1965

Latent class analysis and information retrieval

George Loyd Jensen

Department:

LATENT CLASS ANALYSIS
AND
INFORMATION RETRIEVAL

BY
GEORGE LOYD JENSEN, 1942

A

THESIS

submitted to the faculty of the

UNIVERSITY OF MISSOURI AT ROLLA

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

Rolla, Missouri

1965

Approved by

Ralph E. Lee (advisor) William K. Winters

Ceren J. Miller R. D. Woodworth

ABSTRACT

With the rapid growth of printed information and the development of modern high speed computers more and more interest has been shown for an information retrieval system. One approach to formulating an information retrieval system is based around the latent class analysis model.

The solution to the latent class analysis problem as originally proposed for use in the information retrieval system is very slow and inaccurate.

The writer has formulated a method to increase the speed and accuracy of the solution of the latent class analysis problem with very little change to the original model.

ACKNOWLEDGEMENTS

The author wishes to express his sincere appreciation to Professor William K. Winters for his help in selecting this thesis topic and also for his guidance and help in this study.

The author also extends his appreciation to Professor Ralph E. Lee, Director of the Computer Science Center for all he has done to make this thesis possible.

A special thanks to Kathy Jensen, my wife, for an excellent job of typing this thesis.

TABLE OF CONTENTS

I. INTRODUCTION

Information retrieval may be defined roughly as a procedure to either locate or physically retrieve a document or documents containing information on a given topic with a high degree of reliability.

Information retrieval is one of the newest fields in computing science. Being so new there are many unexplored areas. The research that has been done has not been standardized beyond the point of the effort which has been slanted toward solving the "Library Problem".

Basically, the "Library Problem" is an attempt to locate a document or documents containing information on a given topic with a fairly high degree of reliability. Solving this problem is quite similar to simulating the process a person goes through when he refers to a card catalog in a library. At the present time, this problem is not too pressing. Considering, however, that the volume of printed material for library use doubles every seven to ten years it will be very time consuming and difficult in the near future to find a particular document.

The main trouble in solving the "Library Problem" lies in implementing techniques that will simulate fairly closely the procedure a person goes through while using a card catalog. It is quite easy to simulate logical thinking and trial patterns but very difficult to simulate intuitive thinking and reasoning. For example, a person might know a

synonym or another related term that pertains to the nature of the document from past experience. This extra knowledge might lead to exactly the right document or documents that will best answer the question or furnish the references desired. An information retrieval system, however, is not capable of having past knowledge of any subject, and must use only the specific information it is given. Thus a major problem in information retrieval is getting the best documents from the given information.

Several different information retrieval methods have been suggested and some research done on them to try and solve the "Library Problem". Of these methods only one is built around a complete mathematical model. This model is "Latent Class Analysis", originally developed by Paul F. Lazarsfeld (1)* to assist in analyzing data received from tests given by social psychologists and sociologists. It was later modified by T. W. Anderson (2) to make the notation clearer and the model more flexible. After examining the latent class model, F. B. Baker (3) saw a close similarity between it and the solution to the "Library Problem". He then proposed a method of applying the latent class model to information retrieval, but did not actually apply his idea to a true working situation due to the complexity of solving large non-symmetric matrices

---

*All numbers (x) refer to the bibliography while the numbers (x.y) refer to equations.

for their characteristic values and vectors.

The purpose of this study is to take this method proposed by Baker (3) and change the non-symmetric matrices to symmetric matrices to make a workable system still based on a sound mathematical model quite similar to the latent class model.

II. REVIEW OF LITERATURE

Information retrieval has applications to many other problems other than the "Library Problem". M. F. Maron (4) suggests the following areas have very direct applications and a pressing need for a solution through an information retrieval system.

Consider the U.S. Patent Department where an exhaustive search must be made of all previous patents to see if each idea is truly new or if it is the same or almost the same as some other idea having been already patented.

People in the medical profession are constantly having a greater need for an information retrieval system. As more is learned about cures, diagnoses and symptoms along with the increase of diseases, a physician might find it helpful to use an information retrieval system. The physician could enter symptoms of a patient into this system and have retrieved all past cases with the same symptoms. From this many diseases could be diagnosed quickly and with more accuracy.

Lawyers, criminologists and large employment placement services have similar problems as the Patent Department and physicians. The entire problem can generally be reduced to finding a few items out of a large file of items which will help solve your present problem.

How should the large file of items be organized so an

accurate and thorough search may be made quickly? This question was answered in general by five premises by Joseph Becker and Robert M. Hayes (5).

The first premise is that the file should contain not just the stored document but also some representation of past requests, requesters and the area of use of this request. The file should also be easily updated by the addition of new documents or new information and by deleting outdated documents and information.

The second premise is that a file should appear homogeneous. This refers to the point that there is no qualitative difference between a document, an abstract referring to that document, or a request. The difference between items in the file should be quantitative. From this premise the difference between "fact retrieval" and "document retrieval" is seen to be a quantitative difference between various depths of response rather than a fundamental difference in the character of the file.

