A Gray Code Based Ordering for Documents on Shelves: Classification for Browsing and Retrieval

Journal of the American Society for Information Science
43(4) 1992, 312–322.

Robert M. Losee
University of North Carolina
Chapel Hill, NC 27599-3360 U.S.A.

losee@ils.unc.edu

June 28, 1998

Abstract

A document classifier places documents together in a linear arrangement for browsing or high speed access by human or computerized information retrieval systems. Requirements for document classification and browsing systems are developed from similarity measures, distance measures, and the notion of subject aboutness. A requirement that documents be arranged in decreasing order of similarity as the distance from a given document increases can often not be met. Based on these requirements, information theoretic considerations, and the Gray code, a classification system is proposed that can classify documents without human intervention. It provides a theoretical justification for individual classification numbers going from broad to narrow topics when moving from left to right in the classification number. A general measure of classifier performance is developed and used to evaluate experimental results comparing the distance between subject headings assigned to documents given classifications from the proposed system and the Library of Congress Classification (LCC) system. Browsing in libraries, hypertext, and databases is usually considered to be the domain of subject searches. The proposed system can incorporate both classification by subject and by other forms of bibliographic information, allowing for the generalization of browsing to include all features of an information carrying unit.

1 Introduction

Mentioning document classification may bring to a layperson's mind the Dewey Decimal system, while to the library professional, it usually provokes thoughts of the problems associated with bringing together similar materials, the costs of cataloging, or the difficulty of justifying one classification system over another. Svenonius (1981) has suggested that "the main use for classification, at least in the United States, has been to facilitate browsing." While classification has other roles in the library, from the mundane to playing "a direct role in the creation of original knowledge" [6], classification as a tool to support browsing activities in both libraries and database systems will be the focus of this research and discussion. Although numerous library classification systems have been developed, most have been developed from philosophical and taxonomical considerations [12, 28]; few have been based on more scientific criteria. A set of precise requirements for a document classification system are provided here as well as a classification system consistent with these requirements. This classification system and an associated measure of classification performance have been developed based on concepts used by information professionals who use document classification systems.

Many different quantitative methods are available for determining similarities between documents and for optimal ordering for documents. Methods similar to the work here include studies of coding techniques that place similar documents near each other, given a query [10], and hashing techniques that maintain order [13]. An information theoretic method is developed here which we believe is easier to interpret and consistent with the needs of document classification systems. Other similarity techniques and measures could be used and need to be explored, e.g., the expected mutual information measure. The method developed here for evaluating a classification system is based on information naturalness concerns. Performance measures using statistical correlation could be satisfactorily used, measuring, for example, the correlation between the ordering provided by the classification system and the ordering provided by a perfectly ordered collection of documents.

Aristotle suggested that a science has as its base a series of predicates that, in effect, define the science. For example, the science of physics uses fundamental predicates such as *position*, *velocity*, and *inertia* on which it builds its theories. If a science of document classification is to be developed, it will probably have at its base predicates like the *shelf-distance* between documents and the *subject-similarity* of documents.

Developing a document classification system results in a rather unique set of problems. The primary function of a document or library classification system is to assign a value to each information carrying unit so that when the items are sorted by this value, like information carrying units are grouped together. Co-location of

documents allows users to benefit from browsing through nearby similar materials once they have located a single potentially relevant item, increasing the precision of the browsing. To provide for browsing capabilities, a library classification system should meet several requirements. Broad requirements for a classification system have been suggested by Wynar and Taylor [28]. We believe that a classification system supporting browsing should

- assign classification values objectively (*objectivity* requirement),
- provide a single classification system capable of classifying all possible documents (*inclusion* requirement),
- provide a linear structure (*linearity* requirement),
- assign values to documents so that when one moves away from any document in either direction on a conceptual shelf or in a database, the documents become increasingly dissimilar (*increasing distance-dissimilarity* requirement).

A classification system meeting these requirements will fulfill the needs of a librarian or database manager wishing to place like items together for browsing. In a library, for example, similar documents may be consecutively retrieved by retrieving books as one moves down a shelf.

Other characteristics of a classification system, while desirable, are not mandatory. A classification system might have the following characteristics:

- be easily (quickly) searched,
- be easy for librarians to use when classifying documents,
- allow for classification by computer,
- be consistent with an existing, popular system,
- provide explanatory power, or
- be readily adaptable to incorporate changes in the materials classified.

These requirements and desirable characteristics, when combined with the predicates of a science of classification, will be used as the basis for the proposed classification system. A guide to the automatic classification literature is provided by [18].

A document or book has a number of *features* which may characterize the topic of the document. Subject indicating features may include library subject headings,

while other features may indirectly indicate the document's subject, such as words that occur in a document's title, the language, date and place of publication, and characteristics of the author that might provide information about the subject of the author's work.

Features are assumed here to be binary. A feature is thus present or absent, 1 or 0, depending on the degree to which the document is about the feature. Below, the expression "the probability of a feature" refers to the probability that the feature has the value 1. Features are assumed to be treatable as statistically independent.

The primary purpose of a classification system is to enable information searchers to browse through documents. Once an initial document has been located on a shelf in a library or in a window in a hypertext system, searchers often choose to examine related materials [2, 5, 8, 22, 23]. Classification provides this clustering of similar materials [9, 17]. Classification and browsing systems typically group items by subject similarity, but clustering procedures may also take into consideration bibliographic features not normally thought of as subject related, such as type of binding. Although the classification system discussed below is capable of incorporating bibliographic features used in known item searches and is not limited to conventional subject clustering, further research will be necessary to determine how useful classification by other than subject-features would be to library patrons and database searchers.

Below, a Gray code based classification system is used to group a set of documents together. The classification procedure groups documents based on the documents' subject-bearing features. A measure is proposed which can be used to evaluate the performance of a classification system. Experiments suggest that the Gray code based classification method places documents closer together than does the Library of Congress Classification system.

