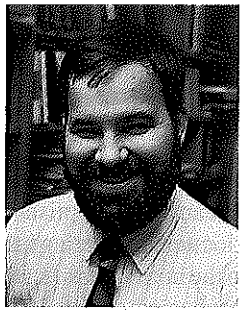
Robert M. Losee, Jr.
School of Information and Library Science, U.N.C., Chapel Hill, NC



Seven Fundamental Questions for the Science of Library Classification

Losee, R.M.: **Seven fundamental questions for the science of library classification.**
Knowl.Org. 20(1993)No.1, p.65-70, 15 refs.
For classification to advance to the point where optimal systems may be developed for manual or automated use, it will be necessary for a science of document or library classification to be developed. Seven questions are posed which the author feels must be answered before such optimal systems can be developed. Suggestions are made as to the forms that answers to these questions might take. (Author)

Introduction

As classification continues to move forward into a largely machine-based information age, it becomes increasingly possible to incorporate theoretically optimal classification procedures into libraries and other information systems. Classification for purposes here is considered the assignment of a value to an entity and the ordering and organization of these entities by these values. Current classification systems are unlikely to be optimal; even if they are optimal, proof of this would be of great benefit. These optimal systems must be developed on a scientific basis, a *science of classification.* This science can make explicit the fundamental parameters of library classification, as well as explain relationships between these components. While this science may rapidly become more formal than current classification systems and, indeed, somewhat more quantitative, this will be due largely to the relative ease with which elegant ordering systems may be developed based on models developed for the more traditional mathematical sciences.

An understanding of these facets of library classification may lead to an understanding of 1) the functional operation of present systems, 2) the effects of modification to these existing systems, and 3) how completely new systems might perform. Another role a science of classification may perform is the answering of practitioners' questions. However, this is not the primary purpose of this, or any other science, and practitioners should neither accept nor reject the basics of a science of classification primarily because of the way it addresses the problems they currently see as being central to the discipline. For a science to progress, it must do so on its own terms; to focus on the needs and concerns of practitioners is to blunt the progress of the discipline.

The science of library classification (1) has been developing at a slow pace. Yet, a clear statement of what classification is and how it can be made better is of critical importance to the further development of classification systems in both traditional libraries and in automated classification systems for fully electronic systems. A classification system developed from scientific principles has many advantages over more *ad hoc* classification systems. For example, one can understand why a theoretically based system performs as it does if it is grounded on scientific principles, providing information professionals with the capability to *explain* what is occurring. The ability of models to *predict* will allow library professionals, particularly technical services staff, to project into the future the co-location capabilities and retrieval performance of a classification system, given a collection organized consistent with scientifically based classification principles. The current state of classification, largely existing as an art and a set of philosophical constructs imposed on a knowledge base, does not allow this sort of prediction or explanation to be made.

Cluster based information retrieval systems have been the object of recent study. These systems group documents by subject but do not (and are not designed to) provide a single classifier system suitable for arranging books on a series of shelves. These clustering systems, as well as many classification systems, often assume that knowledge itself is naturally clustered in some form that is reflected in the structure of the clustering or classification system. While any enumerative system may be clustered at low level, many clustering systems fail to reflect the hierarchical structure of more widely accepted classification systems. Additionally, many classification systems are not explicitly designed to be hierarchical, although any classification system that can assign a written classifier to a document can be treated hierarchically.

Optimal enumerative classification systems can be developed that support patron browsing through user-oriented classification systems. While faceted and synthetic classification systems may eventually prove to be excellent bases for classification systems, enumerative systems may be most easily described and optimized by scientifically based systems. These scientifically based systems may be based on objective criteria and optimized for particular groups of classification system users. Unlike other classification warrants (2), this *functional* warrant

suggests that a particular classification system function might be optimized; in this case, for the end user. Variations of this functional warrant could allow for the optimization of other combinations of classification system components, including both the users' and the librarians' goals and needs, unlike current systems which do not explicitly or implicitly attempt to optimize.