The third premise is that a quantitative model and measure for relevancy is possible. This premise utilizes a mathematical model based on quantified measurement. Becker and Hayes (5) state that at the present time there is not enough sufficient data on which to base a valid model, but several significant steps have been made in this direction.

The fourth premise is that there is no essential relationship between the method of representing an item and the organization of groups of items into a file. The way items

are represented is mainly part of the logical problems of information retrieval. The way items are organized is part of the physical problem. It is assumed that a person would organize items in a large file even if there existed a machine that would decide instantaneously whether a given document was relevant or not to the request. The reason for this assumption is because the machine would have to first access the document. The items would be organized in this case solely to ease the job of access.

The fifth premise is based on the fact that certain items will generally be wanted together. These items then lose their individual identity and assume a group identity. This premise may also be extended to whether items are likely to be wanted or not. This is essentially locating the items most likely to be wanted in an essentially accessible place and those least wanted in an inaccessible place.

These five premises are primarily guides in determining what is meant by a file item, how to represent such an item quantitatively, and how to arrange a file.

The Cambridge Language Research Unit did considerable research regarding the relevancy of two columns of binary numbers. The two columns were defined to be A_i and A_j. Three ways were proposed to find the nearness function of the two columns, A_i and A_j. The first method is to let the nearness function = $N(A_i \cdot A_j)$ = the number of 1's in both columns of A_i and A_j. This definition is suitable if the number of 1's in a column is fixed or only slightly varied.

An example to show this restriction is that the nearness of 1100111 to 1111000 is the same as that of 1110000 to 1101110. The second method is $N(A_i \cdot A_j) + N(\overline{A}_i \cdot \overline{A}_j)$ = the number of places of agreement between 1's and between 0's. This is the number of places where the two columns agree. It avoids the trouble of the first method but it is also ineffectual if the ratio of 1's to 0's is removed far from 1 since the same weight is given to an agreement in a 0 as in a 1. The third method is

$$\frac{N(A_i \cdot A_j)}{N(A_i + A_j)} = \frac{\text{number of 1's in both rows}}{\text{number of 1's in either row}}$$

Since this method attaches no weight to the agreement in 0's it is suitable only when the proportion of 1's to 0's is low. This is the most obvious definition when the columns are of an indefinite length but the number of 1's is statistically limited or fixed.

M. E. Maron and J. L. Kuhns (6) developed the method of "Probabilistic Indexing". A set of words, called key words were used in obtaining the binary vectors A_i and A_j. Key words are words which are decided upon to be used as an indication of the contents of a document. By placing a set of key words in a vector and then comparing the D_i document with these words a binary vector may be obtained. If the document pertains to a certain key word a 1 is placed in that position of the vector A_i if it does not pertain, then a 0 is placed in that position. Having established

the vectors A_i and A_j the following two-way table is set up:

	A_j	$\overline{A}_j$	
A_i	$x = N(A_i \cdot A_j)$	$u = N(A_i \cdot \overline{A}_j)$	$N(A_i)$
$\overline{A}_i$	$v = N(\overline{A}_i \cdot A_j)$	$y = N(\overline{A}_i \cdot \overline{A}_j)$	$N(\overline{A}_i)$
	$N(A_j)$	$N(\overline{A}_j)$	n

$N(A_i \cdot A_j)$ corresponds exactly to method 1 of the Cambridge Language Research Unit. $N(A_i \cdot \overline{A}_j)$ is the number of positions in the vectors A_i and A_j that have a 1 in A_i and a 0 in A_j. Similarly $N(\overline{A}_i \cdot A_j)$ is the number of 0 and 1 combinations in A_i and A_j respectively at the same position. $N(\overline{A}_i \cdot \overline{A}_j)$ corresponds to the number of common zeros. To find a measure of association between the two binary vectors of A_i and A_j it is necessary to find a term $A(A_i , A_j)$ that has: a maximum when A_i is contained in A_j (u=0) or A_j is contained in A_i (v=0) or A_i and A_j give the same class (u=v=0); a minimum when A_j is contained in $\overline{A}_i$ (x=0) or $\overline{A}_i$ is contained in A_j (y=0) or A_i is the compliment of A_j (x=y=0); and a range of values from -1 to 1. Such an equation is :

$$A(A_i , A_j) = \frac{xy - uv}{(xy + uv)}$$

If A_j is the binary vector that is being searched for then $A(A_i , A_j)$ gives a measure of association between the document D_i and the given request. By setting a minimum value for $A(A_i , A_j)$ and retrieving all of the documents, D_i, which have a larger measure of association than the set minimum, the retrieved documents should have a high degree of relevance to the given key word vector. Maron and Kuhns (6) tried their method and found it gave very good results.

H. Edmund Stiles (7) suggested another approach to information retrieval. He suggested and put into operation the association factor. The first step is to develop a list of terms arranged according to their degree of association with a given term. Frequency alone was found to be very unsatisfactory. After considering several formulas,

$$\log_{10} \frac{\left[|fN-AB| - \frac{N}{2} \right]^2 N}{AB\ (N-A)(N-B)} = \text{Association Factor}$$

was decided upon. This formula is a form of the chi-square formula. In this formula,

A = the number of documents indexed by one term

B = the number of documents indexed by a second term

f = the number of documents indexed by the combination of both terms

N = the total number of documents in the collection

Terms that have an association factor of less than one are generally discarded due to their small association and to reduce the size of the term list. The association factor tends to give a higher numerical value to the key index words under the investigated headings that are most likely to be sought by someone using this information retrieval method. The key index words that remain after all of those with an association factor of less than one have been removed are called a "term profile" for this particular heading.