2 Shelf-Distance

For notational simplification, a classification system is understood as an ordering of a series of documents or text fragments on a single conceptual shelf holding N documents. A document or book is denoted as B_i, with the subscript indicating the position of the document on the shelf. The documents are ordered

$$B_\phi, B_1, B_2, B_3, \ldots, B_{\nu-1}, B_\nu, B_{\nu+1}, \ldots, B_{N-1}, B_N$$

where the subscript indicates the position of the document relative to the leftmost document on the shelf, with position ν representing an arbitrarily chosen position under consideration. B_ϕ is a hypothetical document with no subject content of any sort at the left end of the shelf.

Distance measures in most common geometric spaces, such as a conceptual document shelf, must meet several criteria, including:

$$\begin{aligned}\mathcal{D}(B_i, B_i) &= 0 \\ \mathcal{D}(B_i, B_j) &= \mathcal{D}(B_j, B_i) \\ \mathcal{D}(B_i, B_j) &\leq \mathcal{D}(B_i, B_k) + \mathcal{D}(B_k, B_j),\end{aligned}$$

where the distance between B_i and B_j is denoted as $\mathcal{D}(B_i, B_j)$ for all B_i and B_j.

It is always the case that $\mathcal{D}(B_\nu, B_i) < \mathcal{D}(B_\nu, B_j)$ for all i, j, and ν when $j < i < \nu$ or when $j > i > \nu$. Unless indicated otherwise, it is assumed that the distance function, as well as other functions, are only defined over the set of documents to the right of B_ν and over the set of documents to the left of B_ν but not over the set of all documents taken together. Therefore, the distance between B_i and B_j is not defined if B_ν is between B_i and B_j.

3 Distance and Dissimilarity

The *dissimilarity* function, or distance in a conceptual subject space between two documents B_i and B_j, is denoted as $\mathcal{U}(B_i, B_j)$. The subject of a document, what it is *about*, is considered to be determined by all subject related aspects or features of the document. The subject-dissimilarity may be computed as a function of the degree of difference in feature values between two documents.

Because the dissimilarity function may be understood as a distance measure in a conceptual space, the following holds:

$$\begin{aligned}\mathcal{U}(B_i, B_i) &= 0 \\ \mathcal{U}(B_i, B_j) &= \mathcal{U}(B_j, B_i) \\ \mathcal{U}(B_i, B_j) &\leq \mathcal{U}(B_i, B_k) + \mathcal{U}(B_k, B_j).\end{aligned}$$

A classification system may attempt to maximize or minimize factors combining the distances between documents both in physical space (shelf-distance) and subject similarity or dissimilarity between documents. Numerous techniques for combining distance and similarity measures are used in mathematical clustering procedures [1].

A classification function consistent with the requirements for a document classification system, and in particular, the *increasing distance-dissimilarity* requirement, mandates that items be placed in weakly ascending order by the value of a subject-dissimilarity measure as one moves out in either direction from any given document. Weakly ascending order implies that the value for a selected feature

for each item is greater than or equal to the corresponding value for the preceding item. This document arrangement has the effect of ordering documents so that at any one point on the conceptual shelf, the documents to both the right and the left of the point are arranged in order of increasing dissimilarity to the document at the chosen point. This classification function incorporates the shelf-distance between items by requiring ordering by subject dissimilarity; items at a greater distance must have a greater degree of dissimilarity than closer objects.

To be consistent with the *increasing distance-dissimilarity* requirement, documents must be classified so that

$$\begin{aligned} \mathcal{D}(B_\nu, B_i) &< \mathcal{D}(B_\nu, B_j) \\ \mathcal{U}(B_\nu, B_i) &\leq \mathcal{U}(B_\nu, B_j), \end{aligned}$$

for both $j < i < \nu$ and $\nu < i < j$, that is, for documents to the left or the right of B_ν. Given a document at location ν, if i is greater than j and documents at locations i and j are shelved to the left of document ν, then the document at location j is less similar or equally similar to the document at location ν than is the document at location i.

These functions and rules do not make use of the distances to documents on both sides of B_ν at the same time. If the functions were defined over the entire set of documents on the conceptual shelf, it might be the case that for $j < \nu < i$,

$$\begin{aligned} \mathcal{D}(B_\nu, B_i) &< \mathcal{D}(B_\nu, B_j) \\ \mathcal{U}(B_\nu, B_i) &> \mathcal{U}(B_\nu, B_j). \end{aligned}$$

These proposed distance and dissimilarity functions may now be used to develop a possible classification function. A classification function, $\mathcal{C}(B_i)$ takes as its input or argument all the available features of the document, referred to as the document's profile, and returns a number that increases as the position of the document moves to the right on the conceptual shelf. Many existing classification systems, such as the Dewey Decimal Classification (DDC) system or the LCC systems, can be seen as providing an increasing value for those documents further to the right on the single conceptual shelf. Documents ordered by the value of the classification function may be said to be classified.

The distance in classification space between two classified documents B_i and B_j is denoted as $\mathcal{C}_\delta(B_i, B_j)$. Because this is a proper distance function, the following are true:

$$\begin{aligned} \mathcal{C}_\delta(B_i, B_i) &= 0 \\ \mathcal{C}_\delta(B_i, B_j) &= \mathcal{C}_\delta(B_j, B_i) \\ \mathcal{C}_\delta(B_i, B_j) &\leq \mathcal{C}_\delta(B_i, B_k) + \mathcal{C}_\delta(B_k, B_j). \end{aligned}$$

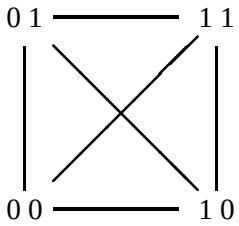


Figure 1: Four documents in a conceptual subject space.

The classification function may now be defined as $C(B_i) = C_\delta(B_i, B_\phi)$, that is, the difference in the classification distance between the document in question and a subject neutral or null document.

If the classification system is consistent with the *increasing distance-dissimilarity* requirement, then $\mathcal{U}(B_i, B_\nu)$ is monotonic with $\mathcal{C}_\delta(B_i, B_\nu)$ and when the value of one function rises or falls the value of the other function also rises or falls. If $j > i > \nu$, then $\mathcal{C}(B_\nu) \leq \mathcal{C}(B_i) \leq \mathcal{C}(B_j)$ and $\mathcal{U}(B_\nu, B_i) \leq \mathcal{U}(B_\nu, B_j)$. The further to the right and the more distant documents are from B_ν, the greater must be their classification value and the greater must be the dissimilarity between a document and B_ν. Similarly, if $j < i < \nu$, then $\mathcal{C}(B_\nu) \geq \mathcal{C}(B_i) \geq \mathcal{C}(B_j)$ and $\mathcal{U}(B_\nu, B_i) \leq \mathcal{U}(B_\nu, B_j)$.

