MACGREGOR, JAMES
--Alternative index systems for Telidon :
recommendations based on behavioral consideration

ALTERNATIVE INDEX SYSTEMS FOR TELIDON: RECOMMENDATIONS BASED ON BEHAVIORAL CONSIDERATIONS

JAMES MACGREGOR

NEWMAN LAM

MARCH, 1982

FOR: DEPARTMENT OF COMMUNICATIONS, OTTAWA, CANADA
UNDER: DEPARTMENT OF SUPPLY & SERVICES CONTRACT SERIAL No. OER1-05088
PERIOD: JANUARY 7 - MARCH 31, 1982

ABSTRACT

The research examined alternative index systems for telidon. A literature review was carried out which resulted in recommendations to shorten search paths, to avoid higher levels in the tree, to provide options for entering the tree at the level of "basic level categories", and to use "prototypes" of basic level categories as keywords to effect that entry.

Two pilot studies were carried out to test the feasibility of these recommendations. In one, subjects were asked to generate index terms in response to 30 questions. In the second, subjects were asked to name places where they would go in search of the information to answer the questions.

The main results of the experiments were that : (a) There were high levels of consistency across subjects in generating index terms and place names in response to questions (the mean percentages of agreements were 67% and 49% respectively); (b) The index terms were significantly more successful in entering a tree than the tree method used alone (the percentages of successes were 82% and 73% respectively). (c) Subjects preferentialy generated items that clustered around a particular level in the tree. (d) A keyword index may be unsuccessful if it fails to match these preferences. Index-6 appeared to use keywords that were too specific given subjects' preferences. (e) There were indications of a simple linear relationship between the percentage of successful entries and the number of keywords in an index, across a

relatively wide range of index sizes.

The results indicated that hybrid keyword/tree index systems may have advantages over tree methods alone in terms of increases in success rate, decreases in search time and decreases in the number of pages accessed during search.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

INTRODUCTION

The report describes an investigation of alternative index systems for telidon. The terms of reference for the research were (a) to search for any existing literature which addresses the problem, and (b) to discover and describe how an alternative index structure may be created.

In addition to meeting these objectives, the report examined the structural characteristics of a tree (Tree-A) to determine whether retrieval might be facilitated through changes in such purely structural properties as the number of levels in the tree, the number of categories at levels, and the number of elements in categories. The report also describes several pilot studies which were carried out as initial tests of the feasibility of alternative index systems. It should be stressed that these studies were pilot studies. They did not have the large numbers of subjects, the tight controls or the sophisticated method of analysis which one would require of a hypothesis-testing experimental design. The results are necessarily tentative.

The final report is organized in the following way. The first section describes the structural analysis of Tree-A. The second section summarizes the literature reviewed, presents recommendations based on the literature, and suggests strategies for implementing the recommendations. The final section describes the pilot studies which were conducted to initially assess the feasibility of the recommended changes.

STRUCTURAL ANALYSIS OF TREE-A

The literature provides indications that hierarchical arrangements play a significant role in human information organization, from the level of immediate perceptual processing (Restle, 1970; Vitz and Todd, 1969) to higher level knowledge structure (Frederickson, 1975). A number of memory information models have been explicitly hierarchical (Collins and Quillian, 1969; Glass and Holyoak, 1974/75; Rosch, 1973, 1975, and 1978; Mandler, 1970 and 1975). Collins and Quillian's model of semantic memory takes on characteristics of taxonomies that are structurally similar to the telidon tree. Glass and Holyoak's ordered search model and Rosch's studies of the internal structure of memory provide additional evidence to support a tree-structure retrieval system. Recently, a number of empirical findings have, however, questioned the efficiency of hierarchical systems (Broadbent, Cooper and Broadbent, 1978; Mathews, Schoenfelt and Valentine, 1982; Mathews, Lee and Coursey, 1981). In light of these counter-arguments, the efficiency of the telidon tree structure was theoretically evaluated. This section provides an efficiency analysis of tree-structure retrieval based on structural considerations alone. In this analysis, human behaviorial aspects were not taken into consideration. The simplifying assumption was made that human retrieval errors would never occur. This is true only if users have perfect knowledge about the content of a tree's categories. As this assumption is not likely to be met under all circumstances, behavioral aspects are considered in the next section.

The efficiency analysis performed here is based on an existing search time model (MacGregor, 1981). An abstract of this model is attached as Appendix II. The model examines the mathematical relationships between search time, database size and the number of vertical divisions in an hierarchical structure. The relationship is described as follows:

$$T_i = i \left(\frac{n^{1/i} + 1}{2} \right)$$

where T_i is the mean search time required in a system with i levels and n elements in the database.

Tree-A, which is used throughout this study, has about 1200 elements in the base and 9 levels in the structure. It is possible to organize a database of this size into tree-structures having different numbers of levels. Keeping the size of the database constant and varying the number of levels, the above model was applied to determine the theoretical consequences on search time.

Database size (n) = 1200

Number of levels (i)	Relative search time (T_i)
1	600.5
2	35.6
3	17.4
4	13.8
5	12.9
6	12.8
7	13.1
8	13.7
9	14.4

Table I: Relative search times for trees with different numbers of levels (for constant database size).

The above table indicates that search time is minimal when there are six levels, for this size of database. Since Tree-A has 9 levels, it appears to deviate markedly from the theoretically optimal structure. However, a comparison of the search time required (14.4-12.8) indicates that the effect of this deviation is likely to be small. Tree-A approximates the most efficient structure.

Another method for analyzing structural efficiencey involves determining whether the sizes of categories in Tree-A correspond to optimal category size. The procedure for this analysis was complicated by the fact that one of the model's assumptions was not met by Tree-A. The model assumes that all elements are stored at the bottom level of the tree, but Tree-A has branches discontinued at different levels of the hierarchy. Owing to this irregularity of structure, there exist different optimal sizes for categories at different levels of the tree. In order to take into account these irregularities of structure, backward analysis procedures were introduced to modify MacGregor's model:

Let A_i be the number of branches discontinued at level i

Let N_i be the number of branches that exist at level i given the optimal category size for level i + 1 is known

Let C_i be the optimal category size at level i

Let m be the number of levels in the structure

At level m, $$N_m = A_m \qquad (1)$$

The optimal category size at level m can be calculated by substituting (1) into the equation developed in MacGregor's model

ie. $$C_m = N_m^{1/m} \qquad (2)$$

Given that C_m is known, at level (m-1), $N_{(m-1)}$ can be computed

$$N_{(m-1)} = A_{(m-1)} + \frac{N_m}{C_m} \qquad (3)$$

If $N_{(m-1)}$ is known, $C_{(m-1)}$ can also be computed

$$C_{(m-1)} = N_{(m-1)}^{1/m-1} \qquad (4)$$

Following these procedures the optimal category sizes that different levels were identified. These optimal values were compared with the respective average category sizes in Tree-A. The comparison is presented in Fig 1.

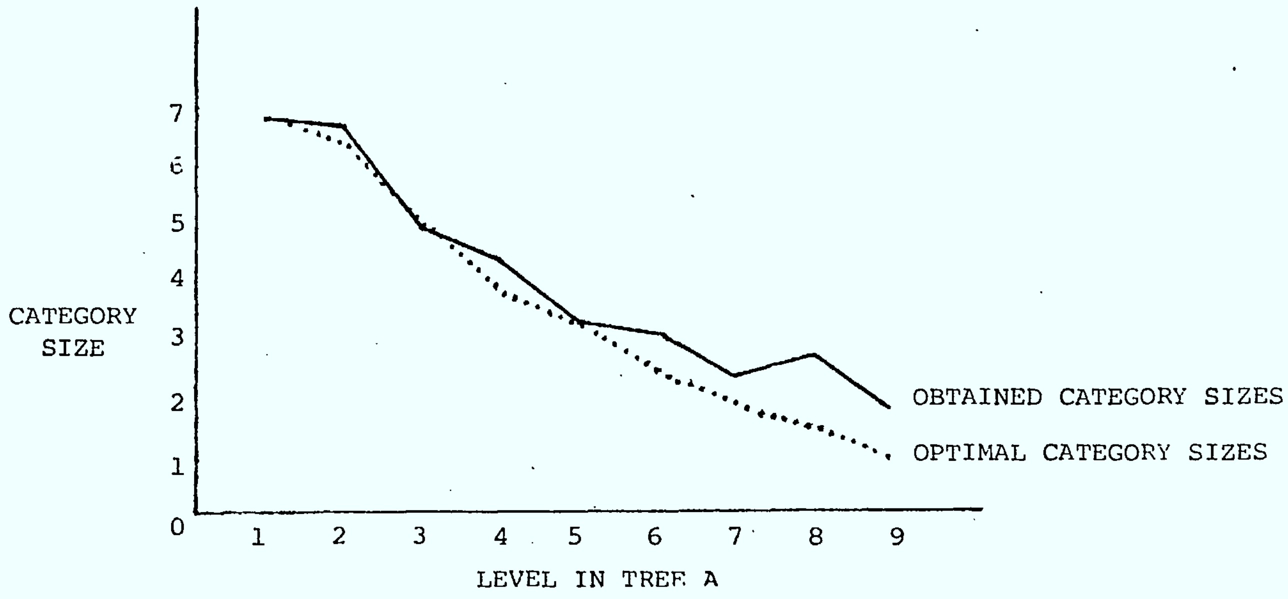

Fig. I: Comparison between theoretically optimal category sizes and the obtained category sizes for Tree-A at different levels in the hierarchy

It can be seen by inspection that the obtained and theoretical values are highly similar. The result indicates that in terms of its structural characteristics at least, Tree-A approaches the optimal form of organization.

LITERATURE REVIEW

Introduction

The structural model of information retrieval described in the previous section demonstrated that tree structures in general can provide the most efficient form of retrieval. An analysis of a particular tree (Tree-A) indicated that, in terms of its structural parameters at least, the tree closely approximated the theoretically optimal form. However, it is important to bear in mind that the model is based on the assumption of perfect knowledge of category content for a given category label. Difficulties with tree-searches may arise because of uncertainty about the contents of categories given only the category labels (Latrémouille and Lee, 1981; McEwen, 1981). The implications are that the search characteristics of users and the content properties of the category system are more likely to provide a basis for system improvement, than the structural characteristics of the tree.

