An Investigation of Automatic Term Weighting Techniques

Hyun-Hee, Kim*

초 록

本 연구는 두 개의 중요한 目的들을 가지고 있다. 첫째 目的은 새로운 單語 加重技法을 고안하는 것이다. 두번째 目的은 제안된 單語 加重技法과 다른 네개의 單語 加重技法들의 문헌검색결과들을 평가하는 것이다. 본 연구에서 실행된 실험결과는 비교적 간단한 스파크 존스(Sparck Jones)의 역문헌빈도 加重技法과 제안된 單語 加重技法의 검색결과들이 더 복잡한 계산을 요하는 다른 세개의 單語 加重技法들의 검색결과들보다 더 나았다.

ABSTRACT

The present study has two main objectives. The first objective is to devise a new term weighting technique which can be used to weight the significance value of each word stem in a test collection of documents on the subject of "enteral hyperalimentation." The next objective is to evaluate retrieval performance of proposed term weighting technique, together with four other term weighting techniques, by conducting a set of experiments.

The experimental results have shown that the performance of Sparck Jones's inverse document frequency weighting and the proposed term significance weighting techniques produced better recall and precision ratios than the other three complex weighting techniques.

CHAPTER I. INTRODUCTION

1. AIM AND OBJECTIVES

The general aim of the proposed study is to bring about a better understanding of automatic term weighting techniques from the information retrieval point of view. There are four specific objectives in the present study. They are the following:

1) To devise a new "term significance"

* 전남대학교 도서관학과
접수일자 : 1984. 11. 15.

weighting technique which can be used to weight the significance value of each word stem in a document collection.

2) To utilize the three term weighting techniques proposed by previous investigators. They are the following:

i) Salton's term discrimination weighting technique

ii) Sparck Jones's inverse document frequency weighting technique

iii) Harter's compound poisson weighting technique using the method of moments.

3) To modify Harter's compound poisson weighting technique by utilizing the method of maximum likelihood to estimate parameters.

4) To evaluate retrieval performance of the above-mentioned five weighting techniques by conducting a set of retrieval experiments.

2. HYPOTHESES

1) HYPOTHESIS 1···A proposed "term significance" weighting technique is more effective in performance than the more complex weighting techniques such as Salton's term discrimination and Harter's compound posson weighting techniques.

2) HYPOTHESIS 2···Sparck Jones's inverse document frequency weighting technique produces better retrieval than the more complex weighting techniques such as Salton's term discrimination and Harter's compound poisson weighting techniques.

3) HYPOTHESIS 3···A modified compound poisson weighting technique based on the method of maximum likelihood is more effective in retrieval performance than the compound poisson weighting technique based on the method of moments.

CHAPTER II. FIVE TECHNIQUES OF TERM WEIGHTING

This section will discuss a feature extraction process using five techniques of term weighting. A feature extraction process is identified by the following four stages: 1) first stage: term extraction, 2) second stage: term normalization, 3) third stage: term selection, and 4) fourth stage: term weighting. The feature extraction process in the five techniques of term weighting