Thus, some characteristics of a classification function meeting the initially established requirements can be analytically described.

4 Unachieveable Requirements

In some cases the ordering requirements placed on a classification system by the *increasing distance-dissimilarity* requirement cannot be met. Therefore, a classification system based upon a set of requirements should not use the *increasing distance-dissimilarity* requirement as part of its foundation, although the requirement may serve as a helpful guide.

Consider a simple situation with four documents, with characteristics $00, 01, 10,$ and 11, arranged at a subject distance from each other as shown in Figure 1. Each vertical or horizontal line represents 1 unit of subject distance or dissimilarity; the name of the document represents the document's coordinates, with the first digit representing a feature graphed on the x axis and the second digit representing a feature graphed on the y axis. The documents located at ends of diagonal lines are further apart than are the document's immediate neighbors on either vertical or horizontal lines. Thus, if the distances in Figure 1 are measured geometrically, $\mathcal{U}(B_{00}, B_{11}) = \mathcal{U}(B_{01}, B_{10}) = \sqrt{2} = 1.414$.

By attempting to arrange these documents in a linear fashion consistent with the classification requirements, it becomes obvious that the *increasing distance-dissimilarity* requirement cannot be met. Assume that the four shelf positions capable of holding a document are numbered, from left to right, $1, 2, 3, 4$. The document with profile 11 is assumed to be on the right half of the shelf in either positions 3 or 4. If 11 is in position 4, then either it must be next to 00, to which it is more dissimilar than it is to 01 and 10 in positions 1 and 2, violating the *increasing distance-dissimilarity* requirement, e.g., $(01, 10, 00, 11)$ or $(10, 01, 00, 11)$, or either 01 or 10 is in position 3, e.g. $(?, ?, 01, 11)$ or $(?, ?, 10, 11)$. Either of the values in position 3, i.e., 01 or 10, is closer to 00 than to the other, so 00 must be in position 2, e.g. $(10, 00, 01, 11)$ or $(01, 00, 10, 11)$. This forces the one of the 01, 10, pair that is not in position 3 to be in position 1. However, the document in position 1 is now closer to the document in position 4 than to the document in position 3, e.g., 10 is more about 11 than is 00 in $(10, 00, 01, 11)$, while 01 is more about 11 than is 00 in $(01, 00, 10, 11)$, both violating the *increasing distance-dissimilarity* requirement.

If 11 is in shelf position 3 instead of position 4, then either 00 is in position 4, e.g.,$(?, ?, 11, 00)$ or one of the 01, 10 pair is in position 4, e.g., $(?, ?, 11, 01)$ or $(?, ?, 11, 10)$. If 00 is in position 4, it is more dissimilar to the document in position 3 than it is to either of the two possible documents in positions 1 and 2, no matter what the arrangement of the latter two, e.g., $(01, 10, 11, 00)$ or $(10, 01, 11, 00)$. If 01 is placed in position 4, then to maintain decreasing similarity, 10, which it is most unlike, must be in position 1 and 00 must be in position 2, e.g. $(10, 00, 11, 01)$. However, document 01 at position 4 is now more similar to 00 at position 2 than it is to document 11 at position 3, violating the *increasing distance-dissimilarity* requirement. The same problem arises if 10 is placed in position 4, e.g. $(01, 00, 11, 10)$. Other arguments are possible, including those based on a mirror image of the above example, but it is clear that the proposed *increasing distance-dissimilarity* requirement often cannot be met.

This has a particularly important impact on classification for probabilistic retrieval [3, 20, 24, 27]. Systems consistent with these probabilistic document retrieval models rank documents in decreasing order of their probability of relevance. A static document arrangement which meets the *increasing distance-dissimilarity* and other requirements would place those documents with the highest rankings closest to a given point, while as one moved away from this point, one would encounter documents with decreasing ranks. Because the *increasing distance-dissimilarity* requirement cannot be met in some cases, a static arrangement of documents for browsing that is consistent with the goals of probabilistic information retrieval cannot be designed.

The *increasing distance-dissimilarity* requirement for a classification system

can be modified to provide a requirement with which a classification system can be consistent. This modified requirement mandates that when documents are being classified, the unclassified document which is most similar to the last document classified, the rightmost document on the conceptual shelf, is placed on the shelf as the new rightmost document. This does not place any restrictions on the dissimilarity of pairs of documents with intervening documents, except through the transitivity of similarity; that is, if A, B, and C are adjacent documents and in this order, and B is highly alike A and C is very similar to B, then obviously C is rather similar to A, although the similarity between $\{A, C\}$ is less than or equal to the similarity between the document pairs $\{A, B\}$ or $\{B, C\}$.

The following modified requirement is suggested in lieu of the *increasing distance-dissimilarity* requirement:

- place documents on a shelf, left to right, so the next document placed on the shelf is the document not already on the shelf most similar to the rightmost document (*modified increasing distance-dissimilarity* requirement).

If the classification system is consistent with the *modified increasing distance-dissimilarity* requirement, then for $j = \nu + 2$ and $i = \nu + 1$, $\mathcal{C}(B_\nu) \leq \mathcal{C}(B_i) \leq \mathcal{C}(B_j)$, $\mathcal{C}_\delta(B_\nu, B_i) \leq \mathcal{C}_\delta(B_\nu, B_j)$, and $\mathcal{U}(B_\nu, B_i) \leq \mathcal{U}(B_\nu, B_j)$. The further to the right and the more distant documents are from B_ν, the greater is the value of the classification function and often the greater is the dissimilarity between the documents and B_ν. If $j = \nu - 2$ and $i = \nu - 1$, then $\mathcal{C}(B_\nu) \geq \mathcal{C}(B_i) \geq \mathcal{C}(B_j)$, $\mathcal{C}_\delta(B_\nu, B_i) \leq \mathcal{C}_\delta(B_\nu, B_j)$, and $\mathcal{U}(B_\nu, B_i) \leq \mathcal{U}(B_\nu, B_j)$.

