J of Lib & Inf Sc, Vol 5 No 1 June 1980

A Mathematical Approach to Relations in Thesauri

M S SRIDHAR*

Presents a brief background of set theory and relations including equivalence and ordering relations. Proposes four working hypotheses concerned with thesauri relations (USE, UF, NT, BT, GS & RT) and examines these relations against the characteristics of relations defined in Mathematics. Explores the possibility of generating equivalence class of descriptors from a single descriptor to aid information/document retrieval. Provides a new mathematical approach to thesaurus and its relations and compares it against traditional way of representing thesaurus. Proposes new areas for investigation in this direction.

0 INTRODUCTION

It is quite obvious that thesaurus and classification scheme are essential tools for Information Retrieval. The conceptual analysis and establishing relations such as NL, BT, RT, etc. among concepts in a thesaurus is largely trial and error based. An examination of characteristics of such relations would provide an insight into the structure of thesaurus. Such a probe would also condition the proliferation of relations in any thesaurus. An attempt is made here to examine relations of thesauri in the light of characteristics of relations enumerated in Mathematics. Thus the working hypotheses of this study could be stated as follows :

1 What are the elementary and compound relations found in thesauri?
2 What characteristics of relations as defined in Mathematics are satisfied by thesauri relations?
3 Are there any equivalence relations in thesauri?
4 Can we generate equivalence class of descriptors from a single descriptor?

*Librarian, ISRO Satellite Centre, Bangalore, 560058.

The study is largely confined to the relations and examples of NASA Thesaurus. Eventhough reader is assumed to have preliminary knowledge of set theory and thesaurus, a quick brush up of necessary background of set theory is provided.

1 SET AND RELATIONS

A SET is a collection of well defined distinct objects. These objects are called ELEMENTS of the set. The vowels of the alphabet : a,e,i,o and u is a set. All postable terms in a thesaurus is also a set. Sets are usually denoted by capital letters like A,B,X,Y, etc. and elements by lower case letters like a,b,x,y, etc.*

Just like a collection of elements is called a set, a collection of sets is called a CLASS and a collection of classes is called a FAMILY for convenience.

Often, we notice certain relations among elements of the same set and also among elements of different sets. Since there is not much difference in properties of relations defined on elements of a single set and relations defined on elements belonging to different sets we shall concentrate on relations among the elements of the same set for our study. The relations such as 'father of', 'brother of', 'less than', 'is similar to', etc. which consider only two elements at a time are called BINARY relations and relations like 'is between', which consider three elements in demonstrating the relation are called TERNARY relations. Here we are more concerned with binary relations. However, ternary relations are very useful in fixing relative position of a descriptor in the hierarchical or generic structure of the thesaurus.

BINARY RELATION (or simply RELATION for our purpose) is an association or a rule for association of two elements of a set or different sets.

The relation itself is a set. To illustrate this we shall consider the following example. If M represents set of all men and W, the set of all women and the relationship of marriage (i.e "husband of") is represented by R (please note that there could be relations among elements of the same set),

*For other concepts like universal set, null set, sub set, set operations, etc. please refer books cited in the bibliography.

then R is a set of all couples like (P_1,P_2) such that P_1 belongs to M and P_2 belongs to W. Thus a binary RELATION is a set of all ordered couple elements or ordered pairs. Notationally P_1 RP_2 indicates that P_1 is related to P_2 in the way R is defined (here it is 'husband of'). For our purpose of applying these concepts to thesaurus the notation $P_1 R_{P_2}$ is more appealing than P_1RP_2.

Since R itself is a set, it could have subsets (which are again relations) and all set operations on R and its subsets result in sets which are nothing but relations.

2 EQUIVALENCE RELATION

A relation defined on elements of a set is said to be EQUIVALENCE RELATION if it is REFLEXIVE, SYMMETRIC and TRANSITIVE.

Reflexivity : A relation R defined on a set A is reflexive if aRa is true for any element a of A.

e.g : The relation "is contemporary of" defined on a set of people is a reflexive relation
The relation "is husband of" defined on a set of people is *not* a reflexive relation

Symmetry : A relation R defined on a set A is said to be symmetric if for any two elements a and b of set A the relation aRb is true then bRa must also hold good.

e.g : The relation "is married to" defined on a set of people is a symmetric relation
The relation "is brother of" defined on a set of people is *not* a symmetric relation

Transitivity : If a,b,c are any three elements of a set A on which a relation R is defined and if aRb and bRc imply aRc then the relation is said to be transitive.

e.g : The relation "is brother of" defined on a set of people is a transitive relation
The relation "is a son of" defined on a set of people is *not* a transitive relation

The properties 'Reflexivity', 'Symmetry' and 'Transitivity' of a relation are independent of each other. As against these we have 'Irreflexive', 'Asymmetric' and 'Intransitive' properties i.e. if a relation is not reflexive, it is said to be irreflexive and so on. Further, if the condition for reflexivity holds good for some elements and fails on other elements of the set, then the relation is said to be 'meso reflexive' (i.e. it is neither reflexive nor irreflexive). Similarly we could see 'meso symmetric' and 'meso transitive' properties.