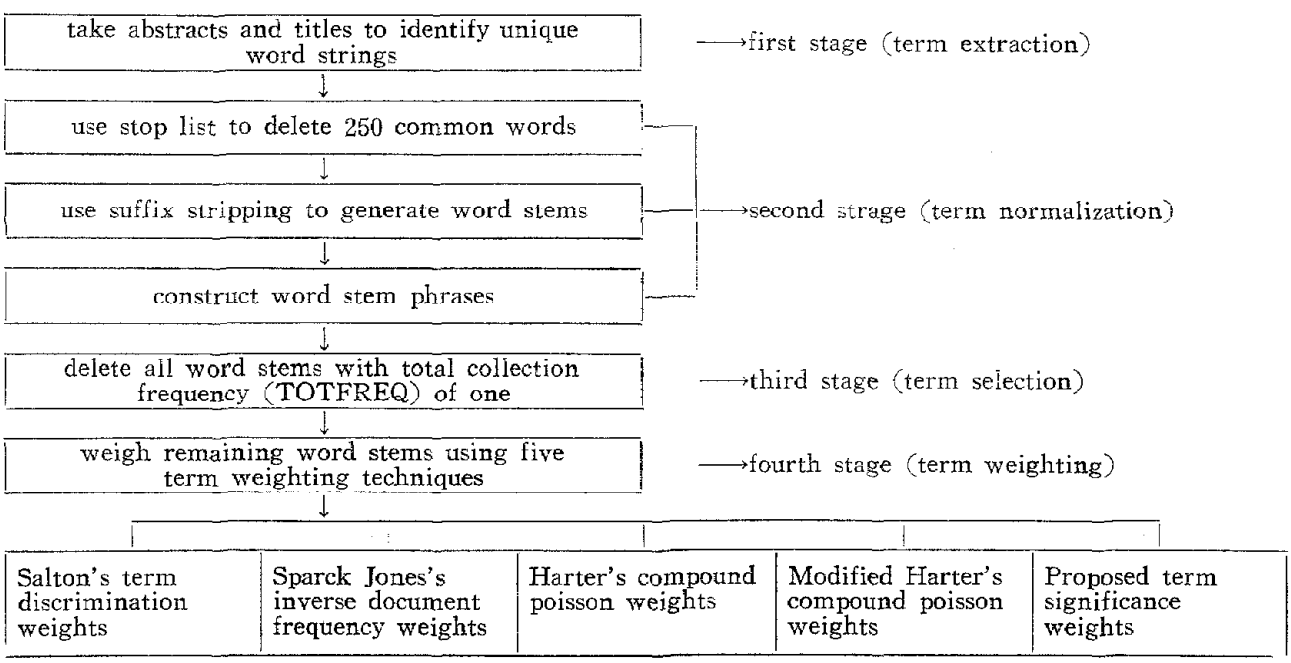

FIGURE 1. FEATURE EXTRACTION PROCESS IN FIVE TECHNIQUES OF TERM WEIGHTING

is shown in Figure 1.

1. SALTON'S TERM DISCRIMINATION WEIGHTING TECHNIQUE

Salton has hypothesized that a good index term used in a document collection renders the documents in the collection as dissimilar as possible, and a bad term renders the documents in the collection as similar as possible, when assigned to a collection of documents. (1) He devised a term discrimination value which can measure the degree to which the use of the term will help to distinguish the documents from each other.

He further has suggested that the best discriminators are those terms which have a medium document frequency over the collection but a skewed frequency distribution in which the term tends to occur frequently in the same document with respect to the group of documents in which they occur.

Salton's term weights utilized both a "global" value and a "local" value. The "global" value indicates the term discrimination value which is an estimate of the overall effectiveness of each word stem in the entire collection. The "local" value indicates the proportion of frequency of a word stem with regard to a given document. The weight value of a word stem in a given document is called "WEIGHT" which consists of the product of the term discrimination value and the "local" value. In the present study, the "WEIGHT" value will be used to weight each word stem extracted from the document texts.

Consider, in particular, a collection of documents. Let D_i and D_j represent two documents each represented by a set of index terms. A similarity measure can be used to represent the degree of similarity between the two documents. There are a number of different measures of similarity between documents(e.g., simple matching correlation, cosine correlation, etc.). In the present study, the cosine correlation will be used to measure the similarity between documents. The cosine correletion between document i (D_i) and document j (D_j) can be defined by:

$$\text{COSINE}\ (D_i, D_j) = \frac{\sum_{k=1}^{t}(\text{TERM}_{ik} \times \text{TERM}_{jk})}{\sqrt{\sum_{k=1}^{t}(\text{TERM}_{ik})^2 \sum_{k=1}^{t}(\text{TERM}_{jk})^2}} \quad \cdots\cdots(3)$$

Where D_i is represented by terms, TERM_{i1}, $\text{TERM}_{i2}, \cdots \text{TERM}_{it}$, where TERM_{ik} is assumed to present the weight, or importance, of term k assigned to document i

D_j is represented by terms, TERM_{j1}, $\text{TERM}_{j2}, \cdots \text{TERM}_{jt}$, where TERM_{jk} represents the weight, or importance, of term k assigned to document j

To demonstrate Salton's term weighting procedures, a sample collection of four documents

TABLE 7. LIST OF WORD STEM FREQUENCIES DERIVED FROM A SAMPLE COLLECTION.

DOCUMENT 1		DOCUMENT 2		DOCUMENT 3		DOCUMENT 4	
WORD STEM	FREQ	WORD STEM	FREQ	WORD STEM	FREQ	WORD STEM	FRFQ
a	4	c	2	a	2	a	2
b	1	d	1	b	1	b	4
e	2			c	1		
				d	2		
				e	7		

is given. Table 7 shows a list of word stem frequencies derived from a sample collection. Each document in the sample collection is represented by a subset of the total set of five unique word stems (e.g., a, b, c, d, and e).