Reclassification can be accomplished through the development of a new document ordering based on information not available when the initial classification was made, such as changes in frequencies of specific terms in recently published documents or bibliographic records. Procedures that could automatically and optimally reclassify an entire OPAC and produce new spine labels might be useful in reclassifying existing collections, if the significant effort expended in reclassification was justified by the increased access to the collection. As the number of online full-text systems increases, the ability to reclassify electronically stored documents will increasingly require *no* physical changes in the system, such as relabelling books or retyping catalog cards, with reclassification being simply a matter of running a computer program.

There are seven basic questions that need to be answered and further explored if a science of classification is to develop such that patrons and librarians can rely on the classification system to perform in the most efficient way possible, a degree of reliability not found in current systems. Designed to articulate with the methodologies of other scientific disciplines studying the organization and access of materials, these questions are both explicit and objective. The questions, directions that might be pursued in seeking their solutions, and examples of answers and the form that answers might take, as well as some of their variants and related problems, are as follows:

1. What should be the form of subject-indicating representations?

What *best represents* a document's subject if the representation is to be incorporated into a classification number assigned to a document? These representations take the form of combinations of letters, digits, and punctuation marks in LC, Dewey, and many other classification systems. If library classifiers are intended to group documents by subject for browsing and a classification number is to represent directly or indirectly the subject of the document, then understanding how a subject indicting variable is used in a call number is essential to understanding a classification system. Let us accept that bibliographic materials may not be completely describable by a single simple statement; the subject of an intellectual product has many facets or characteristics and thus a representation of what it is about must include the values of several subject-indicating variables. Note that determining the subject of a document is not strictly part of the science of classification, although determining the subject of a document is a part of the day-to-day work of a practicing classifier (3, 4, 5). It is studying the *representations* of these subjects and their subsequent *ordering* and *organization* that is the domain of scientific classification.

The subject indicating variables can hold fixed length representations, such as a fixed number of letters, integers (whole numbers), or binary numbers, or variable length or infinite representations, such as real numbers (numbers with decimal points) which, for example, represent 1/3 using an infinite string of digits. Because a real number can be infinite in length, its use as a representation to be printed on a spine label or stored in a computer record is excluded. A limited range of integers, for example, 0 to 9, could represent a set of 10 possible levels of subject-aboutness in a finite sized representation. A binary representation simply represents whether a document is about a subject or not. Representations can always be finite in size given a finite set of document-subjects.

Systems such as the Library of Congress and the Dewey Decimal Classification Systems (and their variants) implicitly limit the set of features which a document might exhibit, and thus may be represented using a finite representation system. However, they do not comfortably represent documents being equally about two different subjects or having two equally important attributes; their inherent hierarchical nature arbitrarily places one subject over another in terms of importance. By choosing one initial letter for the LCC for a document, one is excluding a set of other possibilities. While this may be seen as a strength, it may be a weakness when documents are interdisciplinary in character.

Allowing a document classification to have any possible combination of feature values can result in very long classification "numbers." These numbers may be decreased in size through the use of data compression (6), which reduces the size of a representation without information loss. The question of which form of data compression is the best for classification applications is strictly outside the boundaries of classification science, *per se*, and the size of classification values for individual documents should not be a major factor in designing or selecting a classification system. Given data compression techniques, the representations provided for a given document by any of a number of different possible classification systems may all be reduced to the same average size, assuming that all the classification systems represent the same universe of possible document-subjects.

Which subjects should be used in a classification and representation scheme? If inclusion of all potential subjects is a requirement of the system, the question becomes "Given a very large set of possible subject-indicating variables, which subset or alternative set of variables can most efficiently represent the subject content of all documents?" One answer to this question is to choose all and only those features that are *statistically independent* and thus provide no information about each other, minimizing redundancy. For example, features such as *cataloging* and *classification* are related and not independent in most library catalogs. To choose a variable that is in part dependent on another variable would be to select variables that do not carry as much information as is possible, thus

forcing the system to provide larger representations than necessary.

