V_3} d(v_i,v_j) \right\} \\
&\quad = \frac{k}{k+q} \text{Distance}(V_1,V_3) \\
&\qquad + \frac{q}{k+q} \text{Distance}(V_2,V_3).
\end{aligned}$$

When $\text{Distance}(V_1,V_3) = \text{Distance}(V_2,V_3)$, (4) follows immediately. When the distances are different, inequalities (5a) and (5b) follow because both k and q are positive.

Lemma 2: Let U be a nonempty subset of V and $U_{\max}$ a subset of V such that

$$\text{Distance}(U,U_{\max}) = \text{Distance}(U,\phi),$$

then for all $u \in U_{\max}$, we have

$$\begin{aligned}
\text{Distance}(\{u\},U) &= \text{Distance}(U_{\max},U) \\
&= \text{Distance}(U,\phi).
\end{aligned}$$

Proof of Lemma 2: a) By putting $V_1 = \{u\}$, $V_2 = U_{\max} - \{u\}$, and $V_3 = U$ in the previous lemma, we conclude from (11) that if

$$\text{Distance}(\{u\},U) = \text{Distance}(U_{\max} - \{u\},U),$$

then

$$\begin{aligned}
\text{Distance}(\{u\},U) &= \text{Distance}(U_{\max},U) \\
&= \text{Distance}(U,\phi).
\end{aligned}$$

b) Let us prove that it is always the case. Assume that

$$\text{Distance}(\{u\},U) \neq \text{Distance}(U_{\max} - \{u\},U),$$

then, according to Lemma 1,

$$\begin{aligned}
\text{Distance}(U_{\max},U) &> \min\{\text{Distance}(\{u\},U), \\
&\qquad \text{Distance}(U_{\max} - \{u\},U)\} \\
&< \max\{\text{Distance}(\{u\},U), \\
&\qquad \text{Distance}(U_{\max} - \{u\},U)\}.
\end{aligned}$$

This implies that either $\{u\}$ or $U_{\max} - \{u\}$ is further from U than $U_{\max}$, which contradicts the definition of $U_{\max}$.

Going back to the proof of Theorem 2, let $v_{\max}$ be an element of V such that

$$\text{Distance}(\{v_{\max}\},U) = \max_{v \in V} \text{Distance}(\{v\},U)$$

By definition of $\text{Distance}(U,\phi)$, we have

$$\text{Distance}(U,\phi) \geq \text{Distance}(\{v_{\max}\},U)$$

or, for some set $U_{\max}$,

$$\text{Distance}(U,U_{\max}) \geq \text{Distance}(\{v_{\max}\},U).$$

Let us assume that

$$\text{Distance}(U,U_{\max}) > \text{Distance}(\{v_{\max}\},U).$$

According to Lemma 2, for all $u \in U_{\max}$, we have

$$\begin{aligned}
\text{Distance}(\{u\},U) &= \text{Distance}(U_{\max},U) \\
&= \text{Distance}(U,\phi).
\end{aligned}$$

Therefore, for all $u \in U_{\max}$,

$$\text{Distance}(\{u\},U) > \text{Distance}(\{v_{\max}\},U),$$

which contradicts the definition of $v_{\max}$. Therefore we have

$$\text{Distance}(U,U_{\max}) = \text{Distance}(U,\{v_{\max}\}).$$

Theorem 3: Distance is a metric on the sets of concepts defined on a semantic net.

Proof of Theorem 3: 1) By definition, for all $U \supseteq V$, $\text{Distance}(U,U) = 0$. 2) Regarding symmetry, when neither V_i $(i=1,2)$ is empty, it is easy to see that $\text{Distance}(V_1,V_2) = \text{Distance}(V_2,V_1)$. This can be done by interchanging the order of summation and using the symmetry of the distance d. 3) For all V_1 and V_2 subsets of V, $\text{Distance}(V_1,V_2)$ is positive because it is computed as an averaged sum of positive numbers. When one of the subsets is empty, Distance is then a maximum of such averaged sums:

$$\begin{aligned}
\text{Distance}(V_1,\phi) &= \max_{U \supseteq V} \{\text{Distance}(V_1,U)\} \\
&= \max_{U \supseteq V} \{\text{Distance}(U,V_1)\} \\
&= \text{Distance}(\phi,V_1).
\end{aligned}$$