The decision process followed by a user in selecting a category may be divided into three stages:

(1) Category membership verification: This involves an attempt to select from category labels those categories perceived to lead to the required information.

(2) Relationship Comparison: In cases where an item is perceived to be potentially a member of more than one category an evaluation takes place of the relative "strengths" of the possible

category/member relationships.

(3) Decision: A decision process takes place on the basis of (1) and (2).

In performing these three stages, the system may provide a telidon user with only category labels. Since the processes of verification and comparison require more information than this, the additional information must be provided by the user. A human memory search is involved which examines the possible linkages between the required information and the given category labels. When the human memory search selects category-member linkages which do not correspond to those of the system, errors will arise.

The literature review, therefore, examined information on human memory search and retrieval, on category membership identification, and on category differentiation. The review is organized in terms of (a) a description of the major models in the field, and of related experimental findings; (b) a discussion of the implications of these for system improvements; and (c) a set of recommendations.

Major Models

Spreading-Activation Theory - The theory was first introduced by

Quillian (1968) and later expanded by Collins and Loftus (1975). In Quillian's model, concepts are considered to be interrelated within a network (see Figure 2).[1] The links connecting concepts together are referred to as "pointers" which move in both directions between concepts.[2] Links are established through common properties of the linked concepts.[3] The strengths of associate links are represented by a numerical indicator called "criteriality". A single link may have more than one criteriality. In a connection between "bus" and "vehicle" for example, the criteriality for the link from "bus" to "vehicle" would have a higher value than the link from "vehicle" to "bus". The reason for the difference is that while a "bus" is a "vehicle", not all "vehicles" are "buses".

Quillian proposes five different types of links: (a) superordinate and subordinate, (b) modifier, (c) disjunctive sets, (d) conjunctive sets, and (e) a residual class of links. Superordinate and subordinate links are the critical determinants in a category-member verification task. In performing a verification task, the Quillian model proposes that the search process commences by "spreading" in all direction through the links that the two words have with neighbouring concepts, and continuing to more distant relationships. Each concept activated by the search receives an "activation tag" indicating the starting concept and the immediate predecessor. The search process ceases when a connection between

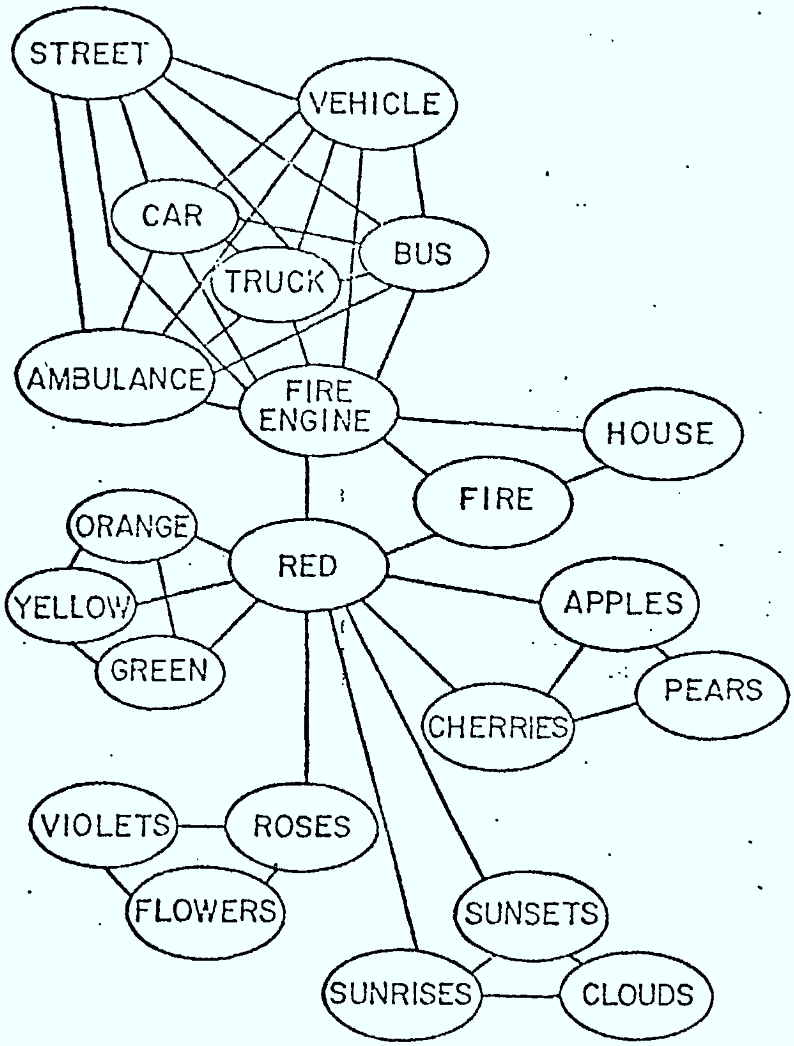

Figure 2. A schematic representation of concept relatedness in a stereotypical fragment of human memory-where a shorter line represents greater relatedness. (adapted from Collins and Loftus, 1975)

the tags from the two starting concepts is discovered. The path leading to this intersection is re-structured by following the tags back to the starting concepts and is then evaluated to determine whether it satisfies constraints imposed by syntax and context.[4] Quillian suggests that the search process can involve more than two concepts. This means that more than one category-member relationship can be verified at the same time. The amount of activation allocated to each connection, however, would be reduced.[5]

Introduction of this model in 1968 gave rise to both complementary and rival models during the following decade. The former are best characterized by Glass and Holyoak's ordered search model (1974/75), the latter by Smith, Shoben and Rips' (1974) feature comparison model (1975). Collins and Loftus (1975) argue that the rival models actually misinterpret Quillian's model. As previously mentioned, they expanded Quillian's model to encompass strategies suggested in more recent studies. The result is a complicated model with sufficient generality to account for many experimental paradigms.

Ordered Search Model - The model was proposed by Glass and Holyoak (1974/75) and suggests that a concept is represented in memory by a single semantic marker.[6] It assumes that semantic markers are interrelated. Hence, parallel to a concept network exists also a marker network. (see Figure 3). When verification of a category-member relationship is

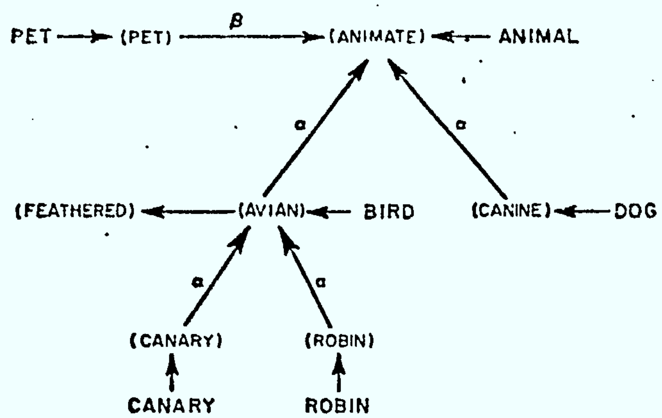

Figure 3. A representative portion of a semantic marker network, showing word to marker and marker to marker links. (Adapted from Glass and Holyoak, 1974/75)

required, the markers representing the concepts are first identified. Next, an ordered search occurs within the marker network. The search ceases upon discovery of information which logically confirms or disconfirms the relationship. Confirmation occures when a member's marker dominates the category's marker.

The model has significance in that the marker network indicates an explicit representation of class inclusion relationships among concepts. Information that represents a category-member relationship is, hence, directly stored in the marker network. For example, the fact that "dog" is a subset of "animal" is directly indicated by the connection of "canine" to "animate". (see figure 3). Subordinate markers of the same class are not necessarily connected. This allows an explicit indication of the distinction between dissimilar category members. The fact, for instance, that "canine" and "avian" are not connected indicates that "dog" is not the same as "bird".

Glass and Holyoak supported their theory with findings about a relationship between "verification reaction time" and "production norms". These observations indicated that longer reaction times were required as the number of levels between concepts increased. The reaction time for verifiying "All robins are birds", for example, was longer than the time required in verifying that "All robins are animals".(7) The implication is that difficulties experienced in a category-member verification process are dependent on the vertical distance between the category names and the required information in the hierarchy. Such a proposition is also implied in Quillian's spreading-activation model.

Feature Comparison Model - Smith, Shoben, and Rips (1974) take a different approach in explaining memory search behavior. [8] They propose that a concept is represented by a set of semantic features. These features may vary along a continuum from "essential" or "defining" to "accidental" or "characteristic". "Defining" features are those possessed by all members of a category, "characteristic" features by only some of the members. Characteristics are therefore features which distinguish members in a category from one another. These features also indicate the member's "typicality" or semantic relatedness to the category name. The model assumes that category-member relationships are verified through a two-stage process. The first stage involves a comparison of the defining and characteristic features of the two concepts. This yields an estimate of the relatedness of the two concepts reflected by the proportion of features they share. If the relatedness of the two concepts is higher than a pre-determined upper level criterion, the search process will cease and provide a "positive" response. When the relatedness falls below a lower-level criterion, a "negative" response will be given. If the estimate falls between the upper and lower-level thresholds, a second comparison stage is initiated where only defining features are considered.

A "positive" response results if the defining features of the two concepts are identical.

The feature comparison model is supported by finding that sentences such as "All robins are birds" are verified more rapidly than sentences such as "All penguins are birds". The finding suggests that robins have certain properties which make them more readily identifiable than penquins as members of the category "birds". The notion of characteristic features, in this instance, provides an explanation.[9]

Property Comparison Model - The model was proposed by McCloskey and Glucksberg (1979), and is also based on the assumption that concepts may be represented by a set of properties. Each property consists of an attribute (e.g. size, colour, shape) and a set of one or more values for the attribute (e.g. red, yellow, green for the attribute colour). Among the "exemplars" (ie. category members) of a concept exist typical and atypical types. Typical exemplars are those most representative of the concept whereas atypical exemplars are the least representative ones. The set of values for a concept attribute includes only those shared by the concept's typical exemplars. The colour attribute for the concept "fruit", for example, may have values red (e.g. cherries, strawberries, apples), green (e.g. pears, melons), and so on. The attribute value of a particular exemplar may therefore differ from those of the concept.