The second step is to find key index words that appear in a certain percent or all of the term profiles if a multi-term key word selection is used. These will be referred to as "first generation terms". If a single key word selection is used, all of the members of the term profile will be first generation terms.

The third step is to take the first generation terms and treat them as a multi-term key word selection and repeat the first two steps. This will yield "second generation terms", among which will be terms that are closely related such as synonyms.

The fourth step is to make a table of association factors for the expanded list of requested terms. The sum of the association factors for each term is called its "weight". This weight indicates the degree of association between that term and the complete request.

The last step is to compare the list of expanded requested terms with the index terms for each document in the collection and add the weights of the terms that match. The sum of the weights is called the "document relevance number". This number is used to present the documents to the requester in the order of their probable relevance to the request.

According to tests run using the association factor, the results were very good. The documents which were predicted to have the highest relevance did have a direct bearing to the original requested terms.

Harold Borko and Myrna Bernick (8) have experimentally studied the applicability of factor analysis to the generation of categories. They chose a group of 405 abstracts of computer literature to work with and 90 key words. A matrix was constructed so that the A_{ij} element equaled the number of times the j-th key word appeared in the i-th document. Based on the frequencies in the data matrix the correlation coefficients were computed for each of the 90 index terms and correlated with each of the other terms. This resulted in a 90 x 90 correlation matrix. The matrix was then analyzed by means of factor analysis. Factor analysis is a technique to reduce the original correlation matrix to a smaller number of factors or eigenvectors. In their experiment, Borko and Bernick (8) after trying values between 15 and 32 decided on 21 factors. These factors were general group headings under which the key words were to be

grouped. The documents were then placed automatically into the 21 groups. The formula that was used to do this was:

$$P = (L_1 \times T_1 + L_2 \times T_2 + \ldots + L_n \times T_n)$$

where

P = the predicted classification

T_n = the number of occurrences of the n term

L_n = the normalized factor loading of term n for a given catagory

After automatically classifying the documents they were classified by humans who did not know the results of the automatic classification. The automatic classification agreed with the human classification about sixty-five percent of the time in one study and just under fifty percent in another. Although these results are not as good as had been hoped for by Borko and Bernick (8) it does provide a good starting place for document classification.

Frank B. Baker (3) approached information retrieval by using as a basic model for his work, "Latent Class Analysis". Latent class analysis was originated by Lazarsfeld (1) and was later improved by Anderson (2). It was originally designed to analyze the responses a person gave to questions on a sociologist test. Latent class analysis is based on a mathematical model which assumes the population may be classified into homogeneous

sets. Each set has independent responses for the different items. The test will contain a certain number, k, of questions or items. Each of the k items may have either a positive or negative response.

Let the probability that an individual in the population will give a positive response on the i-th item be designated by π_i (or $\pi_{\bar{i}}$ if a negative response is desired). π_{ij} means a positive response to the i-th and j-th item by an individual. This may continue to $\pi_{1,2,\ldots,k-1,k}$ if a positive response was desired from each of the k items from an individual. This indicates that there are 2^k possible response patterns for each individual.

It is assumed that the population could be divided into n mutually exclusive subpopulations or classes. Let ν^α (α=1,2,...m) be the probability that a person drawn at random is from the α subpopulation. Let λ_i^α represent the probability that a person in the α class responds positively to the i-th item, ($\lambda_{\bar{i}}^\alpha$ is a negative response), then

$$\pi_i = \sum_{\alpha=1}^{m} \nu^\alpha \lambda_i^\alpha .$$

Similarly,

$$\begin{aligned}
\pi_{ij} &= \sum_{\alpha=1}^{m} \nu^\alpha \lambda_i^\alpha \lambda_j^\alpha && i \neq j \\
\pi_{ijk} &= \sum_{\alpha=1}^{m} \nu^\alpha \lambda_i^\alpha \lambda_j^\alpha \lambda_k^\alpha && i \neq j,\ j \neq k,\ k \neq i \\
\pi_{\bar{i}jk} &= \sum_{\alpha=1}^{m} \nu^\alpha \lambda_{\bar{i}}^\alpha \lambda_j^\alpha \lambda_k^\alpha && i \neq j,\ j \neq k,\ k \neq i \qquad (2.01)
\end{aligned}$$

are known as the accounting equations. For convenience, $\lambda_i^\alpha \lambda_j^\alpha \lambda_k^\alpha$ will be written as λ_{ijk}^α, for example:

$$\pi_{ijk} = \sum_{\alpha=1}^{m} \nu^\alpha \lambda_{ijk}^\alpha \quad i \neq j,\ j \neq k,\ k \neq i.$$

The class to which an individual belongs can not be observed directly. The class may only be inferred by the response pattern which the individual gives. In order to make this inference, the investigator must use the ν's and λ's of the latent structure together with the observed response pattern of each individual. The investigator may, however, determine the π's which are the underlying probabilities of response by an infinitely large sample. The value of π_i may be calculated by counting the number of positive responses to the i-th item and dividing by the number of questionaires in the sample. π_{ij} may be calculated by counting the number of positive responses to both items i and j by one person then dividing by the total number of questionaires in the sample. This introduces the problem: how does one find the ν's and λ's given the values of π for all possible subscript combinations?