If the similarity measure is computed for all pairs of document D_i and document D_j except when $i=j$, an average value AVERAGE-SIMILARITY is obtainable. This represents the average document-pair similarity for the collection:

$$\text{AVERAGE-SIMILARITY} = (1/\text{CONSTANT}) \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \text{SIMILAR}(D_i, D_j) \quad \cdots\cdots(4)$$

Where AVERAGE-SIMILARITY=the average document-pair similarity
n=the number of documents in collection
CONSTANT=$n(n-1)/2$
SIMILAR(D_i, D_j)=similarity between document i and document j (calculated by the cosine correlation in Eq. (3))

The foregoing expression reflects the density of the document space, that is, the degree to which the documents are bunched up in the "space" of documents. In practice the computation of average document-pair similarity (AVERAGE-SIMILARITY) requries $n(n-1)/2$ comparisons for n documents. Thus in the given collection, 6 comparisons have to be conducted to compute the AVERAGE-SIMILARITY for four documents (e.g., SIMILAR(D_1, D_2), SIMILAR(D_1, D_3), SIMILAR(D_1, D_4), SIMILAR(D_2, D_3), SIMILAR(D_2, D_4), SIMILAR(D_3, D_4)).

However, the space density can be computed more efficiently by constructing an artificial, "average" document $\bar{D}$ as the CENTROID, in which the set of index terms representing the CENTROID $\bar{D}$ are assumed to exhibit average frequency characteristics of all the index terms used in the space. The average frequency of each index term is defined as

$$(\text{AVERAGE FREQ})_k = (1/n) \sum_{i=1}^{n} \text{FREQ}_{ik} \quad \cdots\cdots(5)$$

Where (AVERAGE FREQ)$_k$=the average frequency of term k in a collection
n=number of documents in a collection
FREQ$_{ik}$=within document frequency of term k in document i

The given sample document collection can be then representable by a matrix indicating the CENTROID $\bar{D}$. The matrix is shown in Table 8.

TABLE 8. DOCUMENT-WORD STEM RELATION MATRIX INDICATING THE CENTROID

	T_a	T_b	T_c	T_d	T_e
D_1	4	1	0	0	2
D_2	0	0	2	1	0
D_3	2	1	1	2	7
D_4	2	4	0	0	0
CENTROID($\bar{D}$)	2	1.5	1	1	3

The fifth row in Table 8 gives the average frequency of each word stem. Thus it can be seen that the CENTROID $\bar{D}$ consists of the average frequency of each word stem derived from the document collection.

The average document-pair similarity (AVGSIM) is then obtained from dividing total sum of similarities of each document with the CENTROID $\bar{D}$ by the number of documents in a collection. AVGSIM is defined as:

$$\text{AVGSIM} = (1/n) \sum_{i=1}^{n} \text{SIMILAR}(\bar{D}, D_i) \cdots(6)$$

Where AVGSIM=the average document-pair similarity
n=the number of documents in a

collection

SIMILAR($\bar{D}, D_i$)=the similarity between the centroid and document D_i (calculated by the cosine correlation in Eq. (3))

Without using the CENTROID $\bar{D}$, $n(n-1)/2$ comparisons for a collection of n documents need to be conducted to compute the average-pair similarity (AVGSIM). However, this can be greatly reduced to only n comparisons by using the CENTROID $\bar{D}$. In the given example, only four instead of six comparisons need to be done to calculate the AVGSIM (.eg., SIMILAR ($D_1, \bar{D}$), SIMILAR ($D_2, \bar{D}$), SIMILAR ($D_2, \bar{D}$), SIMILAR ($D_3, \bar{D}$), and SIMILAR ($D_4, \bar{D}$). In this case, the AVESIM is 0.65.

Consider now the original document collection with a given term k removed from all the documents and let $(AVGSIM)_k$ represent the space density in that case. If term k had been a broad, high-frequency term with a fairly even frequency distribution, it is likely that it would have appeared in most document descriptions; therefore, its removal will reduce the average document-pair similarity. This case is clearly unfavorable, because when such a high-frequency term is assigned to the documents, the average similarity will increase and the document space is compressed. On the other hand, if term k had been assigned a high weight in some documents, but not in others, its removal would be likely to increase the average similarity between documents. Salton's discrimination value (DISCVALUE$_k$) of a term is based on the difference between the two average similarity values. It can be computed as

$$DISCVALUE_k = (AVGSIM)_k - AVGSIM \quad \cdots\cdots (7)$$

Where DISCVALUE$_k$=the discrimination value of word stem k

$(AVGSIM)_k$=the average document-pair similarity with word stem k deleted.