The attribute value "yellow" for bananas, for example, is inconsistent with the attribute values for fruit mentioned above. This does not mean, however, that bananas would not be considered as fruit since other attributes are also compared. The property comparison, hence, rejects the distinction between defining and characteristic features and thereby avoids the inability to specify the defining features of concepts (cf. Collins and Loftus 1975, Wittgenstein 1953). In addition, the assumption upon which it is based allows us to account for the differences among category members in terms of their typicality. Typical exemplars are those with many attribute values identical to the values of the category concept.

When a category-members relationship is presented for verification, McCloskey and Glucksberg assume that category and member properties are identified and compared. Each comparison yields either positive evidence (that the relationship is valid) or negative evidence (that the relationship is invalid). Such positive and negative evidence accumulates as the comparison process continues. The process ceases when the cumulated evidence has exceeded a decision criterion.

In the comparison process, positive evidence is produced when one or more category attribute values correspond with the member's attribute values. If the values do not correspond, negative evidence is said to have been found.(10) This verification task, hence, involves a process of probability estimation. The probability estimates are made according to Bayes' Theorem (Collins and Loftus 1975; Fitt, 1966). The probability that a relationship is valid, given the available evidence, is denoted by the equation:

$$P\ (T/E) = \frac{P(E/T)}{P\ (E/T) + P\ (E/F)}$$

where P(E/T) is the probability of obtaining the available evidence given the relationship is valid
and P(E/F) is the probability of obtaining the evidence given the relationship is invalid.

The significance behind McCloskey and Glucksberg's comparison model rests on this assumption of probability estimation. It implies that verification error will likely be made if the comparison process involves an atypical member of the category. It also indicates that category membership is a matter of degree rather than all or none. [11]

Related Studies - A study by Anisfield and Knapp (1968) provides an explanation for the occurrence of verification error. Subjects in this study were presented with a list of words appearing in sequential order. At any particular point in time, the subject was exposed to only one word and asked to determine whether the word had appeared before. It was found that subjects made errors when the word provided was semantically related to one or more words displayed previously. [12] Based on this finding, Anisfield and Knapp claimed that only a subset of the word's features is examined in the verification process. False recognition is attributed to the possibility that the set of features selected may correspond to those of the preceding words. This hypothesis is supported by Tversky's studies (Tversky and Kahneman 1973, 1981, Kahneman and Tversky, 1973) which discovered that only available heuristics (i.e. relevant issues that come to mind) are considered in a decision process.

A study by Rips (1975) provides evidence concerning judgemental aspects in category membership generalization. In this study, subjects were informed that an unknown property (or disease) was possessed by a species of animal. They were requested to predict other animals possessing the same property. It was discovered that subjects were inclined to generalize the property to other species with similar features.

Implications of the models

As previously mentioned, a user's memory search process in telidon-type situations may involve:

(1) A verification process, where users verify the category-member relationship between a concept and a category name.

(2) A comparison process, where alternative category-member relationships are evaluated.

(3) A decision process, involving the selection of a category, given (1) and (2).

The literature review, whether directly or indirectly, addressed all of these areas. Based on the information reviewed, approaches for improving the system will be identified.

(1) Verification process--The network models (Quillian, 1969; Collins and Loftus, 1975; Glass and Holyoak, 1974-1975) suggest that the

verification task is completed by searching for connections within a network of concepts (marker network: see Glass and Holyoak, 1974/75). Rival models propose that relations among concepts are verified through a feature comparison process. Collins and Loftus argue that the feature comparison aspect is also included in Quillian's network model. Since associative links are represented by the concepts' criterial properties, they claim that the parallel search along the relational links inevitably leads to a feature comparison process. Taking this view, it may be assumed that the verification task is guided by a network search process. Quillian (1969) and Collins and Loftus (1975) have indicated that verification time is directly related to the distance separating the concepts in the network. Glass and Holyoak observed that sentences such as "All robins are animals" require a longer verification time than sentences such as "All robins are birds". This appears to be an indication of the relationship between distance and verification time. Based on this, we might expect that users will experience difficulty in the verification process if the required information is located many levels below the root frame (first level category). Consequently, one would also expect a direct relationship between error rate and the distance between concepts. A relationship of this kind is consistent with experimental findings when subjects search through tree-structures (Whalen and Latrémouille, 1981). The pilot studies reported later in the present paper found a strong positive relationship between the probability of error at a node and the number of levels between that node and the goal ($\underline{r}$ = .94, p < .001).

The models and the experimental findings therefore agree that the verification of a category-member relationship appears to be influenced by the number of intervening nodes in a tree or a network which are required to establish that relationship. One recommendation for improving system effectiveness would therefore be to seek ways of shortening the path between the level at which search commences and the goal nodes.

(2) Comparison process- All the literature reviewed leads to the conclusion that concept properties (or features) are used as the means for verifying a relation. It is therefore the consensus that the process of comparing relationships and of establishing category-member relationships involves some kind of property comparison. Given this, users would be expected to experience difficulty in a comparison task when categories have only a few shared features with the required information, and hence little basis for comparison. In Smith et al's study (1975), findings on the relationship between reaction time and number of defining features substantiated this view. Since general categories (i.e. those in the upper levels of the tree) usually contain only a small number of features shared with their category members, the critical determinant for the success of a category selection depends on the extent to which those general category names elicit "defining" features as opposed to merely "characteristic" features, to use the terms of Smith et al (1975). McCloskey and Glucksberg (1979), in their probabilistic search model, argue that defining features of categories do not always exist. Studies of false recognition and concept relatedness (Rips, Shoben, and Smith , 1975, Rosch, 1973, Smith, 1967, Wilkins, 1971), provide evidence

supporting this view. All of these studies conclude that category names are not necessarily well-defined. Furthermore, the assumptions put forth by Quillian in his spreading-activation model suggest that a smaller amount of activation would be allocated to each concept if the number of the concepts to be verified increases. This indicates that verification difficulty will increase when the number of categories increases. Based on the models and experimental findings reviewed, it appears, then, that:

(a) A comparison task becomes difficult when one or more elements in the comparison process is a general category.

(b) The difficulties lie in the fact that only a few shared features are identifiable.

(c) Errors are more likely to occur in this kind of comparison since categories are not well defined. Features representing one relationship, for example, may be identical to those representing another.

(d) Since only a subset of features is selected, it is possible that a perfect overlap of features from different categories may occur.

(e) As the amount of activation for each concept is reduced when comparing many relationships, the probability of selection error is further increased.

(3) Decision Process- Glass and Holyoak (1974/75) suggest that a category-member relationship is verified through a logical connection of defining features. Quillian (1969), and Collins and Loftus (1975) assert that the verification procedure is based on an evaluation process. Smith et al (1975) claim that upper and lower criteria are involved. Studies by Anisfield and Knapp, Rips, and Tversky et al provide indications

that the process involves also some forms of judgement. In one of the experiments reported later in this paper, evidence supporting a judgemental process was found. In the experiment, it was found that 54.2 % of the subjects repeated an incorrect choice at least once in the search process. Such confidence in the incorrect choice implies the existence of some forms of subjective judgement. It is reasonable to assume that the extent to which subjective judgement is used is directly related to the amount of information available. The kind of information required for this decision task is related to the concept's specific features. Since the number of specific features is in turn, related to the hierarchical level in which the categories are found, we find a further reason to recommend that system effectiveness may be increased by establishing ways to bypass the upper level categories in the tree.

Recommendations for System Improvement

The user of a tree-method of information retrieval has to make a series of successful matches between superordinate and subordinate categories which extend in a "path" from the root frame to the goal node of the tree. The literature review suggested the main ways in which the probability of a successful overall match might be increased:

(1) By decreasing the number of levels in the path.

(2) By avoiding if possible the upper more general category levels.

The number of levels in a path could be decreased in at least two ways; by decreasing the total number of levels in the tree or, via the second recommendation, by bypassing the upper levels in the tree.

Decreasing the total number of levels in a tree would require either increasing the number of categories, or increasing the number of items in the categories, or both (assuming no changes in the size of the database). The structural model described earlier would argue against any large scale alterations of the structural parameters of Tree-A. In addition, empirical findings provide evidence that search difficulty increases with increases in category size (Landauer and Myer, 1972; Wilkins,1971). A better solution may therefore be to leave the general structure of the tree unchanged and to provide users with options for bypassing the upper levels. This would achieve both recommended changes in one step.

Since the purpose of avoiding high level categories is to avoid decisions associated with high degrees of uncertainty, the problem then arises of what level in the tree entry should be made to avoid high degrees of uncertainty? Decisions concerning this may be facilitated by consideration of theoretical and empirical investigations related to the concept of "basic level" categories (Posner, 1969; Rosch, 1973, 1974, 1975; Rosch and Lloyd, 1978; Rosch and Mervis,1975).

"Basic level categories" are supposed to exist at characteristic levels in the vertical dimension of a hierarchy of concepts. Theoretically, basic level categories are at the conceptual level most frequently used when members of a population or subpopulation refer to concepts. Examples of concepts with basic level characteristics are "chair", "car", and "school", which are used more frequently than either the corresponding superordinate categories, "furniture", "vehicle", and "educational institution", or the subordinate categories, "rocker", "sports coupe" and "grade school".

A basic level category is at that level of abstraction at which "cue validity" is maximized (Rosch, 1978). "Cue validity" is a measure of the total number of attributes which are associated with a particular category and <u>not</u> associated with other categories. A category with higher cue validity will therefore be better differentiated from other categories than a category with lower cue validity. Superordinate categories tend to have low cue validity because they have few common attributes. (eg. members of the class "furniture" may share relatively few common attributes) Subordinate categories tend to have low cue validity because, while their members may share many common attributes, they also tend to share those attributes with contrasting subordinate categories (eg. "kitchen chair" shares most of its attributes with other kinds of chair).