Independent variables may be computed through procedures such as factor or principal components analysis (7, 8). These procedures begin with data, such as natural language text, and compute a set of independent features, features that do not correlate with each other or provide information about each other. These features have the least amount of redundancy possible; encoding the subject of a documents will provide the greatest amount of information per character in the representation.

A conceptually different procedure, dimension reduction, explicitly reduces the number of features in a data set. Two examples of dimension reduction are provided by the field of cartography. Terrain maps are capable of showing three dimensional terrain in two dimensions. Similarly, maps using the Mercator projection show a three dimensional, round planet on a two dimension surface. Both of these methods distort and leave something to be desired, but they do diminish the number of dimensions or features represented by the descriptive mechanism, which is useful in cartography. Dimension reduction in classification involves removing factors of lesser importance and leaving those of greater significance. More specifically, one should continue to reduce dimensions as long as no information is lost. This is done by almost all human-based systems, which find it easy to distinguish the wheat from the chaff, but it is far less easily done with automated systems.

Classification scholars have suggested varying numbers of dimensions as being best for representation and ordering. For example, empirical support exists for Ranganathan's 5 dimensional space for his PMEST model which uses 5 basic categories (personality, matter, energy, space, time) (9). Whether these models are optimal, close to optimal, or adequate are empirical questions whose answer requires knowledge of the optimal classification system, given a set of desired constraints.

1a. What representational factors may be easily manipulated and remembered by human beings?

When selecting a feature set upon which to base a classification system, the library scientist may wish to choose those features that are easiest for humans to remember and manipulate, all other factors being equal. The ease of use of a classification system is a factor only when the classification system is a *visible classification system*, a system that needs to be examined and used directly by the patron. *Invisible classification systems* place similar documents near each other without displaying the classification number. These classification numbers are for the internal use of the computer or information system only. Invisible classification systems are likely to become increasingly important as computer-based systems provide retrieval through one mechanism, e.g., Boolean searching, and document organization to support browsing through documents through another mechanism. One might use Boolean queries to locate an initial interesting document in a database; neighboring similar documents might then be examined next. This notion of classification is more limited than existing library and book-based visible classification systems, which function both as browsing and finding tools, with users conducting known item searches, moving from the catalog to the book stacks with transcribed call numbers.

2. How should distance and dissimilarity values between individual document features be combined?

When making decisions about document classification, it is necessary to consider factors such as how alike two books are and how far apart they are in a collection. Yet, given a universe of different subject-variables, it is not obvious how one should combine the dissimilarity values that might be computed for each feature or variable to provide a composite document-similarity measure.

Two approaches to measuring similarities between multiple featured variables have been used widely in information theory (10). The first is to merely count the number of features by which two representations differ. This is often referred to as the Hamming distance, as well as being called the Manhattan or city block distance. The latter names are derived from the idea that if each possible representation is taken as a corner in a city, the distance between representations may be measured as the distance moving from one intersection to another by using streets and not cutting across lots, and where it is understood that each block has a length of 1. The other method of measuring distance, based on more traditional geometric considerations, computes distance "as the crow flies," that is, where one is not limited to traveling on streets. Both of these methods assume that one can combine distances associated with different variables.

3. What between-document or shelf distance should be minimized?

The distance between documents is a critical factor in a classification system. If documents are to be browsed, it is necessary for similar documents to be a short distance from each other while it may be desirable to explicitly place less similar documents at a greater distance. Thus, the study of document distance is a critical part of placing similar documents "close" together. Attention to distance may take many forms, the two extremes of which are examined here.

Ignoring the choice of taking no account of distance in a theory, the least attention one might pay to distance is to only examine adjacent documents, that is, only consider distances of 1. In essence, a collection may be classified by randomly selecting a document, placing next to it the document with the greatest degree of similarity, placing next to this document the document with the greatest degree of similarity to it, and so forth. This method is relatively stable, given the addition or deletion of documents, especially in large collections.

At the opposite extreme is taking a global view, taking into account similarities and distances between all possible pairs in a collection, adjacent or not. This method is somewhat sensitive to small changes; adding a single document can cause a ripple effect, modifying the order throughout the entire collection.