5 The Gray Code

An ordering principle and a classification system consistent with the modified requirements can be obtained by representing document features by the binary Gray code. Each feature is represented by a 1 or a 0 in the Gray code representation for each document. The Gray code provides a representation for each ordered item such that there is only 1 character difference between the representation for an item and the representation for the next item [15]. This may be more formally explicated by defining the *Hamming distance* between two individual binary representations for items as the number of features by which they differ [21]. For example, the Hamming distance between 01101 and 00111 is 2 because they differ in 2 positions, the second and the fourth. The Gray code may be more formally defined as a code consisting of a series of representations $R_1, R_2, \ldots, R_n$ ordered by their numeric value such that the Hamming distance between R_i and R_{i+1} is 1 for all i.

Table 1: Reflecting Gray code.

Decimal	*Binary*	*Gray Code*
0	000	0
1	001	1
2	010	11
3	011	10
4	100	110
5	101	111
6	110	101
7	111	100

Gray code representations for the numbers from 1 to 8 are provided in Table 1. There are several different Gray codes which may be applied to document classification [14]. The form used here is often referred to as the *reflected Gray* code. A listing of the codewords or representations can be split into two equal portions at several points with the top half having a 0 as the leftmost bit and the bottom half having a 1 as the leftmost bit, assuming the same number of bits in numbers in both halves [11]. This can be done to the data in Table 1 by taking either the first two, four, or eight numbers and splitting them into equal sized top and bottom halves. Following convention, leading zeroes are suppressed.

Simple procedures exist for converting numbers from the standard binary representation of numbers to a Gray code representation. For example, a standard binary number can be converted into the Gray code by moving from the rightmost bit to the leftmost bit and changing the value of a bit if the bit to its left is a 1. Thus, the binary number 110 becomes 101 in the Gray code, with each of the rightmost two bits transposed in value because the bit to the left of each of the two bits is a 1. A number in Gray code can be easily changed into a standard binary number, again while moving from right to left, by changing the value of a bit if the sum of the bits to its left is an odd number. Continuing with the above example, 101 in Gray code is changed back into the standard binary number 110 because the sum of the bit values to the left of each of the two rightmost bits is an odd number, changing the value of the two rightmost bits.

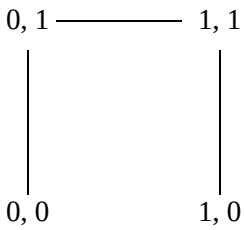


Figure 2: Four documents arranged in order by the value of their position in space.

6 Classification Using the Gray Code

A classification system can be implemented using the Gray code to represent the presence or absence of features of the document being classified. Documents are then ordered by the value of the Gray code. Each profile is written as a number, with each position representing a feature's value. Using the reflecting Gray code in Table 1, a document with profile 11 would be placed before one with characteristics 10, because the former proceeds the latter in value.

The four profiles in Figures 1 and 2 may be treated as Gray code representations and ordered as suggested by Table 1. The ordering is indicated by the lines in Figure 2. The Hamming distance between each adjacent profile is 1, the lowest possible Hamming distance for non-identical profiles. This is an intrinsic property of consecutive representations in the Gray code. Classification using the proposed method requires that the presence or absence of all features be represented directly or indirectly by binary values at a certain position in the Gray code.

A sample set of documents shown in Table 2 illustrates how documents may be arranged using the arithmetic (regular) counting sequence and the Gray code counting sequence. The arithmetic or alphabetic counting sequences are used in most library classification systems such as the Dewey Decimal and Library of Congress Classification systems. The Hamming distances, representing the subject dissimilarities between adjacent documents, are greater with the arithmetic counting sequence than with the proposed Gray code based system. The average distance in the case of the Gray code is 1 while the average distance between documents using the more traditional ordering is 1.73.

The shortest possible code is that with no redundancy between features. This is found only when document features are statistically independent, that is, when one feature doesn't provide an information about another feature. One can force independence of represented features by using as features factors generated through factor analysis, which computes statistically independent features underlying a set of statistically dependent features [4, 7]. For purposes here, we have chosen not

Document	Binary Code	Hamming Distance	Document	Gray Code	Hamming Distance
A	0 0 0 0	–	A	0 0 0 0	–
B	0 0 0 1	1	B	0 0 0 1	1
C	0 0 1 0	2	D	0 0 1 1	1
D	0 0 1 1	1	C	0 0 1 0	1
E	0 1 0 0	3	G	0 1 1 0	1
F	0 1 0 1	1	H	0 1 1 1	1
G	0 1 1 0	2	F	0 1 0 1	1
H	0 1 1 1	1	E	0 1 0 0	1
I	1 0 0 0	4	M	1 1 0 0	1
J	1 0 0 1	1	N	1 1 0 1	1
K	1 0 1 0	2	P	1 1 1 1	1
L	1 0 1 1	1	O	1 1 1 0	1
M	1 1 0 0	3	K	1 0 1 0	1
N	1 1 0 1	1	L	1 0 1 1	1
O	1 1 1 0	2	J	1 0 0 1	1
P	1 1 1 1	1	I	1 0 0 0	1

Table 2: Document ordering with binary and Gray codes. Hamming distances represent the distance from a document (with the distance indicated) to the adjacent document immediately above it, understood as the document immediately to its left on a shelf. The average Hamming distance between adjacent documents using the binary code is 1.73, while the average distance between adjacent documents is only 1 for Gray code ordered documents, indicating that the latter is superior at grouping documents by subject.

to optimize the code for the shortest length possible. This is not to suggest that the use of factor analytic techniques would not improve the proposed classification system.