4) For the triangle inequality, we have to prove that

$$\text{Distance}(V_1,V_3) \leq \text{Distance}(V_1,V_2) + \text{Distance}(V_2,V_3).$$

a) Let $V_1 = \{v_1^1, \cdots, v_1^k\}$, $V_2 = \{v_2^1, \cdots, v_2^q\}$, and $V_3 = \{v_3^1, \cdots, v_3^m\}$ be three nonempty subsets of V. We have

$$\begin{aligned}\text{Distance}(V_1, V_2) &= \frac{1}{kq} \sum_{i=1}^{k} \sum_{j=1}^{q} d(v_1^i, v_2^j) \\ &= \frac{1}{kqm} \sum_{i=1}^{k} \sum_{j=1}^{q} m \times d(v_1^i, v_2^j) \\ &= \frac{1}{kqm} \sum_{i=1}^{k} \sum_{j=1}^{q} \sum_{p=1}^{m} d(v_1^i, v_2^j). \quad (6)\end{aligned}$$

Similarly, we can write Distance(V_2, V_3) as

$$\text{Distance}(V_2, V_3) = \frac{1}{kqm} \sum_{i=1}^{k} \sum_{j=1}^{q} \sum_{p=1}^{m} d(v_2^j, v_3^p). \quad (7)$$

Adding (6) and (7) yields

$$\begin{aligned}&\text{Distance}(V_1, V_2) + \text{Distance}(V_2, V_3) \\ &\quad = \frac{1}{kqm} \sum_{i=1}^{k} \sum_{j=1}^{q} \sum_{p=1}^{m} d(v_1^i, v_2^j) + d(v_2^j, v_3^p). \quad (8)\end{aligned}$$

Because d satisfies the triangle inequality, we have

$$d(v_1^i, v_2^j) + d(v_2^j, v_3^p) \geq d(v_1^i, v_3^p). \quad (9)$$

From (8) and (9) it follows that

$$\begin{aligned}&\text{Distance}(V_1, V_2) + \text{Distance}(V_2, V_3) \\ &\quad \geq \frac{1}{kqm} \sum_{i=1}^{k} \sum_{j=1}^{q} \sum_{p=1}^{m} d(v_1^i, v_3^p). \quad (10)\end{aligned}$$

Because the summand does not depend on the index j, the summation over j is equivalent to multiplying the summand by q. Implementing this change and eliminating q yields:

$$\begin{aligned}&\text{Distance}(V_1, V_2) + \text{Distance}(V_2, V_3) \\ &\quad \geq \frac{1}{km} \sum_{i=1}^{k} \sum_{p=1}^{m} d(v_1^i, v_3^p) \quad (11) \\ &\quad \geq \text{Distance}(V_1, V_3). \quad (12)\end{aligned}$$

This establishes the triangle inequality in the case where the three sets are nonempty.

b) When V_2 is empty, then there exists at least one subset of V, $V_{\max}$, such that

$$\text{Distance}(V_1, \phi) = \text{Distance}(V_1, V_{\max}).$$

In the previous steps we proved that

$$\begin{aligned}\text{Distance}(V_1, V_3) \leq{}& \text{Distance}(V_1, V_{\max}) \\ &+ \text{Distance}(V_{\max}, V_3).\end{aligned}$$

From the definition of Distance(ϕ, V_3), we know that

$$\text{Distance}(V_{\max}, V_3) \leq \text{Distance}(\phi, V_3).$$

Therefore,

$$\text{Distance}(V_1, V_3) \leq \text{Distance}(V_1, \phi) + \text{Distance}(\phi, V_3).$$

c) When V_3 is empty, we write Distance(V_1, ϕ) as Distance$(V_1, V_{\max})$ and reduce this case to the previous one for which the triangle inequality was proven to hold.

References

[1] R. Forsyth and R. Rada, *Machine Learning: Expert Systems and Information Retrieval*. London: Ellis Horwood, 1986.

[2] H. Mili and R. Rada, "Merging thesauri: Principles and evaluation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-10, no. 2, pp. 204–220, 1988.

[3] R. Rada and E. Bicknell, "Ranking documents with a thesaurus," *J. Amer. Soc. Inform. Sci.*, in press.

[4] R. H. Hopkins, K. B. Campbell, and N. S. Peterson, "Representations of perceived relations among the properties and variables of a complex system," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-17, pp. 52–60, Jan./Feb. 1987.