3a. Should distance be treated as a linear quantity? (11).

How individuals react to distances in the library is an empirical question that remains largely unexplored. Patrons appear to browse through a small area on a classified shelf, examining books located within a few centimeters to a few meters apart, at most. It is not obvious that a book located 200 meters from the start of browsing is twice as costly to retrieve as one located 100 meters from the starting point, that is, that the distance or effort function should be treated as linear. This is in part due to retrieval procedures when finding documents at some distance from each other. These differ from procedures used when browsing through sets of very close documents. Documents of interest on the same shelf are likely to be located through browsing, while documents several ranges away will need to be located through other procedures, such as through a known-item searched using information provided by a catalog record.

4. What is the cost of placing similar or dissimilar documents a certain distance apart?

Several factors potentially useful in developing classification systems are objectively determinable. The distance between documents is one such objective factor. The informational similarity between documents is likewise objective if information is objectively measured as is done by Shannon and some others (10). Given these objective quantities, how might a classification system be "customized" for a particular collection, assuming that classification systems should not be the same for all collections? A collection is most easily customized by noting that the difference between library collections indicates a difference in the information needs of the patrons; these different needs place different economic constraints on the classification system, suggesting that developing an economics of classification may result in improved adaptive classification systems.

The value of grouping certain documents together or placing two documents next to each other may be empirically determined and will be different in different informational environments. These costs may be further broken down into the costs associated with ordering (in some sense) documents based on a single feature. Once single feature costs have been determined, they may be combined, allowing for the estimation of the costs of placing documents at a certain distance with a given degree of dissimilarity.

5. In what order should documents be placed?

An ordering applied to a set of documents may be understood as a path through a space consisting of all possible documents or all possible document subject combinations. If a library's books were randomly placed, a classification system could be viewed as a path from each book to a randomly selected "next" book, providing an order in which they could be placed on a shelf. This path *is* the ordering principle within the classification system, and understanding the nature of the classification path is fundamental to understanding the nature of the classification system itself.

To examine paths through a collection, let us assume that each of n possible independent features is represented or is not represented in a document, with each value denoted by a 1 or a 0, respectively. The set of possible documents' subjects can be represented as the set of corners or vertices of an n-dimensional cube (12, 13). A classification system might use the binary Gray code (14) to supply an ordering of vertices and a means of representing what the document is about.

A simple binary coded classification system with only one feature in the system would be represented by a line with two ends, one end representing a 1 or the presence of the feature and the other end representing a 0 or the lack of the feature. Similarly, a system with two features would be represented by a square, with each of the four possible values (00, 01, 10, 11) for the two features represented by one of the four corners of the square. The best classification system ordering in this instance would probably be a path around the square.

A set of n features is said to produce an "n-cube." A classification system may thus also be seen as a *path connecting all vertices on the n-cube.* Each corner represents a particular set of coordinates. Thus, each path through this cube represents a counting sequence for the set of all possible coordinates. Classification may thus be viewed both as a path and as a counting sequence. This counting sequence provides the familiar filing order for books.

In many circumstances, multiple paths may satisfy the requirements that constrain a classification system. Consider, for example, a traditional table with four legs and thus with six planes composed of the vertices at the ends of the table legs (on the surface of the 3-cube): the top, bottom, front, back, left and right planes (diagonal planes are ignored here). We assume the distance minimizing criteria proposed earlier, in which the primary concern is to place each document next to another very similar document. This criteria may be met by starting at one corner of the top of the table and moving through all vertices on the table top. When this circuit has been completed, move straight to the floor and then make a circuit of the vertices on the bottom, where the table legs meet the floor. The criteria might also be met by moving around the edges of the front of the table, then, when this is completed, moving

to the back and then making a circuit of the back plane of the table.

These, as well as other possible circuits, all meet the criteria established earlier. This suggests that multiple classification systems may meet the requirements established for a classification system. One of these systems may be arbitrarily accepted, if all perform equally, or other criteria may be introduced to further limit the number of acceptable paths or classification systems.