Let

$$\Lambda = \begin{bmatrix} 1 & \lambda_1^1 & \lambda_2^1 \ldots \lambda_k^1 \\ 1 & \lambda_1^2 & \lambda_2^2 \ldots \lambda_k^2 \\ \ldots & & \\ 1 & \lambda_1^m & \lambda_2^m \ldots \lambda_k^m \end{bmatrix} \qquad (2.02)$$

be the matrix of latent parameters. The elements in the first column of this matrix, λ_0^α , are considered as dummy items and given a value of one. The superscripts which refer to the rows designate the latent class; the subscripts which refer to the columns designate the item.

Define Λ_1 as an m x m matrix consisting of the first m columns of Λ. Define Λ_2 as an m x m matrix consisting of the first column and the next m-1 columns of Λ not used by Λ_1, that is column 0 and columns m to 2m - 2.

Let

$$N = \begin{bmatrix} \nu^1 & 0 & 0 \ldots 0 \\ 0 & \nu^2 & 0 \ldots 0 \\ 0 & 0 & \nu^3 \ldots 0 \\ \ldots & & \\ 0 & 0 & 0 \ldots \nu^m \end{bmatrix} \qquad (2.03)$$

and

$$\Delta = \begin{bmatrix} \lambda_k^1 & 0 & 0 \ldots 0 \\ 0 & \lambda_k^2 & 0 \ldots 0 \\ 0 & 0 & \lambda_k^3 \ldots 0 \\ \ldots & & \\ 0 & 0 & 0 \ldots \lambda_k^m \end{bmatrix} \qquad (2.04)$$

where $k > 2m - 2$.

Since the initial number of the items is arbitrary it is seen that two sets of m-1 items and one additional item, k, have been selected from all of the items. This implies that there must exist at least $2m - 1$ items.

To solve for the selected ν's and λ's the accounting equations π_{ijk} and π_{ij} will be used where $i = 0,1,2,\ldots m-1$ and $j=0,m,m+1,\ldots 2m-2$. When a subscript is zero it is omitted, for example, $\pi_{i0k} = \pi_{ik}$, $\pi_{0jk} = \pi_{jk}$, $\pi_{000}=\pi_0= 1$.

The $\prod$ matrix is formed to be:

$$\prod = \begin{bmatrix} \pi_{0,0,k} & \pi_{0,m,k} & \pi_{0,m+1,k} & \ldots & \pi_{0,2m-2,k} \\ \pi_{1,0,k} & \pi_{1,m,k} & \pi_{1,m+1,k} & \ldots & \pi_{1,2m-2,k} \\ \pi_{2,0,k} & \pi_{2,m,k} & \pi_{2,m+1,k} & \ldots & \pi_{2.2m-2,k} \\ \ldots & & & & \\ \pi_{m-1,0,k} & \pi_{m-1,m,k} & \pi_{m-1,m+1,k} & \ldots & \pi_{m-1,2m-2,k} \end{bmatrix}$$

$$= \begin{bmatrix} \pi_{k} & \pi_{m,k} & \pi_{m+1,k} & \ldots & \pi_{2m-2.k} \\ \pi_{1,k} & \pi_{1,m,k} & \pi_{1,m+1,k} & \ldots & \pi_{1,2m-2,k} \\ \pi_{2,k} & \pi_{2,m,k} & \pi_{2,m+1,k} & \ldots & \pi_{2,2m-2,k} \\ \ldots & & & & \\ \pi_{m-1,k} & \pi_{m-1,m,k} & \pi_{m-1,m+1,k} & \ldots & \pi_{m-1,2m-2,k} \end{bmatrix} \qquad (2.05)$$

and each element has the subscript k. The first subscript and the second subscript (when the subscript is not surpressed) refer respectively to the items of Λ_1 and Λ_2. Define also a matrix $\prod^*$, the same as $\prod$ except π_{ijk} is replaced by π_{ij}.

By matrix multiplication it can be shown that

$$\prod = \Lambda_1^T N \Delta \Lambda_2 \qquad (2.06)$$

$$\prod^* = \Lambda_1^T N \Lambda_2 \qquad (2.07)$$

which corresponds to the elements of π_{ij} and π_{ijk} according to the accounting equations (2.01).