Features may be placed in any order into the Gray code and the modified requirements will be met. However, the expected dissimilarity between documents, as represented by the sum of the expected dissimilarity between features, can be decreased by placing those features with the least expected dissimilarity furthest to the right in the code, while the features with the greatest expected dissimilarity are placed to the left. Figure 3 shows the relationship between the probability of a feature occurring and its expected Hamming distance. The expected Hamming distance may be computed as the product of the proability that a feature will change values times the probability of starting with that feature times the distance associated with that change. The expected Hamming distance of a feature, with probability of having the value 1 is $.5$, may be computed as (the probability of having a 0, $.5$, times the probability of having a change to a 1, $.5$ times the distance of 1,) plus (the probability of having a 1, $.5$, times the probability of having a change to a 0, $.5$, times the distance of 1,) equalling $(.5 \times .5 \times 1) + (.5 \times .5 \times 1) = .5$. The expected Hamming distance of a feature with a probability of $.9$ similarly would be computed as $(.9 \times .1 \times 1) + (.1 \times .9 \times 1) = .18$. In general, if the probability of a feature occurring is p, the expected Hamming distance for that feature is $2p(1-p)$.

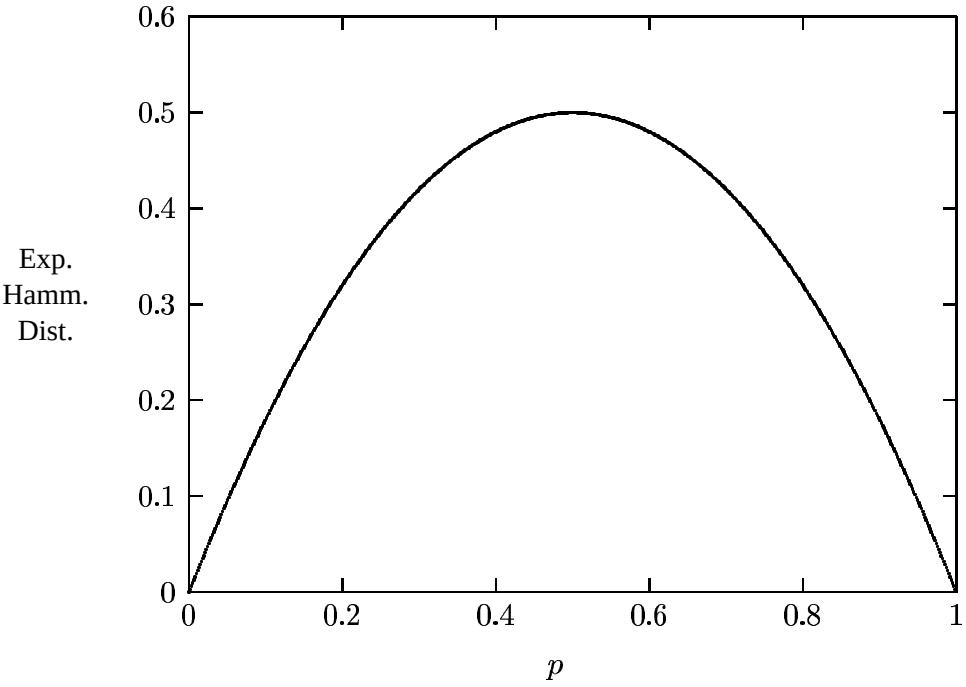
When all features have low probabilities, they can be arranged from left to right in order of decreasing probability of occurrence, which is the same as ordering the features from left to right in decreasing order of their expected dissimilarity. However, in cases where some features have or might have probabilities over $.5$, ordering by decreasing expected dissimilarity, which is theoretically justified, will not produce the same ordering as ordering by decreasing probability.

Note that most natural language terms occur in text with a probability of less than $.1$ and similar low probabilities might be expected of most other features. As the rightmost bits "cycle" more frequently than bits to the left, this arrangement will result in a lower expected dissimilarity between adjacent documents than would be the case if the least probable features with the greatest expected dissimilarity were on the right side and cycled most frequently.

The degree to which a document is "about" a topic may be measured as the information that the document contains on the subject corresponding to the features contained in the document. Information may be measured in this context using the measure developed by Shannon [25, 21]. Information is inversely related to the probability of a feature occurring.

Given this measure of information, the degree of difference between the information in one document and another may be computed as the information contained in the differing features of the document "on the right." A feature always

Figure 3: Expected Hamming distance for a feature with probability p.



carries information, whether its value is 1 or 0. When a change in feature values occurs, when moving from one document to an adjacent document, the difference in information associated with the change to a new feature value may be understood as the information associated with the changed feature. Because the feature values of 0 and 1 will result in different individual information measures, the order in which one moves (left to right or right to left) will result in different measures of information. The average information associated with any new features is used here to avoid this problem; this results in the same measured information whether the documents are analyzed from left to right or from right to left.

The difference in information between the k^{th} feature in documents i and j, where $B_{i,k} \neq B_{j,k}$ and $p_k = \Pr(B_{j,k})$, may be computed as

$$d_k = -p_k \log p_k + (1 - p_k) \log(1 - p_k).$$

Logarithms are computed to base 2 if information is measured in *bits*. The use of other types of difference measures is possible and will be the subject of future research.

The difference between two adjacent documents, i and j, the degree that they are about different subjects, may be computed as

$$\mathcal{U}_I(B_{i,k}, B_{j,k}) = \sum_k d_k$$

where there are n features, numbered 1 through n and for all k where $B_{i,k} \neq B_{j,k}$. Figure 4 illustrates how the expected value of the information dissimilarity, referred to here as the *expected information dissimilarity*, changes as the probability varies from 0 to 1. If features have low probabilities ($p < .5$), the expected informational dissimilarity will decrease as the probability of a feature occurring decreases.

The effect of feature ordering based on expected information dissimilarity is illustrated in Table 3. Two features are indicated for each document: one with the probability (.5) of a feature having the value 1 and the other feature with probability (.25) of the feature having the value 1. The latter feature has a lower expected information dissimilarity value, as can be seen by referring to Figure 4. The left half of Table 3 shows the information dissimilarity when features are ordered from left to right by decreasing expected information dissimilarity. The total information dissimilarity can be seen to be 2.62 bits with an average information dissimilarity of .374 bits. The right half of Table 3 shows the information dissimilarity when features are ordered from left to right by increasing expected information dissimilarity. The total information dissimilarity can be seen to be 2.81 in this case with an average information dissimilarity of .401 bits. The average information dissimilarity is obviously lower for the case where features were ordered by decreasing expected information dissimilarity and is thus superior with this data.