6. Should paths and classification numbers stay constant for all collection variations, e.g., different collection sizes and concentrations?

Given that classification may be viewed as a path along the edges of the *n*-cube, should the varying degrees of density or unequal densities modify the best path? In other words, should different kinds of libraries or different sized libraries use different classification systems? If the classification system (1) is invisible or (2) the patron only uses this one collection or (3) the library or system manager follows the lead of retail stores who organize their goods in the manner they feel is best for their specific needs, then the answer is "Yes!"

6a. Should the path stay constant for all different collection emphases?

Consider a classification system for four possible documents. The documents are conceptually arranged clockwise around a square of the four vertices in the order A, B, C, D, with A in the upper left. Using the Hamming distance to measure both subject dissimilarity and the physical distance apart, the physical distance and subject-dissimilarities between adjacent documents are 1 and the distance and dissimilarities between diagonally opposed documents are 2. If all four documents are present, then the orders {A, B, C, D}, {B, C, D, A}, {C, D, A, B}, and {D, A, B, C} are (equally) optimal under many classification schemes and each ordering would be rationally acceptable. However, if the possibility is decreased or removed of having a document B, because of a shift in collection emphasis, it becomes clear that an ordering including the sequence {A, C, D} is far from optimal, because A is adjacent to a subject-opposite document, C. A better arrangement with B removed is clearly {A, D, C} or {C, D, A}. Therefore, the classification path should not stay constant for all different collection emphases.

6b. Should the path stay constant for all different collection sizes?

The path should stay constant as a collection grows in size only if the growth is uniform, that is, the probability that a document will have a given feature remains constant as the collection grows. If there is a shift as the collection grows, as there almost always is in practice, then the argument presented in the response to Question 6a above would hold, making it necessary to reclassify as a collection grows, if optimal classification is to be achieved.

7. How should classification performance be measured?

The development of a classifier performance measure is essential to the theory and development of a science of library classification. A measure should be developed based on whatever measurement characteristics the user desires to study with a measure. This idea of a measure treats a measure as being related to a model; a measure is developed so that it accurately analyzes the characteristics that the model and its assumptions dictate are important. One might arbitrarily decide, for example, that the measure should have a value of 1 if perfect classification occurs and a value of 0 if random classification occurs (12). This scale might be easily changed so that it would provide a different range of values, for example, 0 to 100, or change the meaning of the top and bottom of the range, changing the bottom value of 0 to occur only with the worst possible classification performance.

7a. What is the best possible classification performance?

An interesting question is whether the "best" system should be a very good practical system or one that might never be fully achieved but is perceived to be the best imaginable. For example, an ideal but never to be achieved system would place adjacent to each other all the documents each user might find of interest. For any realistic set of users, this is unachieveable. Should this be used as the "best" performance obtainable?

7b. What is the worst possible classification performance?

A measure of classification performance may require knowledge of the performance of the worst possible classification system. Accepting the *n*-cube model described above, the worst possible classification system is one where paths constantly move almost directly across the *n*-cube from one corner to an almost-opposite corner. Always moving to the opposite corner is impossible, as one would constantly move back and forth between the same pair of corners. This worst-possible system may be implemented in a classification system through use of the anti-Gray code proposed by Hamming (15).

Conclusions

Library classification, the ordering of materials to facilitate patron browsing, can be studied objectively. Classification systems can be optimized, providing maximal browsing support for the user. Developing these "best" systems to support both traditional book-based libraries and paperless, full-text electronic systems requires answers to the seven fundamental questions examined here. In addition, *as professionals, we should strive* for the use of library systems consistent with models that allow us to *explain* what happens and to *predict* future performance.

These questions have been proposed in such a way that answers may be more easily found, given research in this

field or borrowing results from other scientific disciplines faced with similar problems (librarians are not alone!). In particular, the assumption that subject bearing features are binary has allowed partial answers to some of these questions to be framed in terms of the *n*-cube, an easily analyzed theoretical construct that generalizes from the familiar three-dimensional cube. The author believes that studies of this model will allow for the further development and improvement of objective classification procedures.

