Hemalata Iyer
State University of New York. School of Information Science and Policy, Albany, NY

Subject Representation and Entropy

Iyer, Hemalata: **Subject representation and entropy.**
Int.Classif. 19(1992)No.1, p.15-18, 14 refs.
The paper examines the systems approach to subject structuring. It presents an overview of empirical studies undertaken to test the postulates relating to subject structuring. Entropy provides a measure of disorganization in a system. Structured subject representations are considered as systems and the measure of entropy is applied to determine the extent of distortion in the communication of the intended messages.
(Author)

1.Introduction

Subject representation is the principal basis on which information systems retrieve information. The subject expounded in the document has to be represented as subject headings, subject index terms, class numbers, data structures and other kinds of surrogates. This is done in order to provide access to information in the information system.

Representation of subjects in the form of class numbers, subject headings, etc., is done by the process of analysis of the subject of the document into its constituent elements and assembling them in a preferred order. This process is equivalent to transforming the n-dimensional configuration of the subject into a linear configuration. It involves the arrangement of the component elements of each subject belonging to a subject field, and all subjects belonging to different subject fields among themselves in a sequence helpful to a majority of users, and requires keeping invariant every immediate neighbourhood relation among all the subjects while transforming or mapping the n-dimensional configuration of subjects into a line (1). Thus, subject indexing systems are primarily concerned with analyzing, identifying, and representing relations between the components of a subject of a document. Such a sequence of component ideas in a subject giving rise to a structured pattern, assists communication, learning and remembering. As Jerome Bruner explains,

"In understanding a complex structure, the human intellect finds it helpful to identify the substructures and categorize them. Such pattern recognition, pattern formulation and categorization have been found to be involved in the human learning process and information handling" (2).

Structuring of ideas is therefore a biological necessity.

Anderson and Bower (3) list the criteria helpful in the choice of a structure for representing information. They are

1. The representation should be capable of expressing any conception which a human can formulate or understand.

2. The representation should allow for a relatively efficient search for and retrieval of information; that is, specific information should remain relatively accessible even when the data file grows to encyclopedic proportions.

3. The representation should saliently exhibit the substantive information extracted from a given input. It should not be influenced by the peculiarities of the particular natural language in which that information was communicated. This hope for language invariance amounts to a wish for a universal interlingua in which any conception in any language should be expressed but for which the format would not be specific to a particular language.

4. For reasons of parsimony, the representation should involve a minimum of formal categories. That is, it should make minimal formal (structural or syntactic) distinctions at the outset; more complex distinctions would be built up by the construction rules for concatenating primitive ideas.

5. The representation must allow for easy expression of concatenation operations, by which "duplex ideas" can be constructed out of "simple ideas". This means, for example, that the representation should allow expression of conceptual hierarchies, or multiply embedded predications, or allow one to predicate a new information structure.

Among other things, the structure of indexing languages must aid communication. The structure should provide cues in a modulated fashion, so that the information seeker moves on gradually towards his area of interest.

2. Ranganathan's Categories and Absolute Syntax

Ranganathan's approach to the structuring of subjects is based on the postulational approach. It centers around the concept of the Basic Subject [BS] and the Five Fundamental Categories [FC]: Personality [P], Matter [M], Energy [E], Space [S], and Time [T] with the sequencing of the categories being PMEST on the principle of decreasing concreteness (4). The 'Basic Subject'

specifies the context of the subject in relation to other subjects in the universe of subjects.

'Personality' is the core entity of a subject statement. Ranganathan considered it as the most ineffable one for definition and suggested the method of residues for its recognition. However, this method was found to be inadequate and as Gopinath writes,

> "The problems of recognition of fundamental categories is not definitional, but contextual. The semantic and syntactic aspects in the formation of these compound subjects and the generalizations of these structures to a nodal base ... that is, the basic subject-sets cause the difficulties in the recognition of Personality" (5).

'Matter' connotes a property or materialness of the focal idea of a subject statement. Later, the material constituent was considered to be the qualifier and only 'Property' is considered to be the fundamental category matter.

'Energy' connotes an action in relation to the focal idea. It could denote action, interaction or mutual action.