A "prototype" is considered to be that category member which is most representative of the category <u>or</u> the superordinate category of which it is a member. "Gun", for example, is more prototypical than "hand grenade" of the category "weapon" (Rosch and Mervis, 1975). "Prototypicality" is also considered to be a function of cue validity. The most prototypical member of a category is the member with most attributes in common with other members of the category and least attributes in common with other categories. A prototype is therefore better differentiated than other category members, and the prototype of a basic level category is a well differentiated member of a well differentiated category.

To the extent that these ideas are applicable to tree-structures we would expect that somewhere along a path from the root frame to a goal node there is a page which is, relative to other pages on that path, at the basic level of categorization. This page will be better differentiated from other pages than either the immediately preceding, superordinate, page or the immediately succeeding, subordinate, page. Associated with this "basic level" page will be a category label which is more prototypical of the category represented by the page than other labels. If this is correct, then entering the tree at a basic level category, using the prototype as a keyword, may be the most effective means of bypassing the early pages and decreasing the overall length of the search path. There are several reasons why the basic level may represnt the most effective point of entry.

First, if basic level categories are better differentiated than categories at other levels, then there should be less ambiguity associated with them. Entry at this level is therefore less likely to result in a user entering on the wrong path.

Second, if basic level categories represent the conceptual level at which population members most frequently communicate concepts, then they are likely to be associated with higher levels of inter-subject agreement than categories at other levels. In other words, we would expect a greater consensus within the population about category membership at the basic level than at any other level.

Third, for similar reasons, the prototypes of basic level categories should be the most frequently used exemplars of those categories within a population. If this is true, then keywords based on basic level protypes should be the least ambiguous labels for the least ambiguous categories in the tree.

In summary, the literature review led to the following recommendations and methods of implementation:

(1) Retain tree-structures in their present general form, since they are the most efficient means of information retrieval, so long as ambiguities about category membership do not exceed certain (at present, unknown) tolerances.

(2) Decrease the number of levels between the node of entry and the goal node by bypassing ambiguous higher level categories whenever possible.

(3) Provide options for entering the tree at lower levels which have the characteristics of basic level categories relative to other levels on that path.

(4) Use keywords based on the prototypes of basic level categories to effect entry.

In order to test the feasibility of these recommendations, two pilot studies were carried out. Details of these experiments are presented in the following section.

INTRODUCTION TO EXPERIMENTS

Two experiments were conducted to initially test the feasiblility of the recommendations. The aim of the experiments was to investigate how a hybrid keyword/tree system might be developed using basic level prototypes as keywords. An immediate difficulty lay in identifying basic level prototypes, and the methods used here were based on the assumption that members of the population will spontaneously generate prototypes of basic level categories under certain conditions.

Theoretically, population members will not only use basic level prototypes themselves in communicating concepts, they will also expect other population members to use them. There are reasons to anticipate that this latter characteristic may provide the best route for accessing prototypes from population members. Single individuals may not be, or may not feel themselves to be, representative of the population. If asked what labels they would apply to categories they might readily generate idiosyncratic responses. However, they could hardly expect the population at large to be misrepresentative. Asking what they expect a "normal person" to do might therefore bypass individual deviations and tap into perceived population stereotypes. To the extent that these perceived stereotypes are shared among population members, we would expect to find high levels of agreement between subjects.

A second problem is in identifying suitable contexts in which to ask subjects to generate terms. We adopted two different approaches here. The first involved providing subjects with specific questions and asking them to generate category labels that others would use in accessing information to answer the questions. Theoretically, the subjects should generate basic level categories--the most distinctive or salient categories--and provide prototypes--the most distinctive category members--as referrents to the category.

The second approach adopted here was based on some additional theoretical considerations. Rosch (1978) and Rosch and Lloyd (1975) have argued that basic level categories are frequently concrete objects. This is because concrete objects tend to have high cue validity. Information in databases may frequently be of a relatively abstract kind, and consequently may not have a well-defined basic level of categorization. However, if different kinds of information tend to be strongly associated with different concrete objects, those concrete associates may then provide prototypical terms which can be used to access the information. The problem then would be to identify a class of common objects that members of the population will frequently associate with types of information. For these associations to be widely held within the population, the objects would have to be relatively permanent, salient and commonly experienced aspects of the environment.

A class of objects which may have these properties is that of "places". Restaurants, universities, garages, farms, machine-shops, hospitals and so on are all concrete aspects of the environment which may be perceived as centres of "local expertise" in certain matters. It seemed to us possible, therefore, that population members may readily associate different kinds of information with different kinds of places in the environment. Such place-names might then provide a set of keywords which have the properties of prototypes.

Two pilot experiments were conducted to test these methods of generating potential keywords. Details of the experiments are presented below.

Pilot Studies

Two pilot studies were carried out to obtain preliminary information on the degree of between-subject consistency in generating category labels. The two procedures are described in detail below. In both cases subjects were given the same 30 items which had been randomly selected from the final nodes of a tree-structure (Tree-A).

In the first method, subjects were asked what index terms people would check under in a directory if they were searching for information about the items. The responses to this procedure will be referred to as "subject-generated index terms", or "subject-terms". In the second method, subjects were asked what kinds of places people would go to in search of such information, if no reference materials were available. Responses to this procedure will be referred to as "locations".

The data from these procedures were analyzed in a number of ways. The subject generated index terms were used in conjunction with a tree directory (Directory-A) which was available. The directory provided an alphabetical listing of category labels ("Directory Index Terms"), together with their associated nodes in the tree. If a match was found between a subject-generated term and a Directory Index Term then the entry node was noted. This made it possible to determine:

(a) Whether a subject-term resulted in an entry into the tree,

(b) Whether an entry was on the correct path to the goal node, and

(c) The level of entry into the tree.

Since no directory was available for the subject-generated location terms, a similar analysis was not possible. Our interest in this case was simply in whether there were higher degrees of between-subject agreement when subjects were asked to generate index terms or when they were asked to generate locations.

EXPERIMENT 1

Method

Subjects (i) *Experimental group:* A total of 15 subjects participated, 9 male and 6 female. Ages ranged from 23 to approximately 45 years.

(ii) *Control group:* Eight new subjects, 5 male and 3 female, participated in a control condition. Ages ranged from 22 to 37 years.

Procedure (i) *Experimental:* Subjects responded individually in a group setting. Subjects were given the same list of 30 questions. The questions were based on randomly selected items from the base level of Tree-A. The items themselves, prefixed by the phrase "Find out about..." constituted the questions. (The 30 items are reproduced as Appendix I).

The subjects were instructed to write what terms they considered a "normal person" would check under in an index to find such information. Subjects were asked to provide three different terms (where possible) for each question. The three responses were written in order of preference.

(ii) *Control:* Control subjects were tested individually in a telidon simulation. (Since no videotext system was available, Tree-A was simulated by mounting pages on cards. The approach was similar to that described by Latremouille and Lee, 1981).

Because of the time-consuming nature of the procedure, control subjects were asked to search for only 15 of the 30 items. The original list was divided into two sets of 15, and four control subjects were allocated to each of the two sets. The order of presentation of items was randomized for each subject.

Control subjects were allowed three trials to find answers to each question. A trial was defined as a sequence of choices directed deeper into the tree. A return to a higher level therefore signalled the start of a new trial.

Results and Discussion

In order to analyze the subject-generated index terms, certain decisions were made concerning the coding of responses. The general rule employed was that if a subject provided a term that appeared in the directory, then a match was considered to have occurred. If the subject-generated term consisted of more than one word, and if one of the words appeared in the Directory, a match was considered to have occurred. If, on the other hand, the subject-term consisted of one word and the Directory Term several words, one of which was the subject term, entry was made at the highest common node of the alternatives given in the Directory. So, for example, if a subject gave the term "Baby", matches would be made with three Directory Terms, "Baby illness", "Baby care" and "Baby furniture". The associated nodes would have been 1342, 136 and 1366 respectively, and entry would have occurred at node 13.

Credit was given for plurals and possessives, so "Babies" was considered to match with "Baby", "Children" with "Children's", and so on. However, no credit was given for synonyms.

These decision rules were used since they were of a kind that could relatively easily be incorporated into an automated keyword system.

Entries were considered to be successful if they occurred at a node which was on the path to the goal node for that particular question. If entry occurred at a node which was not on the correct path, then this was counted as an error. Since subjects gave, in most instances, three subject terms for each question in order of preference, these were considered to represent three separate trials to enter the system. If the preferred subject term matched a Directory Term, an entry was made which was either successful or an error. If no match occurred on the first trial, the second subject term was used. Again, if no match occurred, the third and last subject term was used. Subject terms led, therefore, to either "successful entry", "error", or to "no match found". Errors were relatively rare (6% of all entries) and will not generally be distinguished here from "no match found" responses.

The control subjects were also allowed three attempts to reach a goal node. The first analysis compared the percentage of successes for the two conditions summed over the three trials. The percentage of successful entries for the experimental group was 72.9%, for the control group 66.7%. The difference is relatively small, but significant ($t = 2.87$, $p < .01$). The results suggest that using subject terms leads to a slightly higher success rate than using the tree-structure. It should be noted, however, that the comparison is somewhat biased in favour of the experimental condition. In that condition, a success was recorded for any successful entry into the tree, whereas in the control condition, success required reaching the goal node. An attempt was made to adjust for this bias by applying a similar criterion to the control condition, counting a "success" for reaching the entry node achieved by experimental subjects. However, in several of the questions, different subjects in the experimental group successfully entered at different levels in the tree. In such cases, the highest level was taken as the entry node, and successes were counted for control subjects who reached this node. This measure is slightly biased in favour of the control condition. Using scores adjusted on this basis, the performance of the experimental and control subjects are compared in Table 2 below.

	percentage of successful entries			
	Trial 1	Trial 2	Trial 3	Total
Experimental group	54.0	14.7	4.2	72.9
Control group	52.5	12.5	7.5	72.5

Table : Percentage of successful entries on trials 1, 2, 3, and overall, for the experimental and control conditions.