Empirical research is needed to examine human behavior and preferences. For example, questions about user reactions to distances as well as uncertainty about costs of placing documents at a certain distance can only be studied through empirical research. Given both this further empirical work and theoretical analysis of linear classification systems, the science of library classification can be expected to move forward at an increased pace.

References:

(1) Moravcsik, M.J.: The classification of science and the science of classification. Scientometrics 10(1986)No.3-4, p.179-197

(2) Beghtol, C.: Semantic validity: Concepts of warrant in bibliographic classification systems. Libr.Resources & Techn.Serv. 30(1986)No.2, p.109-125

(3) Lancaster, F.W.: Vocabulary control for information retrieval. Arlington, VA: Inform. Resources Press 1986. 270p.

(4) Foskett, A.C.: The subject approach to information. 4th ed. Hamden, CT: Linnet Books 1982. p.574

(5) Kwasnik, B.: The importance of factors that are not document attributes in the organization of personal documents. J.Doc. 47(1991)No.4, p.389-398

(6) Storer, J.A.: Data compression: Methods and theory. Rockville, MD: Computer Science Press 1988. 413 p.

(7) Borko, H.: Research in computer based classification systems. In: Theory of Subject Analysis: A sourcebook. Littleton, CO: Libraries Unlimited 1985. p.287-3005

(8) Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J.Amer.Soc.Inform.Sci. 41(1990)No.6, p.391-407

(9) Iyer, Hemalatha: Subject representation and entropy. Int.Classif.19(1992)No.1, p.15-18

(10) Losee, R.M.: The science of information. New York, NY: Academic Press 1990. 213p.

(11) Boyce, B., Douglass, J.S., Rabelais, J., Shiflett, L., Wallace, D.P.: Measurement of subject scatter in the Superintendent of Documents Classification. Government Publ.Rev. 17(1990)p.333-339

(12) Losee, R.M.: A gray code based ordering for documents on shelves: Classification for browsing and retrieval. J.Amer.Soc.Inform.Sci. 43 (1992)No.4, p.312-322

(13) Cosgrove, S.J.: Item classification using the N-Cube's Hierarchical Knowledge Representation Schema. In: Libraries and Expert Systems. Los Angeles: Taylor graham 1991. p.88-98

(14) Gilbert, E.N.: Gray codes and paths on the N-cube. Bell Ssystem Techn.J. 37(1958)p.815-826

(15) Hamming, R.: Coding and information theory. 2nd ed. Englewood Cliffs, NJ: Prentice Hall 1986. 259p.

Dr.Robert M.Losee, Jr., School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA.

European Academy for Standardization, e.V. (EURAS)

Recently this Academy was founded in Hamburg, Germany, by researchers from different academic areas, such as economics, engineering, law. Members have joint by their interest for standardization in international relations. Aims and activities are directed towards promotion of research in relation to standardization, utilization of the results in education as well as their publication. Promoted is also the construction of an international network of universities and other research institutions in this regard. EURAS does not aim at establishing standards and nor does engage in any such preparatory work. Research activities of the Academy are rather concerned with reasons for the manifold effects of standards as well as with an analysis of the significance of standardization in different disciplines and social areas. - For further information please contact: European Academy for Standardization e.V. (EURAS), GV Mr.A.Inklaar, PF 700 822, Holstenhofweg 85, D-22039 Hamburg.

Museum Computer Network (MCN)

The 1993 Annual Conference of MCN will be held from Nov.3-6, 1993 at the Seattle Sheraton Hotel, in Seattle, WA, USA. Program sessions will address all areas of museum computing, with particular emphasis on multimedia applications for exhibits, education, collections management, conservation, and research. For more information please contact Mrs.Diane Zorich, MCN'93 Program Chair, Peabody Museum of Archaeology and Ethnology, Harvard University, 11 Divinity Ave., Cambridge, MA 02138. Fax: 617-495-7535.