Figure 4: Expected informational dissimilarity (measured in bits) for a feature with probability p.

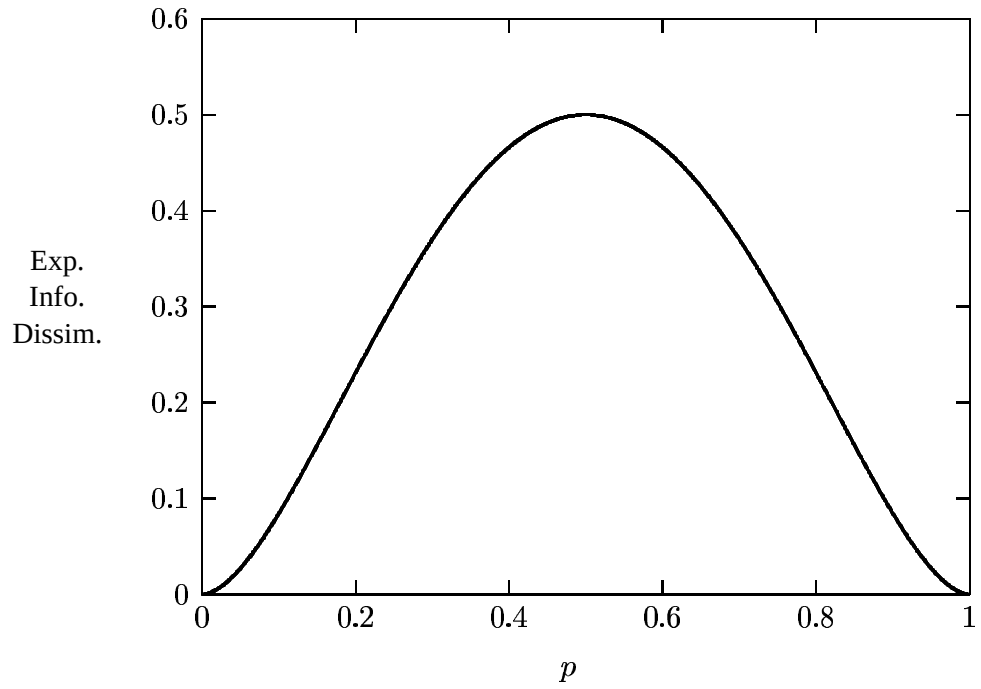


Table 3: Comparison of ordering features from left to right by decreasing expected information dissimilarity and increasing expected information dissimilarity. Ordering by decreasing expected information dissimilarity results in a lower average distance between documents, i.e. adjacent documents are more similar when ordered based on decreasing expected information dissimilarity.

Decreasing Exp. Info. Dis.		Increasing Exp. Info. Dis.	
Document Profile	Information Dissimilarity	Document Profile	Information Dissimilarity
0 0	–	0 0	–
0 0	0.00	0 0	0.00
0 0	0.00	0 0	0.00
0 1	0.81	0 1	1.00
1 1	1.00	0 1	0.00
1 0	0.81	0 1	0.00
1 0	0.00	1 1	.81
1 0	0.00	1 0	1.00
Totals:	2.62		2.81
Averages:	.374		.401

7 Measuring Classifier Performance

The Hamming distance is a suitable subject-distance measure of the dissimilarity between two documents. It can be used to examine the order of a set of classified documents by averaging the Hamming distances between each document and the document to its right, examining each pair of adjacent documents once.

Other measures of the quality of classifier performance may be used to compare classification system performance. For example, one might compute the difference between the two documents using the Shannon measure of information as in the dissimilarity measures discussed earlier.

Both the Hamming distance and the information dissimilarity measures provide values that may be difficult to interpret. A classifier quality (Q) measure is proposed here that is more helpful in evaluating classifier performance. It is normed so that Q is expected to lie in the range from 0 to 1 for an effective classification system. Q is expected to equal 0 when documents are randomly classified, i.e., unclassified, and is expected to equal 1 when documents are perfectly ordered. A negative value is obtained when a *perverse classification system* is used, where performance is worse than would be obtained by randomly arranging documents. A formula yielding this range of values is

$$Q = \frac{E_u - M}{E_u - E_o},$$

where M is the measure of the average dissimilarity for a set of documents, E_u is the expected dissimilarity for an unordered set of documents, where documents are randomly ordered, and E_o is the expected dissimilarity for a perfectly ordered set of documents. Note that M will be greater than E_u in cases when better document ordering is obtained by shelving documents in random locations than by using the classification system, while M is less than E_u, and may approach E_o, when the classification system orders documents so that the subject distance between documents is less than would be obtained if documents were randomly ordered. M, E_u, and E_o are computed in terms of either Hamming distances or information dissimilarity, depending upon need.

E_u, the average expected distance between randomly ordered documents, can be computed from the probabilities that each feature occurs in a document, assuming statistical independence of features. The probability of a change is computed as the probability of a document being selected having a binary feature (p) multiplied by the probability of the neighboring document not having the feature $(1-p)$ added to the probability of a document not having a feature $(1-p)$ times the probability of its neighbor having the feature (p), or $2p(1-p)$. The expected Hamming

distance between two document is the sum of this value over n features, numbered 1 through n,

$$2\sum_{i=1}^{n} p_i(1-p_i).$$

Consider four document profiles, 010, 110, 000, and 100, with features numbered from left to right 1, 2, and 3. Assume that $p_1 = .5$, $p_2 = .25$, and p_3 approaches 0. It can be expected that one half of randomly selected document pairs will differ in the first feature present in the documents. Almost no pairs will be expected to differ in their values at the third feature. The expected Hamming distance for this unordered document set is $.5 + .375 + 0 = .875$.