In addition 'symmetry' has one more variety called 'Antisymmetry'. For example relation "is less than or equal to" defined on the set of integer numbers is an antisymmetric relation becuause if two elements a and b in set A satisfy $a \leqslant b$ and $b \leqslant a$ then we conclude $a = b$ or a and b are identical. (Note that the relation "is less than or equal in height to" defined on a set of students in a class is *not* antisymmetric).

Thus we have noted that any of the ten properties and their combinations could exist with relation.

An important result of an equivalence relation on a set is that it decomposes (or partitions) the set into two or more mutually exclusive (disjoint) subsets called EQUIVALENCE CLASSES. Hence if a is an element in set A which is partitioned by relation R then a is in atleast one of the equivalence classes. The equivalence class to which a belongs can be generated by cyclic process by applying relation R to other elments of A in relation to element a. This equivalence class is called 'EQUIVALENCE CLASS GENERATED BY ELEMENT a'.

Four important characteristics of equivalence classes are :

(i) If two elements a and b are in the same equivalence class then aRb is true,

(ii) Every element a in the set A, must belong to atleast one of the equivalence classes,

(iii) Equivalence classes are mutually exclusive,

(iv) If a and b belong to two different equivalence classes then neither aRb nor bRa is true and a and b are called INCOMPARABLE ELEMENTS.

3 ORDERING RELATIONS

Certain special transitive relations having either reflexivity or symmetry

are called ORDERING RELATIONS.

A relation R on a set A that is reflexive and transitive is called a QUASI-ORDERING RELATION.

e.g: The relation "is atleast as old as" defined on a set of people
The relation "implies" defined on a set of statements

In quasi-ordering relation, there may exist two elements a and b in set A for which neither aRb nor bRa holds. It is already said that such elements are called INCOMPARABIE ELEMENTS. However, in some quasi-ordering relations both aRb and bRa are not excluded. In the first example given above there are no incomparable elements (i.e for any two elements a and b we will have either aRb or bRa). Such a relation which has no incomparable elements is described as CONNECTED RELATION. On the other hand the second example with relation "implies" is not connected.

A quasi-ordering relation R on a set A is said to be WEAK-ORDERING RELATION whenever it is connected. Alternatively, weak ordering relation is one that is connected (and hence reflexive) and transitive.

e.g: The relation "is atleast as tall as" defined on a set of people
Ths relation "is less than or equal to" defined on the set of positive integers.

The second example "is less than or equal to" has an additional property of antisymmetry i.e whenever, $a \leqslant b$ and $b \leqslant a$ we conclude $a = b$. Such a quasi-ordering relation which has antisymmetric property is called PARTIAL ORDERING RELATION.

A connected partial ordering relation is called a SIMPLE ORDERING.

eg: The relation "is less than or equal to" defined on the set of positive integers.

We shall notice and confirm here that simple ordering is a compound relation consisting of two relations. The above example "is less than or equal to" consists of "is less than" and "is equal to". We have already noted that relations are sets and any combination of (or any set operations on) them results in a relation.

A connected relation that is asymmetric (and hence necessarily irreflexive) and transitive is called a STRONG (or LINEAR) ORDERING RELATION.

e.g: The relation "is less than" defined on the set of positive integers.
The relation "is earlier than" defined on a set of events.

A strong ordering relation R on a set A is characterised by : For all elements in A,

(i) aRa does not hold good (irreflexive)
(ii) if a is different from b, then aRb or bRa holds good (asymmetric)
(iii) if aRb and bRc are true, then aRc is also true (transitive).

A relation which is irreflexive and transitive is called a PREFERENCE RELATION.

e.g: The relation "is preferred to" defined on a set of alternative actions.

Here aRa is nonsense since we cannot prefer something to itself. Table 1 and 2 provide list of properties of relation and types of ordering relations explained above.