The latent roots, or eigenvalues of these matrices may be obtained through the determinental equation

$$|\prod - \Theta \prod^*| = 0.$$

This equation may be expanded to:

$$\begin{aligned} |\prod - \Theta \prod^*| &= |\Lambda_1^T N \Delta \Lambda_2 - \Theta \Lambda_1^T N \Lambda_2| \\ &= |\Lambda_1^T N| \cdot |\Delta - \Theta I| \cdot |\Lambda_2 \\ &= |\Lambda_1^T| \cdot |N| \cdot |\Delta - \Theta I| \cdot |\Lambda_2| \end{aligned} \qquad (2.08)$$

To obtain significant results it must be defined that Λ_1, N, Λ_2 are non-singular which implies, since N is a diagonal matrix that each v^α is non-zero. Since Δ is a diagonal matrix the latent roots are therefore the elements of Δ which are $\lambda_k^1, \lambda_k^2, \ldots \lambda_k^m$. For each eigenvalue ($\Theta^\alpha$) a corresponding eigenvector may be found that satisfies

$\left[\prod - \theta^{\alpha} \prod^{*}\right] x^{\alpha} = 0$. By arranging these vectors in columns, a matrix $X = (x^1, x^2, \ldots x^n)$ is formed. Let θ^{α} be the α diagonal element of a matrix Θ. Then

$$\prod X = \prod^{*} X \Theta \tag{2.09}$$

Suppose the θ's in Θ are ordered so that $\Theta = \Delta$ then

$$\prod X = \prod^{*} X \Delta$$

or

$$\Lambda_1^T N \Delta \Lambda_2 X = \Lambda_1^T N \Lambda_2 X \Delta \quad . \tag{2.10}$$

A possible solution then would be $X = \Lambda_2^{-1}$. More generally $X = \Lambda_2^{-1} E_x$ where E_x is a diagonal matrix to account for a multiplication factor for each row. Given a solution X, Λ_2 may be obtained by $\Lambda_2 = E_x X^{-1}$. Since every element in column 1 of Λ_2 is 1 by definition, the diagonal matrix elements of the E_x must be the reciprocal of the elements in the first column of the corresponding row of X^{-1}.

A second set of vectors (y^{α}) may be found which satisfy $(y^{\alpha})^T \left[\prod - \theta^{\alpha} \prod^{*}\right] = 0$ or $\prod^T y_{\alpha} = \theta^{\alpha} \prod^{*T} y_{\alpha}$ by a similar argument as for the X matrix a Y matrix may be found and shown that $\Lambda_1 = E_y Y^{-1}$.

By knowing the elements of Λ_1, Λ_2 and Δ, N may be found in the following manner:

$$(\Lambda_1^T)^{-1} \prod^{*} \Lambda_2^{-1} = (\Lambda_1^T)^{-1} \Lambda_1^T N \Lambda_2 \Lambda_2^{-1} = INI = N \tag{2.11}$$

By having these values the problem is solved, because all values of ν and λ will be known.

III. DISCUSSION

One of the main purposes of this study is to formulate the methods suggested by Anderson (2) and Baker (3) for solving the latent class analysis problem. Solving the latent class analysis problem is primarily solving the accounting equations (2.01). The accounting equations may be arranged to form matrices $\prod$ and $\prod^*$ where $\prod$ and $\prod^*$ are defined by equations (2.06) and (2.07) respectively. $\prod$ and $\prod^*$ are neither defined to be nor would in general be symmetric. Because the solution involves finding the eigenvalues and eigenvectors of equation (2.08), it would be desirable to have $\prod$ and $\prod^*$ symmetric. This is due to the fact that most methods available for finding the eigenvalues and eigenvectors of large systems of non-symmetric matrices are considerably more difficult and not as accurate as methods for symmetric matrices.

The method Anderson (2) suggested was to expand a polynomial of degree m and find the roots of the equation. Then by knowing the roots, a method of co-factors may be used to find the eigenvectors. This method is not only very slow but a small error introduced at any point in the calculations is greatly expanded through succeeding steps. To make the model more suitable for calculations the $\prod$ and $\prod^*$ matrices were changed so they would be symmetric. This was accomplished by setting $\bar{\Lambda}_1 = \bar{\Lambda}_2$. This will improve the model by giving an inner relationship between each element in Λ_1 to itself

when the $\overline{\prod}$ and $\overline{\prod}^*$ matrices are computed. $\overline{\Lambda}$, $\overline{\Lambda}_1$ may be defined by the following equation:

$$\overline{\Lambda} = \begin{bmatrix} 1 & \lambda_1^1 & \lambda_2^1 & \cdots & \lambda_k^1 \\ 1 & \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_k^2 \\ \vdots & \vdots & \vdots & & \\ 1 & \lambda_1^m & \lambda_2^m & \cdots & \lambda_k^m \end{bmatrix} \qquad (3.01)$$

$$\overline{\Lambda}_1 = \overline{\Lambda}_2 = \begin{bmatrix} 1 & \lambda_1^1 & \lambda_2^1 & \cdots & \lambda_{m-1}^1 \\ 1 & \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_{m-1}^2 \\ \vdots & \vdots & \vdots & & \\ 1 & \lambda_1^m & \lambda_2^m & \cdots & \lambda_{m-1}^m \end{bmatrix} \qquad (3.02)$$