There were no significant differences between the two conditions, either in overall performance or on performance at each trial. The results indicated that subject terms were, in this case, at <u>least</u> as successful as the tree-method.

The results in Table 2 seem to indicate a marked decrease in success rates with successive trials. This is in part artificial, since the ratios were calculated over a constant total number of responses in order to sum to the correct overall percentage of successes. A more meaningful picture of performance on successive trials is provided by calculating the percentage of successes on Trial 2, given a failure on Trial 1, and the percentage of successes on Trial 3, given failures on Trials 1 _and_ 2. These results are presented as conditional probabilities of success in Table 3.

	Trial 1	Trial 2	Trial 3
Experimental group	.54	.32	.14
Control group	.53	.26	.21

Table 3: Conditional probabilities of success on Trials 2 and 3 given failure on the previous trials for the experimental and control conditions.

The results for the control group indicate that while successive trials do contribute to a final success, they do so with relatively low likelihood. If the costs of successive trials are high in terms of time, frustration, etc., then the extent to which naive users will make repeated attempts is an issue that should probably be investigated. It seems likely, however, that any method that can reduce the costs associated with repeated trials is likely to enhance the overall attractiveness and effectiveness of the system.

The relatively low probabilities of success on trials, subsequent to a failure on the first trial, also suggest that the extent to which subjects learn when using the tree method may be an area of concern, and should probably be investigated.

One reason why later trials contributed relatively little to the overall success may be because subjects frequently made the same mistakes on later trials than they had made on earlier trials. The responses to the six questions with the lowest success rates clearly indicated this (Questions 10, 14, 16, 27, 29, and 30 in Appendix 1). On average, over 60% of the subjects made perseverative errors on these questions. In addition, <u>different</u> subjects tended to make the <u>same</u> errors on these particular questions. The number of subjects making the <u>same</u> wrong choice was 100% on five of the questions and 75% on the sixth.

The nature and extent of these errors are shown in Table 4.

The consistency and persistence of these errors suggest that a change in the organization of the tree to conform to the subjects' behavior may be the most efficient solution to this kind of problem.

QUESTIONS

	10	14	16	27	29	30
Correct Node	Household and Family (13)	Home and Community (14)	Home and Community (14)	Federal Government (173)	Business (3)	Business (3)
Selected Wrong Node	Home and Community (14)	Household and Family (13)	Education (15)	Provincial Government (172)	Table of Contents (1)	Table of Contents (1)
% of Subjects Selecting This Wrong Node at Least Once	75%	100%	100%	100%	100%	100%
% of Subjects Repeating This Wrong Selection On a Subsequent Trial	50%	100%	50%	0%	75%	50%

Table 4: Patterns of within-and-between subject errors on questions with the lowest success rates (control condition only).

The results for the experimental group in Table 3 indicated that for this condition also successive trials contributed to a final success, but again, at relatively low rates. The time costs for repeated attempts however, are markedly lower in this case. The time required to generate and write a subject-term was, on average, 20 seconds. Assuming that inputing a term as a keyword required the same amount of time again, the total time required for a keyword trial would be lower by perhaps a factor of 10, than that required for a trial using the tree method.

Some indication of the relatively greater efficiency of the experimental over the control procedure is provided by the levels of entry into the tree associated with subject-terms.

The mean depth of entry for successful entries was level 3.71. The mean level of the goal node for the questions was 5.53. In a system operating with these characteristics, users would therefore bypass the first 3 levels and enter into the tree structure at a point where they would be required to examine somewhere between 1 and 2 pages in order to reach the goal node.

The standard deviation around this mean level of entry was only .35, which gives some indication of how consistently subjects entered at the third and fourth levels. It is tempting to interpret this as a characteristic of the subjects, but some caution must be exercised here.

In order for an entry to occur, a Directory Term must exist which corresponds to a subject-term. The level of entry is therefore determined by the entry node for the Directory Term, a decision which was made by the originator of the Directory.

We can, of course, consider the author of the Directory as the sixteenth subject in the experimental group--in which case, what the results indicate are that very high levels of agreement exist (across sixteen individuals) about what constitute reasonable index terms for the 30 question items. This is essentially what the 73% successful entry rate reflects.

Nevertheless, there are characteristics of the true subjects' responses which seem to suggest that they prefer index terms which occur around a particular level of abstraction (at least to the extent that we can take levels in the tree as a measure of abstraction).

It will be recalled that, in addition to successful entries, responses could result in errors (wrong entries) or "no match found". For the latter category, there is, of course, no information about a corresponding level in the tree. In order to measure the level of abstraction of these subject-terms, it would be necessary to carry out separate scaling procedures, which was not possible within the terms of the present research.

However, it is possible to ask the following question: Given that some subjects successfully entered at a given level for a particular question, how many other subjects failed to enter or entered wrongly for that question?

An answer to this would indicate how many failures occurred (given that a successful entry was possible), expressed as a function of the level of the successful entry. There is no reason, *a priori*, to expect this to vary as a function of level unless subjects tended to give responses at a particular level of abstraction. However, if subjects consistently generated terms which were fairly concrete, for example, then they would tend to miss those opportunities to enter when the Directory Term happened to be relatively abstract.

Table 5 shows the data relevant to this issue. The Table shows the number of successful entries at each level in the tree, and the total number of failures which occurred on questions entered at those levels.

LEVEL OF ENTRY

	1	2	3	4	5	6	7
# of successful entries	0	31	113	122	54	8	0
# of failures	0	26	41	33	15	7	0
% of successful entries	0	54%	73%	79%	78%	53%	0

Table 5: Success and failure as a function of the level of successful entry.

The results show a consistent pattern. When successful entries occur either very high or very deep in the tree, then the associated questions show relatively high rates of failure. When successful entries occurred around levels 3, 4 and 5, then the associated question had relatively low rates of failure. The results indicate some systematic preference for generating terms associated with particular levels of the tree. This preference was true of both the experimental subjects and of the author of the Directory.

In summary, the results of Experiment 1 indicated that:

(i) A hybrid keyword/tree structure which maps subject-generated terms onto a Directory can be at least as effective as a pure tree-system in terms of probability of success.

(ii) The hybrid system is more efficient in terms of the number of pages that have to be examined (approximately 70% fewer pages).

(iii) The hybrid system is more efficient in terms of the time that is likely to be required to access information.

(iv) There may be problems associated with subjects' learning in a pure tree-structure. In particular, evidence of perseverative errors was found on repeated trials to find the same item. The fact that the same errors occurred between subjects as well as within subjects suggests that changing the tree to conform to subjects' stereotypes may be the most efficient solution.

(v) Index terms generated by subjects and by the author of the Directory showed patterns of preference associated with particular levels in the tree.

A more extended discussion of these findings will be deferred until after a presentation of the remaining findings.

Supplementary Analyses

Subsequent to the analyses described previously, a second directory for Tree-A became available (Index-6: By Keyword). The subject-generated terms from Experiment 1 were used in conjunction with this second directory in a manner similar to that described previously. Results were again analyzed in terms of percentage of successful entries, and level of entry. We also combined the original Directory-A and Index-6. This provided three different methods for using the subject-generated terms to enter the tree: via Directory-A, via Index-6 and via the Directory-A/Index-6 combination. Table 6 shows the percentage of successful entries using the three methods, along with the percentage of successful entries for the control group using the tree-structure.

Retrieval Method

Directory-A	Index-6	Directory-A/ Index-6	Tree Structure
72.9	62.7	81.5	72.5

Table 6: Percentage of successful entries using four different retrieval methods.

The Directory-A/Index-6 combination proved to be significantly better than its closest rival, Directory-A alone ($t=4.30$, $p<.01$).

Both Directory-A and the tree-structure were significantly better than Index-6 alone (_t_=4.58 and 4.39 respectively, $p<.01$).

The difference between Directory-A and Index-6 may be attributable to differences in the number of terms which they used (355 and 251 respectively). In fact, there appears from the present limited results to be a highly consistent relationship between the percentage of successful entries and the number of terms in an index. Table 7 shows the number of terms in the three indexes and the associated percentages of successful entries into the tree.

Index

	Index-6	Directory-A	Directory-A/ Index-6
# of index terms	251	355	446
# of successful entries	62.7	72.9	81.5

Table 7: The number of index terms and the percentage of successful entries for three different index systems.

Three data-points provide a very limited basis for extrapolating a function, but nevertheless the relationship between percentage success and number of index terms is surprisingly well-described by the function.

$$y = 38.54 + .096\,x,$$

where y = percentage of successful entries
and x = number of index terms

The function accounts for 99.99% of the variance among the data points.

Aside from the dangers in generalizing from such a limited sample, it is clear that this function cannot hold for all values in the domain of x, since it predicts a 39% success rate for zero items in an index. One would also expect increasingly diminishing returns as the number of index terms is increased much beyond the levels used here. However, it raises the possibility that some consistent relationship may exist between success rate and index size, at least within certain ranges, which could be useful in providing rules-of-thumb for index construction. (Note, however, that the size of the data-base was constant in this case. Any relationship between success rate and index size is likely to vary with the size of the data-base.)

A more interesting finding for present purposes was that the percentage of successful entries increased to 81.5% when Directory-A and Index-6 were combined. This represented a significantly higher level of performance than the control group, and suggests that, under certain conditions at least, a hybrid keyword/tree-structure method of retrieval can significantly out-perform the tree-method used alone.

Earlier evidence was presented which indicated that subject-generated index terms tended to cluster around the third and fourth levels of Tree-A, and an analysis of the ratios of successful-to-unsuccessful responses for different levels in the tree supported the idea that subjects preferentially generated index terms around levels 3 and 4. Additional evidence for this was obtained from performances with Index-6. The preliminary argument runs as follows.