[5] G. Loberg, G. M. Powell, A. Orefice, and J. D. Roberts, "Representing operational planning knowledge," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-16, pp. 774–787, Nov./Dec. 1986.

[6] S. Miyamoto, K. Oi, O. Abe, A. Katsuya, and K. Nakayama, "Directed graph representations of association structures: A systematic approach," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-16, no. 1, pp. 53–61, 1986.

[7] E. Reingold, J. Nievergelt, and N. Deo, *Combinatorial Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[8] S. Y. Lu, "A tree-matching algorithm based on node splitting and merging," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 249–256, Mar. 1984.

[9] M. A. Eshera and K. S. Fu, "A graph distance measure for image analysis," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-14, pp. 398–407, May/June 1984.

[10] A. Tversky, "Features of similarity," *Psych. Rev.*, vol. 84, pp. 327–352, 1977.

[11] A. M. Collins and E. F. Loftus, "A spreading activation theory of semantic processing," *Psych. Rev.*, vol. 82, pp. 407–428, 1975.

[12] R. Fikes and T. Kehler, "The role of frame-based representation in reasoning," *Commun. Assoc. Comput. Mach.*, vol. 28, no. 9, pp. 904–920, Sept. 1985.

[13] C. Hoede, "Similarity in knowledge graphs," Dep. Appl. Math., Twente Univ. of Technology, 7500 AE Enschede, The Netherlands, Memor. 550, Jan. 1986.

[14] R. Prieto-Diaz and P. Freeman, "Classifying software for reusability," *IEEE Software*, vol. 4, pp. 6–16, Jan. 1987.

[15] D. Rumelhart and D. Norman, *Representation in Memory*. La Jolla, CA: Center for Human Information Processing, June 1983.

[16] R. Brachman, "What IS-A is and isn't: An analysis of taxonomic links in semantic networks," *Computer*, vol. 16, no. 10, pp. 30–36, 1983.

[17] B. Adelson, "Comparing natural and abstract categories: A case study from computer science," *Cogn. Sci.*, vol. 9, no. 4, pp. 417–430, 1985.

[18] D. Nau and T. C. Chang, "Problem solving knowledge in a frame-based process planning system," *Inter. J. Intell. Syst.* vol. 1, no. 1, pp. 29–44, Spring 1986.

[19] B. Buchanan and L. M. Fu, "Learning immediate concepts in constructing a hiearchical knowledge based," in *Proc. 9th Int. Joint Conf. Artificial Intell.*, 1985, pp. 659–666.

[20] A. S. Pollitt, "A rule-based system as an intermediary for searching cancer therapy literature on MEDLINE," in *Intelligent Information Systems: Progress and Prospects*, R. Davis, Ed. London: Horwood, 1986, pp. 82–126.

[21] P. Shoval, "Principles, procedures and rules in an expert system for information retrieval," *Inform. Processing Management*, vol. 21, no. 6, pp. 475–487, 1985.

[22] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[23] E. Fox, "Extending the Boolean and vector space models of information retrieval with *P*-norm queries and multiple concept types," Ph.D. dissertation, Dep. Comput. Sci., Cornell Univ., Ithaca, NY, 1983.

[24] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Commun. Assoc. Comut. Mach.*, vol. 29, no. 12, pp. 1213–1228, 1986.