Thus K > m-1 which implies there must be at least as many items as there are classes. Let Δ and N be defined as they were previously in equations (2.03)and (2.04). $\overline{\prod}$ and $\overline{\prod}^*$ may be defined as:

$$\overline{\prod} = \overline{\Lambda}_1^T \, N \, \Delta \, \overline{\Lambda}_2 \quad \text{and} \quad \overline{\prod}^* = \overline{\Lambda}_1^T \, N \, \overline{\Lambda}_2$$

since $\overline{\Lambda}_1 = \overline{\Lambda}_2$

then $\overline{\prod} = \overline{\Lambda}_1^T \, N \, \Delta \, \overline{\Lambda}_1 \qquad (3.03)$

and $\overline{\prod}^{*} = \overline{\Lambda}_1^{T} \; N \; \overline{\Lambda}_1$. (3.04)

After the matrix multiplication of equations (3.03) and (3.04) is carried out $\overline{\prod}$ and $\overline{\prod}^{*}$ take the form:

$$\overline{\prod} = \begin{bmatrix} \pi_{0,0,k} & \pi_{0,1,k} & \pi_{0,2,k} & \cdots & \pi_{0,m-1,k} \\ \pi_{1,0,k} & \pi_{1,1,k} & \pi_{1,2,k} & \cdots & \pi_{1,m-1,k} \\ \pi_{2,0,k} & \pi_{2,1,k} & \pi_{2,2,k} & \cdots & \pi_{2,m-1,k} \\ \vdots & & & & \\ \pi_{m-1,0,k} & \pi_{m-1,1,k} & \pi_{m-1,2,k} & \cdots & \pi_{m-1,m-1,k} \end{bmatrix}$$

$$= \begin{bmatrix} \pi_{k} & \pi_{1,k} & \pi_{2,k} & \cdots & \pi_{m-1,k} \\ \pi_{1,k} & \pi_{1,1,k} & \pi_{1,2,k} & \cdots & \pi_{1,m-1,k} \\ \pi_{2,k} & \pi_{2,1,k} & \pi_{2,2,k} & \cdots & \pi_{2,m-1,k} \\ \vdots & & & & \\ \pi_{m-1,k} & \pi_{m-1,1,k} & \pi_{m-1,2,k} & \cdots & \pi_{m-1,m-1,k} \end{bmatrix} \quad (3.05)$$

$$\bar{\Pi}^{*} = \begin{bmatrix} 1 & \pi_1 & \pi_2 & \cdots & \pi_{m-1} \\ \pi_1 & \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,m-1} \\ \pi_2 & \pi_{2,1} & \pi_{2,2} & \cdots & \pi_{2,m-1} \\ \vdots & & & & \\ \pi_{m-1} & \pi_{m-1,1} & \pi_{m-1,2} & \cdots & \pi_{m-1,m-1} \end{bmatrix} \qquad (3.06)$$

which are easily seen to be symmetric matrices. Since the internal elements π_{ij} and π_{ijk} are formed by the accounting equations (2.01) and

$$\sum_{\alpha=1}^{m} \nu^{\alpha} = 1$$

and all the λ_1^{α}'s are probabilities it follows that $0 < \pi_{ij} \leq 1$ and $0 < \pi_{ijk} \leq 1$. From these it may be concluded that each element is real and positive.

If equation (3.04) is premultiplied by $(\bar{\Lambda}_1^{T})^{-1}$ and post multiplied by $\bar{\Lambda}_1^{-1}$ the following equation is obtained:

$$(\bar{\Lambda}_1^{T})^{-1} \bar{\Pi}^{*} \bar{\Lambda}_1^{-1} = (\bar{\Lambda}_1^{T})^{-1}(\bar{\Lambda}_1^{T})\, N\, \bar{\Lambda}_1\, \Lambda_1^{-1} = I\, N\, I = N \qquad (3.07)$$

N is positive definite by definition since all ν^{α} $(\alpha=1,2,\ldots m) > 0$. Therefore, the quadratic form $(\bar{\Lambda}_1^{-1})^{T} \bar{\Pi}^{*} \bar{\Lambda}_1^{-1}$ is positive definite, thus $\bar{\Pi}^{*}$ is positive definite.

In a like manner, if equation (3.03) is premultiplied by $(\bar{\Lambda}_1^T)^{-1}$ and post multiplied by $(\bar{\Lambda}_1)^{-1}$ it is seen that

$$(\bar{\Lambda}_1^T)^{-1} \bar{\bar{\prod}} \bar{\Lambda}_1^{-1} = N \Delta,$$

where

$$N\Delta = \begin{bmatrix} \nu^1\lambda_k^1 & 0 & 0 & \dots & 0 \\ 0 & \nu^2\lambda_k^2 & 0 & \dots & 0 \\ 0 & 0 & \nu^3\lambda_k^3 & \dots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \dots & \nu^m\lambda_k^m \end{bmatrix} .$$

But by observing that $\nu^\alpha\lambda_k^\alpha > 0$ for all (α=1, 2, ..., m) it is seen that N Δ is positive definite as is the quadratic form $(\bar{\Lambda}_1^{-1})^T \bar{\bar{\prod}} \bar{\Lambda}_1^{-1} = N \Delta$. Therefore, $\bar{\bar{\prod}}$ is positive definite.