Questions were randomly selected from the tree. Subjects generated index terms in response to those questions which in turn "sampled" index terms from the Directory. Consequently, if subjects had no systematic preferences for a particular class of index terms, we would expect to obtain in return a random sample of index terms from the Directory. Consequently, we would expect differences between the sample distribution of the Directory terms and the total distribution of Directory terms to lie within normal sampling error. The same agreement applies, of course, to Index-6. This leads to the expectation that there will be no differences between the mean level of entry node for subject-generated terms and the mean level for all terms in a directory. The obtained means and variances for the distributions of entry nodes are shown in Table 8.

	Entry Levels for Subject-Terms Using:		Mean Level for all Terms in:	
	Index-6	Directory-A	Index-6	Directory-A
Mean:	3.56	3.68	4.44	3.89
Variance:	1.48	0.88	2.16	1.28

Table 8: Mean levels of entry contrasted with mean level of all terms in the directories.

The mean level of successful entries into the tree when subject-terms were used to access terms in Index-6 was 3.56. The mean level for all terms in Index-6 was 4.44. The difference between the means was significant (t=7.86, $p<.001$). The result indicates that subject-generated terms addressed significantly higher levels in the tree than would be expected by a random sampling of terms from Index-6.

The results for Directory-A showed the same pattern, though somewhat less strongly. The mean level of subject-generated terms was 3.68, the mean level for all terms in the Directory was 3.89. The difference was again significant (t=2.67, $p<.01$).

The results lend further support to the hypothesis that subjects' preferentially generated index terms associated with particular levels in the tree. If these findings turn out to be generalizable, then it will be important to find methods for describing these preferential levels of abstraction so that they can be used in developing optimal-level keywords for hybrid systems.

An additional finding of interest was that the mean level of items in Directory-A was significantly different from those in Index-6. (t=5.55, $p<.001$). The terms in Directory-A were, on average, closer to the mean level of entry for subject-terms than were the terms in Index-6.

The implication of this is that terms in Directory-A may be at a more appropriate level of abstraction than those in Index-6. Compared to subjects' responses, the terms in Index-6 are, on average, about one level too deep.

In summary, the supplementary analyses in general supported and extended the findings of Experiment 1. Specifically, they suggested that:

1. Hybrid systems can be more effective in successfully accessing information than tree-methods alone;
2. Hybrid systems are likely to be substantially more efficient in terms of access time and the number of pages examined;
3. Further evidence was obtained that subjects preferentially generated index terms corresponding to a narrow range of levels in the tree;
4. Directories of keywords may differentially match this pattern of subject preferences. In particular, Index-6 employs terms that are significantly deeper and further from subjects' preferences than Directory-A;
5. There were tentative indications of a linear relationship between the number of terms in an index and the percentage of successes associated with that index.

EXPERIMENT 2

Introduction

The aim of Experiment 2 was to discover to what extent subjects consistently associate certain information with certain kinds of locations in the real world. If high levels of agreement across subjects exist, then such location names might be useful as keywords.

At the present time no directory of location names exists for accessing the tree, and the evaluation of responses was restricted to measures of between-subject variability. The variability in location responses from Experiment 2 was compared to the variability in index terms responses from Experiment 1.

Method

Subjects

Ten subjects were used, 7 male and 3 female. Their ages ranged from approximately 25 to 45 years.

Procedure

The procedure was identical to that of the first experiment except for the instructions to subjects.

Subjects were given the question list and asked to consider what places a "normal person" might go to find the necessary information to answer each question. Subjects were asked to assume that (i) no reference materials were available, and that (ii) people at those locations would be helpful in answering questions.

Subjects were again requested to give three responses to each question written in order of preference.

Results and Discussion

The initial coding of responses presented some difficulties since similar but non-identical responses were relatively frequent.

Since our interest in this case was not in immediately using the responses as keywords but rather in assessing to what extent user stereotypes might exist, we paid attention to both the form and the meaning of responses. Two responses were considered to be identical if their judged meanings were the same and if at least one of their main terms was common to both. (This meant, for example, that "Federal Government" and "Canadian Government" were judged to be the same. while "Federal Government" and "Provincial Government" were judged to be different). For purposes of comparison, the responses from Experiment 1 were recoded using the same criteria.

A second problem arose in comparing the results from the two experiments because of the differences in sample sizes. To equate the groups, a random sample of 10 subjects' responses was selected from Experiment 1.

The data were at a nominal level of measurement which limited analyses to fairly crude estimates of variability.

The first measure used was the number of <u>different</u> responses necessary in order to include all 10 subjects' most preferred responses for a question. If all 10 subjects agreed, the score would be 1; if all 10 subjects disagreed, the score would be 10, indicating minimum and maximum variability respectively.

For the Experiment 1 responses, the mean score was 4.8, (S.D.=2.1); for Experiment 2, 5.8, (S.D.=1.5). The difference between the means was significant (t=2.09, p<.05).

The results suggest that more keywords may be required when based on locations than when based on index-terms, *if* the criterion is 100% success. This is a demanding level of performance, and the measures used here fail to reflect the fact that in many cases a majority of subjects agreed on a single term. For this reason a second measure was made, based on the degree of consensus for the single most frequent response to each question. The highest number of subjects agreeing on one term was counted for each question. In this case the maximum possible score was 10, representing zero variability, and the minimum was 1, representing maxiumum variability. For the subject-generated index terms the mean value was 6.73, indicating an average level of agreement of 67.3% of the subjects. For the location terms, the mean was 4.93, indicating a 49.3% level of agreement. The difference between the means was significant ($\underline{t}$=3.21, p<.01).

A final comparison was made by examining the number of alternative responses which subjects gave for each question. It will be recalled that in both experiments, subjects were asked to provide three alternative terms for each question, but in many cases, subjects failed to give second or third choices. Since we know from Experiment 1 that successive choices increase the overall probability of success, a failure to attempt alternatives may reduce the potential effectiveness of a keyword method.

The mean number of alternatives given by subjects in the index-term condition was 2.28 out of 3, in the location-term condition, 1.99 out of 3. The difference between the means was significant ($\underline{t}$=4.49, p<.01). Apparently subjects are less willing or less able to provide alternative locations than alternative index-terms, but the present results do not provide any clues as to the dynamics behind this apparent difference. Possibly, index-terms are more cognitively available than locations, in which case, an open-ended multiple response experiment should produce more index-terms than locations. Another possibility is that subjects feel a greater degree of subjective certainty about locations than about index-terms. Having produced what they feel to be a "good" response, they may be less inclined to generate "inferior" alternatives. Some measure of a subjects' confidence in their responses would permit a test of this. In either case, however, unless the first response has a very high objective probability of success, a failure to generate alternatives is a factor which is likely to diminish the overall effectiveness of keyword-based methods.

Summary

If the present results were to be interpreted in terms of which method of generating keywords is the better of the two, then the evidence seems to favour index-terms over locations. There appear to be higher levels of consistency between subjects, and a greater potential for generating alternatives within the individual subject.

It must be emphasized, however, that the present pilot studies were not designed to rigorously test between alternative methods, but rather to explore the potential usefulness of both. A better evaluation given the present information is that both may be promising techniques for obtaining subject-generated keywords. The consensus levels of 67% and 49% would seem to represent fairly high levels of agreement across individuals, and suggest that it may be worthwhile to pursue these techniques under more rigorous and realistic search situations.

Perhaps one of the most important variables, which the present findings suggest should be further investigated, is the subjects' perceptions of the task. What is the search strategy that the subject adopts and how might this influence performance? The reason why this appears to be a potentially powerful factor is because the differences in the present experiment were obtained by very simple changes in instructions which attempted to manipulate the subjects' perceptions of the situation. Whether the manipulation had the intended effect would require further study. But just this kind of experiment might be highly cost-effective, since the experimental costs are low and the potential information-value could be relatively high. No change is made in the structure of the search task itself, only in the users' cognitive representation of the task. If manipulations of this kind can have marked effects on performance and can be systematically induced, then their implications as an economical strategy for system improvement should certainly be further explored.

CONCLUSIONS

The first section of the report described a structural analysis of a tree conducted to discover if changing the structural characteristics alone - the number of levels, the number of categories and the number of elements in categories- was likely to lead to more efficient retrieval. The results of this analysis indicated that for Tree-A at least its general structural properties closely approximate the theorectically optimal form. This indicated that improvements might better be sought in changing either the content of categories and/or the mode of interaction between the user and the tree.

The second section examined literature related to human memory search and retrieval, to category member identification and to category differentiation. As a result, recommendations were made that the number of levels between the node of entry and the base level of a tree should probably be reduced, and that options should be provided for the user to bypass the upper levels in the tree. It was additionally recommended that an attempt should be made to provide a means of entering the tree at the level of basic level categories, and that this might best be done by using the prototypes of basic level categories as keywords. Such keywords would, theoretically, be salient descriptors of salient categories and have a high degree of consensus within a population.

Two pilot experiments were carried out to test the feasibility of these recommendations. The main findings are summarized below.

(1) Subjects generated index terms and location terms with relatively

high levels of between subject agreement. The mean percentage of subjects generating the same terms in a simulated search for the same information were 67% and 49% for index terms and location terms respectively.

(2) Using subject generated index terms as keywords for entering Tree-A led to a significantly higher number of successful entries than a control group using the tree method alone.

(3) It was found that keyword entry was likely to substantially reduce search time for information. The present results indicated that an average of 70% fewer pages would have to be examined using a hybrid method than using the tree method alone.

(4) The most preferred index term of the three which subjects generated, was the most successful of the three terms in entering Tree-A. This indicated that the subjectively "best" term was also the objectively "best" term (as measured by matches with Directory-A and Index-6).

(5) Subjects generated index terms which clustered around the fourth level of Tree-A. One reason why Index-6 may have been less successful than Directory-A for entering subject terms is that the keywords in Index-6 were on average addressed one level deeper in the tree than the subject generated terms.

(6) There were indications of a simple linear relationship between the number of items in an index and the probability of successful entry.

It was stressed previously that these results were obtained from pilot studies and should be replicated if there is any interest in using them as a basis for system improvement. However they do suggest a number of avenues for further investigation. These are discussed briefly below.