'Space' represents geographical areas and physiographic features.

Absolute Syntax: Ranganathan suggested a facet syntax for compound subjects that is free from linguistic and cultural influences. Such a syntax would reflect the arrangement of ideas simulating the mental process of a normal human intellect. This he called absolute syntax (6) and suggested the structure "BS, PMEST" to parallel the absolute syntax. In his investigation A. Neelameghan found parallels in formal linguistics in terms of the deep structrure of a sentence, especially in the research work of Chomsky, Fodor, Fillmore and Katz (see 1, p.170).

3. Categorization

Absolute syntax provides a structure that is predictive. It is based on the categorization of concepts. Eleanor Rosch (7) points out that the formulation of categories is for cognitive economy. Categorization is done in order to provide maximum information with the least cognitive effort. To categorize a stimulus means to consider it for the purposes of categorization in preference to other stimuli. Therefore it results in cognitive economy.

There are three ways of establishing relationships among categories. The first is cause-effect; the second is probabilistic; the third and the most recent, is the systems approach which is concerned with the interaction of the system with its environment. This method of understanding is an analytico-synthetic one. It looks at the overall purposes governing the design and functions of a system in order to explain its behaviour. The systems approach is hierarchic in nature and moves from the particular to the general and also vice versa. Although synthesis cannot be separated from analysis and causality, it is different in its approach. Purpose and its fulfillment are its primary concern. Obviously then, priorities in the fulfillment of its purpose become essential. Thus, the representation of a system according to its purposes, its environmental constraints, its actors, their objectives, the function of the system, and the parts that perform these functions take on a hierarchic form. One would then be concerned with the priority impact of any of these elements on the overriding purpose of the system (8).

4. The Systems Approach to Fundamental Categories

Any system can be looked at in terms of parts and elements. Emery Ackoff (8) very succinctly explains these parts and the hierarchy of the system's structure becomes apparent.

Personality: Personality is how an individual converts the choices in the environment into a situation wherein he derives maximum benefits; that is, given several alternatives, chooses the one from which he expects maximum utility. The degree of expectation of the outcome depends on the available alternatives as well as the time and place. Personality is defined in terms of its unique regular and specified responses to its environment, and these responses involve the properties of an individual. The properties change or are made to change due to external action in terms of space and time; hence the idea of property, action, space and time. The specific connotations of these embedded categories may be delineated further.

Property: A property is the potentiality for producing a specified type of response in a subject in a specifically chosen environment.

Event: An event is a change in one or more structural properties of either an object, a system, an environment, or a relationship between them over a time period of specified duration. An *action* is an active event which is capable of making something else happen to the thing or its environment; that is, action is explained in terms of what it does to the object. Action involves space and time.

Time Slice: A time slice is a bound part (volume) of space at a moment of time. Time is a property of events that is sufficient to enable an individual to individuate any two changes in the same property of some individual. A time slice is explained through events.

Space is inclusive of time, for any object which occupies space exists through time. Thus, there is an inclusive relationship between elements thereby giving rise to a hierarchic structure between them.

5.Empirical Validation of Ranganathan's Postulates

Ranganathan's approach to subject representation is primarily a postulational approach. The succeeding sections of this paper will give an overview of studies undertaken to empirically test the postulates relating to subject structuring. This is important for a further progression of the theory. Besides, the value of such studies

is that they offer methodologies for testing postulates and validating the hitherto abstract ideas as empirically verifiable truth statements.

Concreteness of the facets, which is the central idea governing the ordering of the facets, needs to be determined empirically. The methodology used for the purpose involves the following steps:

1. Analysis of subject statements into their component ideas or facets of Ranganathan, e.g. Subject statement:

Improvement of social status of immigrant children in England in the 1930's.

2. Administering the terms denoting each of the component ideas to users, as single terms, for e.g. "Children", and also in combination with terms representing other component ideas in a stratified manner.

3. The respondents are requested to make a subject statement using a component idea.

4. The statements are facet analyzed.

5. The *Variance Factor* is determined for each of the component categories. This is based on the number of differing roles in which the component category has been used by the respondents while making their subject statements, although there was an implicit intended role, e.g. [P] type idea used in the following roles: [P] as [P]; [Sp] to [P]; [Sp] to [E]. The intended role is [P] of [P].