E_o becomes very small given a good classification system. We treat it here as though it were 0. This allows us to use the following simplified measure:

$$Q = 1 - \frac{M}{E_u}.$$

8 Analyzing an Existing Classification System

A set of experiments was conducted to test the efficacy of the proposed classification method. The proposed classification system has been used to classify five sets of documents. Several measures of document closeness were then used to compare the quality of the proposed classification system to the quality of the Library of Congress classification system. The results suggest that the proposed system can be efficiently used as a classification system for documents. These results should not be understood as claiming superiority of the proposed method over the LCC; they merely suggest that the proposed system can be used to classify documents with some degree of success, especially in environments where classification of documents by computer is useful or necessary.

The five databases used consist of bibliographic records retrieved from searches of the UNC-Chapel Hill Library's on-line catalog. Four arbitrarily chosen subject headings were used to conduct searches for all bibliographic items included in the UNC-CH bibliographic database assigned these Library of Congress Subject Headings (LCSH): (1) History, Ancient, (2) Information Science, (3) Rationalism, and (4) Personality Tests. The four sets of retrieved records constitute four databases. The fifth database is the union of the four other databases. Those few documents with non-LCC numbers, such as unclassified local theses, dissertations, and microforms, were excluded from the databases. For the purposes of the experiments conducted here, each document may be understood as consisting of an LC call number assigned or accepted by UNC-CH catalogers and a set of features, with the features here being equated to the presence or absence of the LCSHs assigned to

Table 4: Comparison of inter-document distances in LC and Gray code based classification systems.

	Database	*Hamming Distance*	
Data Set	*Size*	*LC*	*Gray*
Pers. Tests	60	2.017	1.729
Hist., Anc.	111	2.127	1.791
Info. Sci.	82	2.691	1.951
Rationalism	135	4.134	3.560
combined	388	3.225	2.457

Table 5: Measuring classification performance of LC and Gray code based classification systems with Q measure computed from expected Hamming distance.

		Q Measure	
Data Set	E_u	*LC*	*Proposed*
Pers. Tests	2.258	.11	.23
Hist., Anc.	2.582	.18	.31
Info. Sci.	2.814	.04	.31
Rationalism	4.460	.07	.20
combined	4.750	.32	.48

the document in the UNC-CH catalog. Other, more difficult to obtain subject bearing features were not used in these analyses. Probabilities of features occurring are computed as the percentage of documents in that particular database containing that particular LCSH. Note that documents were likely cataloged at different times and may have been assigned both LCC numbers and LCSHs under those sets of procedures applicable at the time the work was cataloged.

The average Hamming distances between adjacent documents in the different databases are shown in Table 4, both with the LCC system and with the proposed system. The Hamming distance between adjacent documents is computed in these cases as the number of LCSHs by which the adjacent documents differ. For example, if two documents had identical headings except that one had the heading *Epistemics* and the other the heading *General semantics*, the Hamming distance would be 2.

Documents for the LC ordering were sorted by their full LC call number, including institution dependent information such as Cutter numbers, some of which are non-subject bearing. Documents were ordered for the proposed system in order

Table 6: Order of features as suggested by theory, in the reverse order of that suggested by the theory, and in alphabetical order. Q measure based on Hamming distance and the information dissimilarity measures.

	Hamming Distance			*Information Dissimilarity*		
Data Set	*Reverse*	*Alphabetic*	*Theory*	*Reverse*	*Alphabetic*	*Theory*
Pers. Tests	.14	.20	.23	.19	.32	.38
Hist., Anc.	.12	.22	.31	.33	.45	.62
Info. Sci.	.17	.23	.31	.29	.36	.54
Rationalism	.05	.13	.20	.13	.26	.50
combined	.22	.35	.48	.44	.69	.92

of the standard binary values of the feature arrays which are treated as Gray codes. Distances between documents were measured by computing the distances between assigned subject headings being used as features. This consisted of the Hamming distance, the number of features with different values in the two documents. With all five databases, the Hamming distance is lower with the Gray system than with the LCC system. This supports the notion that the proposed classification system places similar documents closer to each other than does the LCC system.

Table 5 provides E_u, the expected classification performance when documents are randomly distributed, and Q measures of the performance of both the LCC system and the Gray system. Both systems increase in classification performance as measured by the Q measure when the proposed classification method is used on each database. The greater increase in performance for the combined database is due in part to the increase in E_u, but is also due to the increased effectiveness of both classification systems when database size increases.

Table 6 shows the effects of both ordering the features as suggested by the theory proposed above and ordering the features in the reverse order of that suggested by the theory. The increase in Hamming distance Q and in the information dissimilarity Q between adjacent documents is greatest when the features are sorted by the relative information of the features as suggested by the theory. This is consistent with the notion that documents with the same or most similar information should be classified closest together.

One difficulty associated with this method of analysis is that distances between documents are computed based solely on their subject headings. These distances are used for both classification and measurement. Results for the proposed Gray code based classification system might exceed the results obtained with other methods, such as the LCC system, primarily because the aspects of the system being

measured are exactly those chosen for optimization.

9 Discussion

The proposed classification system, consistent with a set of requirements for a classification system, can be used effectively to classify documents in library and database environments. It also provides a document ordering that can be used strictly for comparative purposes when evaluating other classification techniques. It has several advantages over conventional document classification systems, such as the LCC system, the DDC system, or the Universal Decimal Classification system. Based on theoretical considerations, it has been shown to provide satisfactory document classification in an experimental environment comparable to that obtained with the LCC system. It is also objective, that is, given a set of document features, any classifier would assign the same document classification. The same concerns that drive the proposed classification system have been used to suggest a measure of classification performance. In the experimental tests described above, the proposed method outperforms the LCC system using this proposed measure as the evaluative criteria.

The proposed classification system also provides a theoretical justification for the construction of classification numbers, suggesting that the components be arranged from left to right in order of increasing specificity. A classification system's strengths and weaknesses using this system may be studied by noting the differences between a number in this theoretically based classification system and numbers in an existing system such as the Dewey Decimal Classification system or the LCC system. When directly comparing existing systems with the Gray system, it will be necessary to convert portions of the large binary numbers generated by the proposed system into decimal or character form to allow for comparison. If the first character of an existing classification code has one of twenty six possible characters, the first twenty six binary feature values of the proposed system could be extracted and treated as a single character for comparison. This assumes that the first twenty six binary features in the proposed system never co-occur, that is, they are mutually exclusive.