The main finding was that a hybrid keyword/tree method can out-perform the tree method used alone, but it is also important to note that the tree method alone can out-perform a hybrid method. The determining factors will probably be

(a) What keywords are included in an index and

(b) How users generate keywords.

In the present experiments the Directory-A/Index-6 combination resulted in significantly more successful entries than the tree method alone, while Index-6 alone resulted in significantly fewer entries. The success of a hybrid method therefore depends, not surprisingly, on characteristics of the index associated with it.

The present results indicated that the following factors should probably be investigated in order to establish guidelines for the development of keyword indexes.

First, there were consistent indications that subjects generated keywords that clustered around a particular level in the tree. Since the theoretical basis of the present experiments would have predicted such a result, it is tempting to interpret this finding in terms of basic level categories. However, further experiments would be required to test whether the subjects' responses had indeed the properties of basic level categories and prototypes. However this consistent preference is probably something which should be pursued, for the following reason. The results showed that subjects preferentially generated index terms at a certain level. Index-6 failed to match these preferences, in as far as its keywords were on average one level deeper in the tree than subjects' terms. A mismatch of this kind could seriously reduce the potential effectiveness of a keyword system. It would be extremely valuable in developing keyword indexes if guidelines could be established concerning the level in a tree which the keywords should address. The literature reviewed, particularly the literature in the area of basic level categories, suggested that ambiguities may arise when category labels are at levels of abstraction which are either too high or too low, and the results of the present experiments were consistent with this. Subjects seem to prefer index terms which are neither too general nor too specific, and it is important that keyword indexes reflect this pattern of preferences.

A second important area of investigation if hybrid systems are to be implemented will be the instructions given to the user. The present experiments attempted to manipulate the subjects' cognitive representations of the search task. For instance, they asked subjects to think in terms of what other "normal" people might do. Is this instruction likely to be critical or not? The instruction also attempted to influence the perceived context of the task. One method tried to induce an image of a search through an index, the other a search through an environment of locations. The former task is essentially sequential, the latter, spatial. Is this a difference that is likely to be important? Subjects generated more alternative index terms than location terms. Is this because index terms are cognitively more available, or because location terms are subjectively more certain? These issues are not trivial, because the success of a keyword method depends on the user's cognitive representation of the task as much as it does on the keyword index itself. It is the interaction of the two systems that will determine the outcome, and this interaction is what should be optimized.

The present finding was in general encouraging with respect to the development of hybrid keyword/tree index methods. Although the findings are tentative, the pattern of results was sufficiently consistent to suggest that a relatively small set of keywords may be sufficient to be useful for large percentages of the user population. If keyword indexes are developed in the right way, the present results give every indication that they can be used to access information more successfully, more quickly and with less effort in terms of the number of pages to be inspected than tree methods used alone.

NOTES

(1) The assumption of interconnected concepts is supported by studies by Fillenbaum (1971), Rosch (1978), Osgood (1957), Garner (1974), and Skinner (1957).

(2) Support for the concept of associative links can be found in the work of Bartlett (1932), Piaget (1950), Newell et al (1958), and Miller et al (1963).

(3) The assumption that associative links are represented by concept properties is consistent with Aristotle's notion of differentia (in Topica) and Otto Selz's conception of attributes (cf. De Groot, 1965). Studies by Kelly (1955), Bruner et al (1956) and Osgood et al (1957) also support this assumption.

(4) Decision rules for this evaluation process are described in Quillian (1969) and Collins and Quillian (1972).

(5) The reduction in activation is explained by assumptions (2) and (3) of Quillian's model (see Collins and Loftus, 1975).

(6) The concept of semantic marker is similar to Katz and Fodor's notion of a "distinquisher" (1963). It is argued that these two concepts are indistinguishable (Bolinger, 1965).

(7) The experimental task used here is called sentence verification. Same-different decision task can also be used (see the study by Schaeffer and Wallace, 1969).

(8) This model is similar to models of recognition memory (Atkinson and Juola, 1974) and models of perceptual judgement (Hawkins and Shipley, 1972)

(9) Smith et al's finding is substantiated by the studies of Rosch (1973, 1974, 1975), Rips et al (1973), Rips (1975), Sanford (1974), Loftus (1974), Wilkins (1971), Egeth (1966) and Posner and Keele (1968).

(10) McCloskey and Gluckskerg specify that this comparison process is directional: category attribute values are compared with the member's values, not vice versa.

(11) McCloskey and Glucksberg's assumption of fuzzy categories is consistent with the findings in the following studies: Labov (1972), Miller and Johnson-Laird (1976), McCloskey and Glucksberg (1978), Rosch (1973), Rosch and Mervis (1975), Wittgenstien (1953).

(12) Anisfield and Knapp's finding is consistent with the proposal that only word meanings are stored in memory (Perfetti et al, 1970, Light et al, 1970).

References

Anisfield, M., and Knapp, M. Association, Synonymity, and directionality in false recognition. *Journal of Experimental Psychology*, 1968, 77, 2, 171-178.

Aristotle. Topica. In W. D. Ross, (Ed.), *The Works of Aristotle* (Vol. 1), Oxford: Claredon Press, 1928.

Askoff, R. L., Cowan, T. A., Davis, P., Elton, M. C. J., Meditz, M. L., and Sachs, W. M. *Designing a National Scientific and Technological Communication system.* Pennsylvania Press, 1976, pp. 5-17.

Atkinson, R. C., and Juola, J. F. Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes, (Eds.), *Contemporary Developments in Mathematical Psychology* (Vol. 1), San Francisco: Freeman, 1974.

Bartlett, F. C. *Remembering, a study in Experimental and Social Psychology.* Cambridge University Press, 1932.

Bolinger, D. The antomization of meaning. *Language*, 1965, 41, 555-573.

Bruner, J. S., Goodnor, J. J., and Austin, C. A. *A Study of Thinking*, New York: John Wiley, 1956.

Collins, A. M., and Loftus, E. F. A spreading-activation theory of semantic processing. *Psychological Review*, 1975, 82, 407-428.

Collins, A. M., and Quillian, M. R. Retrevial time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1968, 8, 240-248.

Collins, A. M., and Quillian, M. R. Facilitating retrieval from semantic memory: the effect of repeating part of an influence. *Acta Psychologica* 1970, 33, 304-314.

De Groot, A. D. *Thought and Choice in Chess*, Mouton and Co., The Hague, 1965.

Egeth, H. E. Parallel versus social processes in multidimensional stimulus discrimination. *Perception and Psycho Physics*, 1966, 2, 245-252.

Fillenbaum, S., and Rapoport, A. *Structure in the Subjective Lexicon*, New York: Academic Press, 1971.

Fisher, S. C. The process of general abstraction and its product, the general analyst. *Psychological Monographs*, 90, 1916.

Fitts, P. M. Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Pscyhology*, 1966, 71, 849-857.

Garner, W. R. *The Processing of Information and Structure*. New York: Halsted Press, 1974.

Glass, A. L., and Holyoak, K. J. Alternative conceptions of semantic memory. *Cognition*, 1974-75, 3, 313-339.

Hawkins, H. L., and Shipley, R. H. Irrelevant information and processing mode in speeded discrimination. *Journal of Experimental Psychology*, 1972, 90, 389-395.

Humphrey, G. Thinking. London: Methnew, 1951.

Kahneman, D., and Tversky, A. On the psychology of prediction. *Psychological Review*, 80, 4, 1973.

Katz, J. J. and Fodor, J. A. The structure of a semantic theory. *Language*, 1963, 39, 170-210.

Kelly. G. *The Psychology of Personal Constructs* (Vol. 1), New York: W. W. Norton, 1955.

Kintsch, W. Semantic memory: a tutorial. *Attention and Performance VIII*, 1978, 595-620.

Kruskal, J. B., and Wish, M. *Multidimensional Scaling*. London: Sage Publications, 1978.

Labov, W. The boundaries of words and their meanings. In C-J. N. Bailey, and R. W. Shuy, (Eds.) *New ways of analyzing variation in English* (Vol. 1). Washington, D.C.: Georgetown University Press, 1972.

Landauer, T. R., and Meyer, D. E. Category size and semantic memory retrieval. *Journal of Verbal Learning and Verbal Behavior*, 1972, 11, 539-549.

Latremouille, S., and Whalen, T. The effectiveness of a tree structure index when the existence of information is uncertain. In *Telidon Behavioral Research 2*. Ottawa: Department of Communication, 1981.

Levitt, T. Marketing Myopia. In H. A. Thompson, (Ed.), *The Great Writings in Marketing*. Plymouth: The Commercial Press, 1976, pp. 36-58.

Light, L. L., and Carter-Sobell, L., Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 1970, 9, 1-11.

Loftus, E. F. Activation of Semantic Memory. *American Journal of Psychology*, 1974, 86, 331-337.

Mandler, G. Memory storage and retrieval: Some limits on the reach of attention and consciousness. *Attention and Performance V*, 1975, 409-516.

McCloskey, M., and Glucksberg, S. Natural categories: well defined or fuzzy sets? *Memory and Cognition*, 1978, 6, 462-472.

McCloskey, M. and Glucksberg, S. Decision processes in verifying category membership statements: implications for model of semantic memory. *Cognitive Psychology*, 1979, 11, 1-37.

McEwen, S. A. An investigation of user search performance on a telidon information retrieval system. In *Telidon Behavioral Research 2*. Ottawa: Department of Communication, 1981.

Miller, G. A., and Chomsky, N. Finite models of language users. In R. D. Luce, R. L. Bush, and E. Galanter, (Eds.), *Handbook of Mathematical Psychology*, (Vol. 11), New York: John Wiley, 1965.

Miller, G. A., and Johnson-Laird, P. N. *Language and Perception*. Cambridge: The Belknapp Press of Harvard University Press, 1976.

Newell, A., Shaw, J. C., and Simon, H. A. Chess playing programs and the problem of complexity. *I.B.M. Journal of Research and Development*, 1958, 2.