The variance factor together with the proportion of the reciprocation of the intended role is used to determine the concreteness of a category.

6. *Entropy:* The reciprocation and non-reciprocation of the categories can be interpreted from the information point-of-view. In this context, the variance in reciprocation is taken as a function of the information content. Such variation leads to disorganization. A measure of such disorganization is known as entropy. The entropy formula is:

$$E - (p \log p + q \log q)$$

where E = Entropy
p = Proportion of reciprocated roles from the responses
q = Proportion of unreciprocated roles from the responses

Predictability is a function of the amount of organization in the system, the opposite of entropy. Therefore predictability will be E-1; where E = entropy.

This methodology, when applied in determining the concreteness of categories, resulted in [P] being the most concrete category, followed by [M], [S] and [T]. This is in support of Ranganathan's postulate of Concreteness (9).

Having determined this, the next issue is to empirically test the implications of the postulate of decreasing concreteness. The questions raised in this context are:

- Does subject structuring based on decreasing concreteness of facets minimise entropy; i.e., does it support the communication of the intended meaning of the surrogate to the user? If so, to what extent?

The same methodology involving the variance factor is used for this purpose. The results of the study indicate that the presence of the concrete category [P] increases predictability or decreases entropy. The assembly of categories in the order of decreasing concreteness is the most effective way of ordering facets, resulting in least entropy and in maximum predictability (9).

6. Recall Factor and Concreteness

The human mind conceptualizes through observation and experience, and the human memory requires an organizational structure for assimilation of concepts. Retrievability of concepts depends firstly on the level of processing and secondly on the formation of mental images. Images are easily formed if the concept is a concrete one. From the organized store the concepts are recalled in response to some external stimuli (10, 11).

In the context of information retrieval the recall potential of a term is crucial. Hence, it is important to determine the correlation between concreteness of a category and its recall potential.

The questions that need to be asked are:

- Do concrete concepts, such as [P] type concepts have greater recall potential than the less concrete ones, such as [E] type concepts?

- Do stimuli of concepts of differing concreteness result in varying recall?

- Do stimulus concepts of the "genus" type or concepts representing the "whole" result in recall of the "species" type or "part" type concepts respectively with regard to hierarchically related concepts?

The methodology used is the word-association test. The results indicate that the recall of [P], [M], and [E] type ideas are significantly different. There is a direct correlation between concreteness of a category and its recall potential. However, they are independent of the type of stimuli. Whole-part type of ideas have a greater recall potential than the genus-species type of ideas. As regards the direction of recall of hierarchically related concepts, the movement is from the broader to the narrower concept, i.e. from genus to species or from whole to part (12).

7. Bond-Strength and Categories

Ranganathan looked upon the facet structure as one of decreasing sequence of bond-strengths between basic subjects and successive categories, e.g.,

Facet Structure:

Agriculture [BS], Rice Plant [P]; Disease [M]; Prevention [E]. Madras [S] ' Dry Period [T]

In this facet structure, the bond strength of the concept "Agriculture" is greatest with "Rice Plant". It is less with "Diseases". It is still less with "Prevention". It is still lesser with "Madras". It is least with "Dry Period" (6). Thus there are interesting connections between concreteness and bond strength. Once again an empirical test

of the measure of bond strength between categories is carried out.

Measure of Bond Strength: Bond-strength represents relative contiguity in the association of ideas. One of the approaches of psychologists is to use the reaction time as a measure of associative strength. The reaction time from the onset of the stimulus to the onset of the response is known as the latency factor. This latency factor is used as an indication of bond-strength. A word association test is administered to users, and the bond-strength is determined in terms of inter-facets and level clusters (12, p.58-68).

The results of the study indicate that:

- Bond-strength between [M] and [E] is stronger than between [P] and [E], while there is no difference in the bond strength between [P] and [M], and [P] and [E]. This results in the sequence [P] [M] [E] when ordered in a linear way according to the bond-strength existing between them.