AVGSIM=the average document-pair similarity.

In the example given in Table 7, if we want to compute the discrimination value of word stem "a" (DISVALUE$_a$), we need to compute the average document-pair similarity (AVGSIM) and the average document-pair similarity with word stem "a" deleted ($(AVGSIM)_a$). We already have computed the AVGSIM from the collection of four documents, that is, AVGSIM is 0.65. In order to compute the average document-pair similarity with word stem "a" deleted ($(AVGSIM)_a$), the word stem "a" has to be removed from all the documents listed in Table 8, and then the same computational procedures needed for the computation of AVGSIM should take place. In this case, the $(AVGSIM)_a$ is 0.66. Thus the discrimination value for term a (DISCVALUE$_a$) can be computed by: $DISCVALUE_a = (AVGIM)_a - AVGSIM = 0.66 - 0.65 = 0.01$. Therefore, DISCVALUE$_a$ can be viewed as a "global" value for weighting the term "a" with regard to the collection.

Salton's "WEIGHT" value of a given word stem utilizes both the "global" and "local" values. It is the product of the term discrimination value (DISCVALUE) and the proportion of frequency of the word stem in a given document (FREQ):

$$WEIGHT_{ik} = FREQ_{ik} \times DISCVALUE_k \quad \cdots (8)$$

Where $WEIGHT_{ik}$=weight value of the word stem k in a given document i

$FREQ_{ik}$=proportion of frequency of the word stem k in a given document i

DISCVALUE$_k$=the term discrimina-

tion value of the word stem k

Table 8 on page 9, the within-document frequency of the word stem "a" in Document 1 is 4, and total number of word tokens in Document 1 is 7: i.e., including all occurrences of all word stems (see Table 8). Therefore, the weight value of word stem "a" in Document 1 ($WEIGHT_{1a}$) can be computed by:

$$\begin{aligned} WEIGHT_{1a} &= FREQ_{1a} \times DISCVALUE_a \\ &= 4/7 \quad \times 0.01 \\ &= 0.57 \quad \times 0.01 \\ &= 0.006 \end{aligned}$$

For our first term weighting technique, Eq. (8) will be used to weight each word stem in a document. Salton, Yang, Yu, and Wang have demonstrated that the term discrimination weighting technique can be used to weight term efficiently. Salton's term discrimination weighting technique has proved to be theoretically sound. However, his technique involves complicated computations.

2. SPARCK JONES'S INVERSE DOCUMENT FREQUENCY WEIGHTING TECHNIQUE

Inverse document frequency weights were introduced by Sparck Jones. (2) Inverse document frequency weights (IDF) tend to give more weights to terms which occur in only a few documents. Sparck Jones's system of term weights also combined a "global" value with a "local" value. The inverse document frequency value indicated as IDFValue, which reflects the word stem importance in the entire collection, is based on the document frequency of each word stem; it can be considered as the "global" value. The relative frequency of a word stem within a given document indicated as FREQ, is the "local" value. The weight value of a word stem in a given document is called "WEIGHT" which consists of the product of the inverse document frequency value and the "local" value. In the present study, each word stem extracted from the test collection is weighted by this composite "WEIGHT".

Sparck Jones hypothesized that the importance of a term is inversely proportional to the total number of documents to which each term is assigned. She devised a term weighting formula which assigns the higher weights to the rarer terms, and gives the lower weights to more common terms. Her inverse document frequency weighting value of a given word stem k is defined by the following formula:

$$\text{IDFValue } k = \text{Log}_2(n) - \text{Log}_2(\text{DOCFREQ}_k) + 1 \cdots (9)$$

Where IDFValue k=the inverse document frequency weight value of the word stem k

n=the number of documents in a collection

DOCFREQ_k=the number of documents to which the word stem k is assigned

For example, in one collection of 1,333 documents, consider the word stem "albumen" occurring in 23 documents, and the word stem "abdomin" occurring in 48 documents. The inverse document frequency values of these two word stems are:

$$\begin{aligned} &\text{IDFValue(albumen)} \\ &= \text{Log}_2(1333) - \text{Log}_2(23) + 1 \\ &= 7.20 - 3.14 + 1 = 5.06 \\ &\text{IDFValue(abdomin)} \\ &= \text{Log}_2(1333) - \text{Log}_2(48) + 1 \\ &= 7.20 - 3.87 + 1 = 4.33 \end{aligned}$$