Osgood, E. L., Suci, G. J., and Tannenbaum, P. H. *The Measurement of Meaning*. Urbana: University of Illinois Press, 1957.

Perfetti, C. A., and Goodman, D. Semantic constraint on the decoding of ambiguous words. *Journal of Experimental Psychology*, 1970, 86, 420-427.

Piaget, J. *The Psychology of Intelligence* (translated by M. Cook and D. E. Berlyne). London: Routledge and Kegan Paul, 1950

Posner, M. I., and Keele, S. W. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 1968, 77, 353-363.

Quillian, M. R. Semantic Memory. In M. Minsky, (Ed.), *Semantic Information Processing*, MIT Press, 1968.

Rapoport, A., Rapoport, A., Livant, W. P., and Boyd, J. A study of lexical graphs. *Foundation of Language*, 1966, 2, 338-376.

Rips, L. J., Shoben, E. J., and Smith, E. E. Semantic distance and the verification of semantic relation. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 1-20.

Rips, L. J. Inductive judgement about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 1975, 14, 665-681.

Romney, A. K. Shepard, R. N., and Nerlove, S. B. *Multidimensional Scaling, Theory and Application in the Behavioral Science*. New York: Seminar Press, 1972.

Rosch, E. On the internal structure of perceptual and semantic categories. In T. E. Moore, (Ed.), *Cognitive Development and the Acquisation of Language*, 1973.

Rosch, E. Universals and culture specifics in human categorization. In R. Breslin, W. Lonner, and S. Bochner, (Eds.). *Cross-Cultural Perspectives on Learning*. London: Sage Press, 1974.

Rosch, E. Cognitive reference points. *Cognitive Psychology*, 1975, 7, 532-547.

Rosch, E., and Mervis, C. B. Family resemblance: studies in the internal structure of categories. *Cognitive Psychology*, 1975, 7, 573-605.

Rosch, E. Principles of Categorization. In E. Rosch, and B. B. Lloyd (Eds.) *Cognition and Categorization*, 1978, pp.28-49.

Sanford, A. J., and Seymour, P.H.I.C. Semantic distance effects in naming superordinates. *Memory and Cognition*, 1974, 2, 714-720.

Schaeffer, B., and Wallace, R. Semantic similarity and the comparison of word meanings. *Journal of Experimental Psychology*, 1969, 82, 343-346.

Smith, E. E., Shoben, E. J., and Rips, L. J. Structure and process in semantic memory: a featural comparison for semantic decision. *Psychological Review*, 1974, 81, 214-241.

Tversky, A., and Kahneman, D. Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 1973, 5, 207-232.

Tversky, A., and Kahneman, D. The framing of decisions and the psychology of choice. *Science*, 1981, 211, 453-458.

Wilkins, A. T. Conjoint frequency, category size, and categorization time. *Journal of Verbal Learning and Verbal Behavior*, 1971, 10, 382-385.

Wittgenstein, L. *Philosophical Investigation* (translated by G. E. M. Anscombe). Osford: Blackwell, 1953.

Supplementary References:

Broadbent, D. E., Cooper, P.J., and Broadbent, M. H. P. A comparison of hierarchical and matrix retrieval schemes in recall. *Journal of Experimental Psychology:* Human learning and Memory, 1978, 4, 486-497.

Frederickson, C. H. Representing Logical and Semantic Stucture of Knowledge acquired from discourse. *Cognitive Psychology,* 1975, 7, 371-458.

MacGregor, J. A structural model of hierarchical retrieval. Unpublished Manuscript, University of Victoria, 1981.

Mandler, G. Words, lists, and categories: An experimental view of organized memory. In J. L. Cowen (Ed.) *Studies in Thought and Language,* Tucson, Arizona: The University of Arizona Press, 1970, 99-132.

Mathews, R. C., Lee, T. D., and Coursey, B. J. Why hierarchical retrieval schemes do not facilitate recall: a cue overload explanation. *Canadian Journal of Psychology,* 1981, 35, 347-350.

Mathews, R. D., Schoenfelt, E. L., and Valentine, E. The use of Single-- and multi-level organization schemes in recall. *Journal of General Psychology,* in press.

Restle, F. Theory of serial pattern learning: Structural trees. *Psychological Review,* 1970, 77, 481-495.

Vitz, P. C. and Todd, T. C. A coded-element model of the perceptual processing of sequential stimuli. *Psychological Review,* 1969, 76, 433-449.

APPENDIX I

The 30 items selected from the base level of Tree-A:

	Item	Index Page Number
1.	National Weather Report	11221
2.	Indoor Parking Spaces	12632
3.	Customized Vans For Sale	12824
4.	Life Insurance	13122
5.	Roofing Materials	13226
6.	Colds	134312
7.	Kidney Disease	134433
8.	Shopping For Quality	13521
9.	A New Baby	1361
10.	Loudspeakers	137127
11.	How to Plan A Menu	137221
12.	How to Buy a Blender	1375311
13.	Chimney Cleaners	13813
14.	Recipe For Brownies	142111
15.	Collecting China	142522
16.	Art Course In Theory of Design	1426421
17.	Swimming Pools	143221
18.	Afghanistan Restaurant	146111
19.	Pubs And Taverns	14682
20.	Band Concerts	14731
21.	Universities	1514
22.	Employment Opportunities In Graphics	15524
23.	Local Transit Transfers	161122
24.	Air Canada Arrivals	16212
25.	History of Newfoundland	1681111
26.	Travel Routes In Switzerland	168212
27.	Social Credit M.P.s	171114
28.	Sir. John A. Macdonald	17161
29.	Digital Clocks	3521241
30.	Horses For Sale	37321

APPENDIX II

Appendix II presents a summary of a mathematical model for measuring information retrieval time (MacGregor, 1981). The model is based on seven assumptions.

Definitions:

Search time is represented by the mean number of categories and elements observed before the required information is found.

Elements are the data stored in the system.

Assumptions:

(1) Categories at different levels of the hierarchy contain an identical number of subordinate categories or elements.

(2) All elements are stored at the bottom level of the hierarchy.

(3) Categories or elements are retrieved by a random search process without replacement.

(4) Observation time for different categories or elements are the same.

(5) Search within a category ceases when the item subtending the required information is accessed.

(6) Users have perfect knowledge about category content. Given a category name, the user knows what information it subtends.

(7) Search commences from the top level of the hierarchy and ceases only if the required information is successfully retrieved.

Model

Let C be the number of subcategories or elements contained-in a category. The search time within a category is:

$$T_j = \sum_{x=1}^{c} (x/c) \tag{1}$$

where x is the number of subcategories or elements observed before the correct subcategory or element is found

$\frac{1}{c}$ is the probability that a particular subcategory is the correct subcategory.

The equation can be rewritten as

$$T_j = \Sigma\left\{\frac{1}{c} + \frac{2}{c} + \frac{3}{c} + \ldots\ldots\ldots\ldots \frac{(c-1)}{c} + \frac{c}{c}\right\}$$

$$= \Sigma\left\{\left(\frac{(1+}{(c} \frac{c)}{c)}\right) + \left(\frac{2}{c} + \frac{(c-1)}{c}\right) + \ldots\ldots \left(\frac{c/2}{c} + \frac{c/2 + 1}{c}\right)\right\}$$

$$= \frac{c + 1}{2} \tag{2}$$

If a system has i number of levels in its structure, the search time will be:

$$T_j = \sum_{j=1}^{i} T_j \tag{3}$$

Substitute (2) into (3):

$$T_i = \sum_{j=1}^{i} \left\{ \frac{C+1}{2} \right\} \tag{4}$$

Referring to the assumption (1) of equal category size, the equation can be modified as:

$$T_i = i\left(\frac{c+1}{2}\right) \tag{5}$$

Let n_i be the number of elements in the base, given i number of levels.

$$n_i = c_1 \cdot c_2 \cdot c_3 \cdot \ . \ . \ . \cdot c_i$$

$$n_i = c^i$$

$$c = n_i^{\frac{1}{i}} \tag{6}$$

Substitute (6) into (5)

$$T_i = i\left(\frac{n_i^{1/i} + 1}{2}\right) \tag{7}$$

It is possible here to show that search time will be minimal when category sizes are equal. Assume that category size changes at level (i + 1). The search time at (i + 1), hence, is:

$$T_{i+1} = \left(\frac{c_{i+1} + 1}{2}\right) + i\left(\frac{c+1}{2}\right) \tag{8}$$

and,

$$n_{i+1} = c^i \times c_{i+1} \tag{9}$$

or,

$$c_{i+1} = \frac{n_{i+1}}{c^i} \tag{9.1}$$

Substitute (9.1) into (8)

$$T_{i+1} = \left\{ \frac{n_{i+1}/c^i + 1}{2} \right\} + i\left\{ \frac{c+1}{2} \right\}$$

$$= \frac{n_{i+1}/c^i}{2} + \frac{ic}{2} + \frac{i}{2} + \frac{1}{2} \tag{10}$$

To find the optimal value for c, the function T_{i+1} is differentiated with respect to c

$$\frac{d(T_{i+1})}{dc} = -\frac{in_{i+1}/c^{i+1}}{2} + \frac{1}{2}(i) \tag{11}$$

The optimal point appears when $\frac{d(T_{i+1})}{2}$ is equal to zero.

Therefore,

$$-\frac{in_{i+1}/c^{i+1}}{2} + \frac{1}{2}i = 0$$

$$n_{i+1} = c^{i+1} \tag{12}$$

Substitute (9) into (12)

$$c^{i+1} = c^i \cdot c_{i+1}$$

$$c = c_{i+1} \tag{13}$$

Substitute (12) into (8)

$$T_{i+1} = \left\{\frac{c+1}{2}\right\} + i\left\{\frac{c+1}{2}\right\}$$

$$= (i+1)\left\{\frac{c+1}{2}\right\} \qquad (14)$$

Hence, an hierarchical structure is optimally efficient when category size at different levels is the same. The relation $T_i = i\left\{\frac{n^{1/i}+1}{2}\right\}$ is true for any form of hierarchical structure.

DATE DUE