Since $\bar{\bar{\prod}}^*$ has been shown to be real symmetric and positive definite, there exists a non-singular matrix P such that $P^T \bar{\bar{\prod}}^* P = I$. Therefore $\bar{\bar{\prod}}^*$ is equivalent to I under congruence operations. Finding a matrix P which satisfies the desired conditions is a process of performing the same operations on the rows as on the columns of (3.08) until the original matrix $\bar{\bar{\prod}}^*$ is transformed to I which yields

the form (3.09).

$$\left[\begin{array}{c|c} \bar{\Pi}^* & I \\ \hline I & \phi \end{array}\right] \tag{3.08}$$

$$\left[\begin{array}{c|c} I & P \\ \hline P^T & \phi \end{array}\right] . \tag{3.09}$$

To reduce the equations

$$\left[\bar{\Pi} - \theta \bar{\Pi}^* \right] X = 0$$

or

$$\bar{\Pi} X = \theta \bar{\Pi}^* X \tag{3.10}$$

to a form where the eigenvalues are readily found is a procedure of matrix multiplication as shown in equation (3.11).

$$(P^T \bar{\Pi} P - \theta P^T \bar{\Pi}^* P) X^* = 0 \tag{3.11}$$

since

$$P^T \bar{\Pi}^* P = I,$$

then

$$(P^T \bar{\Pi} P - \theta I) X^* = 0. \tag{3.12}$$

These operations may be more easily seen by the use of the determinental equations:

$$|\bar{\Pi} - \theta \bar{\Pi}^*| = 0.$$

$$|P^T| \cdot |\bar{\Pi} - \theta \bar{\Pi}^*| \cdot |P| = 0.$$

$$|P^T \bar{\Pi} P - \theta P^T \bar{\Pi}^* P| = 0.$$

Equivalent matrices have the same eigenvalues therefore, the θ's are the desired eigenvalues of equation (3.10). To find the necessary eigenvectors of equation (3.10), equation (3.11) is factored to the form

$$P^T \left[\bar{\Pi} - \theta \bar{\Pi}^*\right] P X^* = 0 \qquad (3.13)$$

and multiplying by the matrix $(P^T)^{-1}$ as follows:

$$(P^T)^{-1}(P^T) \left[\bar{\Pi} - \theta \bar{\Pi}^*\right] P X^* = (P^T)^{-1} \cdot 0 \qquad (3.14)$$

or

$$\left[\bar{\Pi} - \theta \bar{\Pi}^*\right] P X^* = 0 \qquad (3.15)$$

thus

$$X = P X^* . \qquad (3.16)$$

By using a procedure such as the Jacobi method on equation (3.12) the values of θ and X^* are obtained. By equation (3.16) the eigenvalues and eigenvectors of equation (3.10) are known. By using procedures such as the Jacobi method on the modern day computers the values of θ and X^* may be obtained very accurately and quite rapidly. The speed and accuracy of this procedure is much better than the method suggested by Anderson (2) when doing both procedures on the same computer.

Since $\bar{\Lambda}_1 = \bar{\Lambda}_2$ only one set of eigenvectors is needed. As shown in Chapter II, $X = \bar{\Lambda}_2^{-1} E_x$. By matrix multiplication one obtains

$$\bar{\Pi}^* X = (\bar{\Lambda}_1^T N \bar{\Lambda}_1)(\bar{\Lambda}_1^{-1} E_x) = \bar{\Lambda}_1^T N E_x .$$

Since $N E_x$ is a diagonal matrix, and the first row of $\bar{\Lambda}_1^T$ is defined to be all 1's. $\bar{\Lambda}_1^T$ may be found by dividing each column of $\bar{\Pi}^* X$ by the first element in that column. The elements in the first row of $\bar{\Pi}^* X$ are observed to be the diagonal elements of $N E_x$. Then

$$\bar{\Lambda}_1 X = \bar{\Lambda}_1 (\bar{\Lambda}_1^{-1} E_x) = E_x = N^{-1} (N E_x) \qquad (3.17)$$

since $(N E_x)$ is a diagonal matrix. $(N E_x)^{-1}$ may be found by replacing each diagonal element by its reciprocal. By post multiplying both sides of equation (3.17) by $(N E_x)^{-1}$ equation (3.18) is obtained.

$$(\bar{\Lambda}_1 X)(N E_x)^{-1} = N^{-1} \qquad (3.18)$$

Since N is a diagonal matrix N^{-1} must be diagonal and N is found by taking the reciprocal of each of the diagonal elements.

As explained in chapter II, $\Theta = \Delta$. The problem has now been solved since all of the required values may be found in N, $\bar{\Lambda}_1$, and Δ. The above procedure for finding $\bar{\Lambda}_1$, and N was used to keep the error as small as possible. The fewer operations used the less the induced error is,

and matrix inversion takes considerably more operations than matrix multiplication.

A workable model that is easily solved has now been transformed from the original latent class model.

It is now desired to apply the given workable model to information retrieval. It will be assumed that there is a reliable and efficient means available to select key words to identify each document. The presence or absence of key words would place each item, book or article to be retrieved in a category orgroup with other items to be retrieved that have the same response to the selected key words. A convenient way to represent a positive response to a key word is with a "+" and a negative response with a "0". Then by letting the "+" be "1" the given response to the key words would become a binary number formed from the responses to the ordered key words. There are in this case "k" key words or items which implies there would be 2^k possible categories or groups (each separate from the rest) into which numerous articles, books or items could be arranged.