[25] A. Ortony, "Beyond literal similarity," *Psych. Rev.*, vol. 86, pp. 161–180, 1979.
[26] M. R. Quillian, "Semantic memory," in *Semantic Inform. Processing*, M. Minsky, Ed. Cambridge, MA: MIT Press, 1968.
[27] R. J. Brachman and J. G. Schmolze, "An overview of the KL-ONE knowledge representation system," *Cogn. Sci.*, vol. 9, pp. 171–216, 1986.
[28] E. J. McCluskey, "Minimization of Boolean functions," *Bell Syst. Tech. J.*, vol. 35, no. 6, pp. 1417–1444, 1956.
[29] D. Touretzky, "The mathematics of inheritance systems," Ph.D. dissertation, Dep. Comput. Sci., Carnegie–Mellon Univ., Pittsburgh, PA, May 1984.
[30] E. E. Smith, E. J. Shoben, and L. J. Rips, "Comparison processes in semantic memory," *Psych. Rev.*, pp. 214–241, 1974.
[31] R. Rada, L. Darden, and J. Eng, "Relating two knowledge bases: The role of identity and part-whole," in *The Role of Language in Problem Solving*, vol. 2, R. Jernigan, Ed. Amsterdam, The Netherlands: Elsevier, 1987, pp. 71–91.
[32] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Comptuer Algorithms*. Reading, MA: Addison-Wesley, 1974.
[33] D. B. McCarn, "Medline: An introduction to on-line searching," *J. Amer. Soc. Inform. Sci.*, vol. 31, no. 3, pp. 181–192, May 1980.
[34] J. Backus, S. Davidson, and R. Rada, "Searching for patterns in the Mesh vocabulary," *Bull. Med. Lib. Assoc.*, vol. 75, no. 3, pp. 221–227, July 1987.
[35] N. Library and Inform. Assoc. Council, *Guidelines for Thesaurus Structure, Construction, and Use*. New York: Amer. Nat. Standards Inst., 1980.
[36] J. P. Schwartz, J. H. Kullback, and S. Shrier, "A framework for task cooperation within systems containing intelligent components," *IEEE Trans. Syst. Man Cybern.*, vol. 16, no. 6, pp. 788–793, Nov./Dec. 1986.
[37] S. Siegel, *Nonparametric Statistics*. New York: McGraw-Hill, 1956.
[38] R. Rada *et al.*, "A vocabulary for medical informatics," *Comput. Biomed. Res.*, vol. 20, pp. 244–263, 1987.
[39] J. Sammet and A. Ralston, "The new (1982) computing reviews classification system—Final version," *Commun. Assoc. Comput. Mach.*, vol. 25, no. 1, pp. 13–25, Jan. 1982.
[40] R. Rada, "Gradualness facilities knowledge refinement," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-7, no. 5, pp. 523–530, Sept. 1985.
[41] S. Lester and R. Rada, "A method of medical knowledge base augmentation," *Methods of Inform. Med.*, vol. 26, no. 1, pp. 31–39, 1987.
[42] J. Holland, K. Holyoak, R. Nisbett, and P. Thagard, *Induction: Process of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press, 1986.
[43] A. M. B. de Groot, "The range of automatic spreading activation in word priming," *J. Verbal Learning Verbal Behavior*, vol. 22, pp. 417–436, 1983.
[44] W. Walker and W. Kintsch, "Automatic and strategic aspects of knowledge retrieval," *Cogn. Sci.*, vol. 9, pp. 261–283, 1985.

Roy Rada received the B.A. degree in psychology from Yale University, New Haven, CT, the M.D. degree from Baylor College of Medicine, Houston, TX, the M.S. degree in computer science from the University of Houston, Houston, TX, and the Ph.D. degree in computer science from the University of Illinois at Urbana.

He was an Assistant Professor of Computer Science at Wayne State University from 1981 to 1984. He worked from 1985 to 1988 as Editor of *Index Medicus* at the National Library of Medicine in Bethesda, MD and currently holds a Chair in Computer Science at the University of Liverpool. His research interests focus on intelligent information systems.

Hafedh Mili received the B.S. degree in mathematics and physics from Lycee Mixte de Jemmal, Tunisia, a Diploma in applied mathematics from Ecole Centrale de Paris, France, and the Ph.D. degree in computer science from Geoge Washington University, Washington, DC.

He has been employed on an NSF Research Associateship and an IBM Fellowship. His main interests are in knowledge representation and intelligent information systems.

Ellen Bicknell received the B.S. degree from Rice University, Houston, TX, and the Ph.D. degree from Brown University, Providence, RI, both in chemistry.

She held Post-Doctoral Fellowships in Florida and Oregon and was an Associate Professor of Computer Science at Wayne State University in Detroit, MI, from 1977 to 1986. From 1986 to 1988 she was a Special Expert at the National Library of Medicine. She is interested in both computers and chemistry.

Maria Blettner received the Diplom and Ph.D. degrees in statistics from the University of Dortmund, Germany.

She worked as a Biostatistician at the International Agency for Research on Cancer in Lyon, France from 1983 to 1985 and the National Cancer Institute in Bethesda, MD, from 1985 to 1988. Currently, she is working as a Lecturer in the Department of Statistics and Computational Mathematics at the University of Liverpool, Liverpool, England. Her main interest is in the application and development of statistical methods in biomedicine.