Table 1

Properties of Relation

Reflexive	Irreflexive	Mesoreflexive	—
Symmetric	Asymmetric	Mesosymmetric	Antisymmetric
Transitive	Intransitive	Mesotransitive	—

4 THESAURI

New let us turn our attention to thesauri and examine the relations used in them. A thesaurus, in popular terms, is a controlled vocabulary in which descriptors of a particular field are systematically grouped and presented. Eventhough pure precoordinate and pure postcoordinate vocabu-

Table 2

Ordering Relations

(i.e. Transitive + Reflexive or Symmetric properties)

Reflexive	+	Transitive	=	Quasi-ordering relation
Quasi-ordering relation	+	Connectedness	=	Weak-ordering relation
Quasi-ordering relation	+	Antisymmetric	=	Partial ordering relation
Partial ordering relation	+	Connectedness	=	Simple ordering
Connected relation	+	Asymmetric	=	Strong or Linear ordering relation
Irreflexive	+	Transitive	=	Preference relation

laries do not exist in practice, we shall restrict ourselves to a fairly (relatively) postcoordinate vocabulary since precoordinate vocabulary necessarily injects tree or hierarchy structure and shows more of 'preference relation' than other types of relations.

41 Traditional Approach to Thesauri Relations

Traditionally thesaurus is described as concept mapping in the form of trees or hierarchies or networks of descriptors. Schematically it is represented as arrowgraphs, circular diagrams, or rectangular matrix.

Such presentations lack clear understanding of characteristics of relations in thesaurus. In certain cases combined relations are not distinguished from elementary relations. The nature of all elementary relations are not identified. Quite often relations are not explicitly defined with rules of association. Above all, the model do not provide for generating an equivalence class of descriptors from a given descriptor to aid document/information retrieval. Normally, generation of a set (or class) of descriptors for retrieval purpose is largely descretionary and less consistant.

Thus conceptualising thesaurus as a mathematical model having certain elementary relations among the descriptors enables us to see characteristics of elementary as well as combined relations and their effects on thesaurus. Our hypotheses could be tested and answered by considering thesaurus as a large set of terms (both postable and nonpostable) on which certain relations are defined. Many relations could be defined by way of ascribing rule to associate terms in this set. In defining relations we may encounter a problem of inconsistency due to overlapping of concepts represented by terms. However once relations are fairly defined we can gain better insight into the structure of thesaurus as demonstrated in this study. The present study limits itself to examine the existing relations in thesauri.

Table 3 gives certain frequently used elementary and combined relations and their properties to answer our first two hypotheses in nutshell. Here relation SY is an exception since it is not used in any thesaurus. SY is constructed to demonstrate convenient equivalence classes it generates eventhough it is not useful at retrieval stage. In addition to these relations there could be other relations but most of them are manifestations of these elementary relations. Similar to SY we can construct many combined relations from elementary relations and see how their properties vary from the properties of their constituents.

To answer our third hypothesis, there do not exist any perfect equivalence relation in any thesaurus. However, RT and GS are called quasi-equivalence relations in Table 3 as they do not satisfy transitivity in its real sense. Hence terms in thesaurus are not related in transitive manner as far as RT & GS are concerned. If transitivity is allowed, probably each term in the thesaurus is related to the other term and hence resulting in the entire thesaurus as one equivalence class. Restricting transitivity in RT & GS was necessary in order to make thesaurus manageable. Hence transitivity is satisfied only over few steps and as such relation is called quasi-equivalence relation.

If the number of steps over which transitivity is allowed is more, then the resulting equivalence class of descriptors would be large causing more recall in the retrieval. On the other hand, if the steps are very few the equivalence class of descriptors would be very small resulting in more precision.

Table 3
Thesauri Relations and Their Properties

Sl. No.	Relation: In abbreviation	Relation: In expanded form	Elementary or Combined	Converse of	Reflexivity	Symmetry	Transitivity	Type of relation
1.	USE	Use	Elementary	UF	Irreflexive	Asymmetric	Transitive (in a telescopic way)	Preference relation
2.	UF	Used for	Elementary	USE	Irreflexive	Asymmetric	Transitive (in a telescopic way)	Preference relation
3.	NT	Narrower Term	Elementary	BT	Reflexive*	Antisymmetric	Transitive	Partial ordering relation
4.	BT	Broader Term	Elementary	NT	Reflexive*	Antisymmetric	Transitive	Partial ordering relation
5.	RT	Related Term	Eelmentary	—	Reflexive	Symmetric	Transitive (in a limited sense)	Equivalence relation (Quasi)
6.	SY	Synonymous or near synonymous	Combination of USE & UF	—	Reflexive	Symmetric	Transitive (in a telescopic way and in a limited sense)	Equivalence relation (Quasi)
7.	GS	Generic structure (belongs to the hierarchy ladder)	Combination of NT & BT	—	Reflexive	Symmetric	Transitive (in a limited sense)	Equivalence relation (Quasi)

*NT & BT could be treated as having reflexivity on the anology that relation 'subset' has reflexivity.