- Inter-facet bond-strength is stronger than the one between level clusters of the same category; i.e., the bond-strength between [P] and [M] is stronger than the one between [P] and [P1]. This is contrary to the Postulate of Level Cluster which states that the occurrence of the same facet on different levels demands a grouping together resulting in the facet structure [P], [P2], [P3]; [M], [M2], [M3]:[E]...

8. Facet Analysis and Search Strategies

What is the relevance of these findings for the process of search and retrieval in online bibliographic databases? An experiment of comparative retrieval strategies was conducted whereby the faceted search model was compared with two other types of searches: Quorum Function Search and Online Boolean Search (13). In the Quorum Search proposed by Cyril Cleverdon, the system looks for items with all desired terms present. If no item is retrieved then one of the terms is dropped and the search is performed again. This process is repeated dropping each term in turn until a match is found. If none occurs, then two terms are dropped, and so on. This is an unstructured search as it involves random dropping of terms. This contrasts very well with the faceted model search whereby the questions are facet analyzed using Ranganathan's theory and the search strategies are developed on that basis. This represents a highly structured search process. The Online Boolean Search represents the database searching as performed by the search intermediaries. User evaluation of relevance of the output is used to compute recall and precision measures.

The retrieval results indicate that the faceted model search perform at a higher level of precision and recall than the other two search models.

Structuring of queries using Ranganathan's theory of classification is helpful in the process of searching and retrieval. It serves the following purposes:

- Assists in the choice of concepts from the users' narrative statements representing their information needs.

- Assists in formulating search statements by providing a basis for the use of appropriate Boolean operators. Terms representing different facets are combined with the operator 'AND' and those within a facet, or representing different levels of the same facet are combined with the operator 'OR'. The rationale for this is based on the degree of bond-strength between facets.

- Provides a method for systematically dropping terms, if the search needs to be broadened. Terms are dropped from the right end of the search statement.

9. Conclusion

The studies presented provide an empirical basis for Ranganathan's facet structure, establishing the measure of concreteness of the categories; correlating concreteness with their recall potential, predictability, and bond-strength and the possibilities of application in online bibliographic searching.

References

(1) Neelameghan, A.: Absolute syntax and structure of an indexing and switching language. In: Ordering Systems for Global Information Networks: Proc.3rd Int.Study Conf.Classif.Res., Bangalore: FID/CR 1979. p.168

(2) Bruner, J.S.: Study of thinking. New York: Wiley 1956.

(3) Anderson, J.R., Bower, G.N.: Human associative memory. Washington: V.H.Winston. (distributed by Halsted Press, New York 1973)

(4) Ranganathan,S.R.: Library classification: Fundamentals and procedure. Bombay: Asia 1944.

(5) Gopinath, M.A.: An analysis of the problems in recognition of the manifestations of the fundamental categories in interdisciplinary subjects. Ph.D.Dissertation. Karnatak University, p.120

(6) Ranganathan, S.R.: Prolegomena to library classification. Bombay: Asia 1967. Chapt.RQ

(7) Rosch, Eleanor: Principles of categorization. In: Cognition and categorization. Hillsdale, N.J.: L.Erlbaum Assoc.1978. p.28-29

(8) Ackoff, E.: On purposeful systems. Chicago: Aldine-Atherton 1972.

(9) Iyer, Hemalata: Facet structure of subjects: An empirical study of concreteness and predictability of categories. Libr.Sci.Slant Doc. 19(1982)Paper L

(10) Lindsay, P.H., Norman, D.A.: Human information processing: Introduction to psychology. New York: Academic Press 1977. p.351

(11) Rumelhart, D.E.: Introduction to information processing. Ohio: Merrill 1981.

(12) Iyer, Hemalata: Structure of indexing languages and retrieval effectiveness. Ph.D.Diss., University of Mysore 1984. p.49-57

(13) Iyer, Hemalata: Online searching: Use of classificatory structures. In: Fugmann, R.(Ed.): Tools for Knowledge Representation and the Human Interface. Proc.1st Int.ISKO Conf., Darmstadt. Frankfurt: INDEKS Verl.1990/91. p.159-167

Dr.Hemalata Iyer, School of Information Science and Policy, SUNY at Albany, 135 WesternAve., Albany, NY 12009, USA.