Estimates π_{ij} and π_{ijk}, must be calculated for each value of π_{ij} and π_{ijk} respectively which refer to the probability of drawing from the entire population of books, items and articles a sample that gives a positive(or negative) response to the k-th and/or the i-th and j-th key words. After constructing $\overline{\prod}$ and $\overline{\prod}^*$ the latent class analysis is performed to find the value of all of the ν's and λ's.

Knowing these values the probability that a given response pattern belongs to a particular latent class may be found. This probability is calculated by

$$P_{\alpha} = \frac{\nu^{\alpha} \left[\prod_{i=1}^{k} \lambda_i^{\alpha*}\right]}{\sum_{j=1}^{m} \nu^{j} \left[\prod_{m=1}^{k} \lambda_m^{\alpha*}\right]} \qquad (3.19)$$

where

$$\lambda_i^{\alpha*} = \begin{cases} \lambda_i^{\alpha} & \text{if there is a positive response to the i-th key word in the response pattern being tested} \\ 1-\lambda_i^{\alpha} & \text{if there is a negative response to the i-th key word in the response pattern being tested} \end{cases}$$

The class that has the largest probability is the class to which the response pattern belongs. There are other response patterns that could yield related information to the topic the search is being made on so the retriever may specify what might be called a "relevance level". This would be the probability which any response pattern must have in the chosen class before the documents with that response pattern would be retrieved as relevant documents toward the topic of interest to the user of the system. The higher relevance level, the fewer the number of articles to search through, but the probability would be greater that they pertained to the topic (class being retrieved) of interest.

IV. RESULTS AND CONCLUSIONS

Computer solutions were obtained for solving the latent class problem as proposed by Anderson (2); and to test the latent class problem with symmetric matrices. To test both procedures random numbers were generated between 0 and 1 for all values of v^{α} and λ_i^{α}. The values of π_{ij} and π_{ijk} were calculated as defined by the accounting equations (2.01). Latent class analysis was used to calculate the values of v^{α} and λ_i^{α}. It was found for matrices larger than 6 x 6 the error was greater than .01 for the non-symmetric case. By using symmetric matrices the error was less than .001 for a 20 x 20 matrix. The time required to obtain a solution for the non-symmetric matrix was twice that required by the symmetric case.

Baker (3) only proposed a method of approach to solve the "Library Problem" and did not test his technique. Since he did not actually obtain a solution, he probably was not fully aware of the shortcomings of Anderson's solution to latent class analysis. One of the main problems of Baker's method of information retrieval is the selection of the proper set of key words. There is considerable research presently being done to try and find an efficient method for selecting key words that will properly cover the set of documents. Another problem is the selection of proper values for the probability estimates π_{ij} and π_{ijk}.

Now that a faster and more accurate method is available to solve this problem, Baker's approach to finding a good information retrieval method seems quite feasible. The method as proposed by Baker for an information retrieval system is different from most other methods since it is based on a completely mathematical approach.

The results of this study indicate that the use of latent class analysis with symmetric matrices is a good approach to the development of an information retrieval system.

BIBLIOGRAPHY

1. Lazarsfeld, Paul F., (1950), Measurement and Prediction, Princeton University Press, Princeton, New Jersey, p. 362-472.

2. Anderson, T. W., (1954) On Estimation of Parameters in Latent Structure Analysis, Psychometrica, Vol. 19, No. 1, p. 1-10.

3. Baker, Frank B., (1962) Information Retrieval based upon Latent Class Analysis, Journal of the Association for Computing Machinery, Vol. 9, No. 4, p. 512-521.

4. Maron, M. E., (1961) Information Retrieval: A look at the Logical Framework and some New Concepts, The Rand Corporation, Santa Monica, California, p. 37.

5 Becker, Joseph and R. M. Hayes, (1964) Information Storage and Retrieval: Tools, Elements, Theories, John Wiley and Sons, New York, p. 359-397.

6. Maron, M. E. and J. L. Kuhns, (1960) On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the Association for Computing Machinery, Vol. 7, No. 3, p. 216-244.

7. Stiles, H. Edmund, (1961) The Association Factor in Information Retrieval, Journal of the Association for Computing Machinery, Vol. 10, No. 2, p. 151-162.

8. Borko, Harold and Myrna Bernick, (1962), Automatic Document Classification, Journal of the Association for Computing Machinery, Vol. 10, No. 2, p. 151-162.

VITA

The author was born January 14, 1942 at Independence, Missouri. His primary education was received there, he attended high school in Kansas City, Missouri, graduating in June, 1959. He attended Deep Springs College, Deep Springs, California and received his Bachelor of Science degree in Mathematics from the University of Missouri School of Mines and Metallurgy at Rolla, Missouri in January, 1964.

Since February, 1964 he has been employed as a graduate assistant in the Computer Science Center at the University of Missouri at Rolla, Rolla, Missouri