In Table 3 the relation SY is constructed as a combination of USE and UF and shown as quasi-equivalence relation. SY satisfies all conditions of equivalence relation except transitivity. It satisfies transitivity in a telescopic manner. That is to say all terms of an equivalence class of SY will occur at once place like

HEATING
SY HEATGAIN
PREHEATING
REHEATING
WARMING

Also transitivity fails in few cases. Further, the relation SY is very useful in grouping all synonymous and near synonymous terms as an equivalence class at the time of construction of a thesaurus.

Turning to our last hypothesis, generating an equivalence class of descriptors from a single descriptor is not strictly feasible as there is no perfect equivalence relation in thesaurus. However, in all retrieval methods we attempt to collect an optimum size equivalence class of descriptors balancing recall and precision. This could be better appreciated with the following example :

Suppose we are given an element (descriptor) 'BASE FLOW' and we need to know the equivalence class to which it belongs with respect to relation RT in the set of descriptors in NASA Thesaurus, in order to retrieve all documents on this concept from a system which is indexed using NASA Thesaurus.

We see in the thesaurus;

Step 1 : BASEFLOW
RT HEAD FLOW
WAKES

Cyclically looking under HEAD FLOW and WAKES, we find

Step 2 : HEAD FLOW
RT BASE FLOW
BLASIUS FLOW

INLET FLOW
LIQUID FLOW
PRESSURE DROP

WAKES
RT BACKWASH
BASE FLOW
BUBBLES
CAVITATION FLOW
CONTRAILS
DOWNWASH
DRAFT
DRAG
GROUND EFFECT
STROUHAL NUMBER
TURBULENCE
VORTICES

Right in the second step we are faced with lot of terms loosely related to 'BASE FLOW'. If we include all these we would be aiming at high recall and low precision. Otherwise low recall and high precision.

The relation RT can be further specified with weightages so that we can pass on upto a specified number of steps in collecting descriptors which belong to the equivalence class. Eventhough, this is a matter which needs further inquiry, automatic generation of equivalence class of descriptors from relation RT is highly desirable from the point of view of automated information retrieval.

Lastly we shall see what are Incomparable Elements or Incomparable Descriptors. With respect to relation RT, "BASE FLOW" and "CHARGED PARTICLES" are incomparable descriptors. It is obvious that no incomparable elements need be there in a set of descriptors selected for a search unless the topic of search itself has a relation between the concepts represented by incomparable descriptors.

5 CONCLUSION

Before closing this bird's-eye-view of relations of thesaurus from the

angle of Mathematics we could notice the great potentiality of application of more deeper concepts of Mathematics to explain the structure of thesauri and classification schemes. Analysing relations in classification schemes against the criteria discussed above would throw some light on the nature of relations in classification schemes. Similarly exploring relations in precoordinate thesauri would be desirable. Application of ternary relationship to thesauri (e.g. GS in NASA Thesaurus) and classification schemes may be yet another area for investigation.

The present study has revealed more areas for further inquiry and a good deal of intensive research in this direction may be worthwhile.

BIBLIOGRAPHY

1 GORDON (Charles K) Jr. Introduction to Mathematical Structures. 1967. Dickenson, California.
2 HERSTEIN (IN). Topics in Algebra. Ed. 2. 1975. Vikas Pub. House, New Delhi.
3 LANCASTER (FW). Vocabulary Control for Information Retrieval. 1972. Information Resources Press, Washington, DC.
4 LIPSCHUTZ (S). Schaum's outline of theory and problems of Set Theory and related topics. 1976. Tata McGraw-Hill, New Delhi.
5 LIU (CL). Elements of Discrete Mathematics. 1977. McGraw-Hill Book Co., New York.
6 MIKHAILOV (AI) and GILTAREVESKIJ (RS). An Introductory course on Informatics/Documentation. 1971. Revised & enlarged Ed. FID, The Hague.
7 VICKERY (BC). Classification and Indexing in Science. Ed 3. 1975. Butterworths, London.

About the Author

Dr. M. S. Sridhar is a post graduate in mathematics and business management and a doctorate in library and information science. He is in the profession for last 35 years. Since 1978 he is heading the Library and Documentation Division of ISRO Satellite Centre, Bangalore. Earlier he has worked in the libraries of National Aeronautical Laboratory (Bangalore), Indian Institute of Management (Bangalore) and University of Mysore. Dr. Sridhar has published four books ('User research: a review of information-behaviour studies in science and technology', 'Problems of collection development in special libraries', 'Information behaviour of scientists and engineers' and 'Use and user research with twenty case studies') and 74 research papers, written 19 course material for BLIS and MLIS, presented over 22 papers in conferences and seminars, and contributed 5 chapters to books. **E-mail:** sridharmirle@yahoo.com, mirlesridhar@gmail.com, sridhar@isac.gov.in ; **Phone:** 91-80-25084451; **Fax:** 91-80-25084475.