Several questions remain unanswered about the proposed scientifically based classification system. One is whether the inclusion of large number of features, on the order of tens of thousands, makes the proposed system perform differently than is evidenced by the experiments described earlier. The classification performance with large number of documents also remains to be established; this performance information is needed before a library or database with tens of thousands of documents should consider using such a classification system.

The measures proposed above have been useful for the tests described here, but an experiment comparing the satisfaction of users with real information needs browsing through actual collections would be informative about the relationship between user behavior and needs and the classification system. In addition, browsing in both libraries and hypertext is usually considered to be the domain of subject searches. The proposed system can incorporate both classification by subject and by incorporation of non-subject related bibliographic features, allowing for the extension of browsing beyond subject searching. The usefulness of this form of browsing needs to be empirically examined.

The adaptability of a classification system to new subjects is an important factor to consider when selecting a classification system. The addition and deletion of subjects to the proposed system is simple in theory but may be costly in practice; it may necessitate remarking many documents, a task to be avoided in manual systems [16]. New features are inserted into their proper location in the list of features so they are arranged in order of increasing rarity. This simple action, however, can change the classification number of many documents! This can be remedied by arbitrarily treating any new feature as a rare feature and placing it furthest to the right in the list of features. Usually a new feature, e.g., AIDS, will initially be extremely rare in the literature and thus the *ad hoc* classification technique of placing new features furthest to the right will be consistent with what the theory suggests. However, as these new features become increasingly common in the literature and the theoretically suggested positions shift to the left, the classification system using the *ad hoc* rule will move farther from meeting the requirements of the classification system.

One could ignore new features until a decision is made to reclassify all documents after a period of time. Before reclassification, added documents could be classified by considering only those features already included in the classification system. In the proposed system, documents will then be placed near the most similar documents. This is what most existing classification systems do with documents on new aspects of existing topics.

This problem is certainly not unique to the proposed system. One of the advantages of using a theoretically based model for a classification system is that classification performance can be analyzed. A librarian faced with the problem of deciding when to relabel a set of documents using the proposed scheme can analytically determine how bad the current system performs, with new features omitted or with new features arbitrarily assigned to the rightmost positions. The performance after the reclassification takes place may be computed and used for comparison.

In summary, it is theoretically and practically possible to develop a usable classification system grounded in the scientific concerns expressed here. This classification system can organize documents in a theoretically justified way without

human intervention. This article has discussed some of the characteristics of such a classification system for linearly arranged documents that is capable of grouping similar documents for browsing.

References

[1] Michael R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.

[2] Sharon L. Baker. Overload, browsers, and selections. *Library and Information Science Research,* 8(4):315–329, 1986.

[3] Abraham Bookstein. Information retrieval: A sequential learning process. *Journal of the American Society for Information Science,* 34(4):331–342, September 1983.

[4] Harold Borko. Research in computer based classification systems. In *Theory of Subject Analysis: A Sourcebook,* pages 287–305. Libraries Unlimited, Littleton, Colo., 1985.

[5] J. F. Cover and B. C. Walsh. Online text retrieval via browsing. *Information Processing and Management,* 24(1):31–37, 1988.

[6] Roy Davies. The creation of new knowledge by information retrieval and classification. *Journal of Documentation,* 45(4):273–301, December 1989.

[7] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6):391–407, September 1990.

[8] K. M. Drabenstott, A. N. Demeyer, J. Gerckens, and D. T. Poe. Analysis of a bibliographic database enhanced with a library classification. *Library Resources & Technical Services,* 34(2):179–198, April 1990.

[9] Peter Evans. Browsing as an information seeking technique. Unpublished Report, U. of Maryland, College of Library and Information Services, 1990.

[10] Christos Faloutsos. Gray codes for partial match and range queries. *IEEE Transactions on Software Engineering,* 14(10):1381–1393, October 1988.

[11] Ivan Flores. Reflected number systems. *IRE Transactions on Electronic Computers,* EC-5(2):79–82, June 1956.

[12] Antony Charles Foskett. *The Subject Approach to Information.* Bingley, London, fourth edition, 1982.

[13] E.A. Fox, Q. F. Chen, A. M. Daoud, and L. S. Heath. Order preserving minimal perfect hash functions and information retrieval. In *ACM Annual Conference on Research and Development in Information Retrieval*, pages 279–312, New York, 1990. ACM Press.

[14] E. N. Gilbert. Gray codes and paths on the n-cube. *Bell System Technical Journal*, 37:815–826, May 1958.

[15] Richard Hamming. *Coding and Information Theory.* Prentice-Hall, Englewood Cliffs, N.J., second edition, 1986.

[16] R. P. Holley. Classification in the USA. *International Classification*, 13(2):73–78, 1986.

[17] J. C. Huestis. Clustering LC classification numbers in an online catalog for improved browsability. *Information Technology and Libraries*, 7(4):381–393, 1988.

[18] Karen Sparck Jones. Notes and references on early automatic classification work. *SIGIR Forum*, 25(1):10–17, Spring 1991.

[19] Ray R. Larson. Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly*, 61(2):133–173, April 1991.

[20] Robert M. Losee. Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, 39(1):8–16, January 1988.

[21] Robert M. Losee. *The Science of Information: Measurement and Applications.* Academic Press, San Diego, 1990.

[22] G. Marchionini. An invitation to browse. *Canadian Journal of Information Science*, 12(3/4):69–79, 1987.

[23] Philip M. Morse. Search theory and browsing. *Library Quarterly*, 40(4):391–408, 1970.

[24] Stephen E. Robertson, C. J. Van Rijsbergen, and M.F. Porter. Probabilistic models of indexing and searching. In Robert Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, pages 35–56, London, 1981. Butterworths.

[25] Claude E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Ill., 1949.

[26] Elaine Svenonius. Directions for research in indexing, classification and cataloging. *Library Resources & Technical Services,* 25(1):88–103, 1981.

[27] C.J. Van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[28] Bohdan S. Wynar. *Introduction to Cataloging and Classification.* Libraries Unlimited, Littleton, Colo., seventh edition by Arlene G. Taylor edition, 1985.