Sparck Jones's "WEIGHT" value of a word stem can be obtained from the product of its inverse document frequency (IDFValue) and the proportion of frequency of the word stem

in a given document(FREQ). The "WEIGHT" value of a word stem k in the document i is:

$$WEIGHT_{ik}=FREQ_{ik}\times IDFValue_k \quad \cdots\cdots(10)$$

Where $WEIGHT_{ik}$=weight value of the word stem k in a given document i

$FREQ_{ik}$=proportion of frequency of the word stem k in a given document i

$IDFValue_k$=inverse document weight value of the word stem k

In the given example, if the within-document frequency of the word stem "albumen" in Document 1 is 3, and total number of word tokens in Document 1 is 21: i.e.' including all occurrences of all word stems, then the "WEIGHT" value of "albumen" in Document 1 is:

$$\begin{aligned} WEIGHT_1 albumen &= 3/21\times 5.06 \\ &= 0.14\times 5.06 \\ &= 0.72 \end{aligned}$$

The inverse document frequency weights are relatively simple and easy to apply. Furthermore, the experiment by Salton has shown that the simple inverse document frequency weights are more effective in retrieval than Salton's term discrimination weights whose computations are more complicated.

3. HARTER'S COMPOUND POISSON WEIGHTING TECHNIQUE

The compound poisson model proposed by Bookstein, Swanson, and Harter is based on the basic notion of word frequency counts in a document (3)(4). Salton's term discrimination weights and Sparck Jones's IDF weights are based on the assumption that good index terms may help to distinguish the documents from each other. However, Harter's approach to the concept of good index terms is different from the both of these techniques. Harter suggested that a word can be selected as a good index term on the basis of its frequency of occurrence in a document if its occurrence tends to cluster in a relatively few documents.

After the stop words are deleted, he divided the content-bearing words in the document texts into two classes···specialty words and non-specialty words (3). Specialty words (e.g., Vivonex, cancer, amino acid) possess significance *in representing the document content.* Therefore, these words could be used as good index terms. Although non-specialty words (e.g., see, based, consist) may be content-bearing, they have little value for indexing purpose. On the other hand, specialty words in one collection may be not useful as index terms in another collection. The word "cancer" may be a good index term. The same word "cancer" may appear in a collection of documents on the subject of computer science. In this case, "cancer" may be a non-specialty word in the computer science collection.

He suggested that non-specialty words tend to be distributed at random in a collection of documents. In contrast, specilaty words tend not to be so distributed.(4) He defined a randomly distributed word as one whose distribution among documents may be described by a poisson density function. He further proposed that a specialty word distinguishes more than one class of documents with respect to the extent to which the topic named by the word is treated in the documents in the collection.(4) Thus a specialty word can divide documents in the collection into two classes: class I documents and class II documents. In the class I documents, the subject represented by the word is treated to a relatively greater extent than the topic named by the word as treated in the class II documents.

In Harter's experiment, words with the same

TABEL 9. FREQUENCY DISTRIBUTIONS FOR TWO WORD STEMS

NUMBER OF DOCUMENTS CONTAINING k TOKENS												
k TOKEN	0	1	2	3	4	5	6	7	8	9	10	TOTFREQ	DOCFREQ
albumen	1310	18	3	1	1	0	0	0	0	0	0	31	23
abdomin	1285	37	8	3	0	0	0	0	0	0	0	62	48

TABLE 10. MOMENT ESTIMATES OF h, $m1$, AND $m2$ FOR THE TWO WORD STEMS LISTED IN TABLE 9

WORD STEM	$m1$	$m2$	h
albumen	1.2557	0.0091	0.0114
abdomin	0.5484	0.0000	0.0848

stem(as girl and girls) are considered separately and individually. But, in the present study, Harter's model is being applied to word stems rather than word strings for the following reasons: 1) the size of the total file of index terms can be reduced; and 2) higher recall would result from retrieval using word stems.

Compound poisson weights also incorporate two kinds of values···a "global" value and a "local" value. The "global" value indicates an estimate of the overall effectiveness of each word stem as a potential index term. The "local" value presents the probability that a document belongs to the class I given that it has a certain given occurrence of a given term. Harter called his "global" value, the "Z" value. He combined the "local" value, and called it the "B" value. In the present study, the "B" value was used to weight each word stem in the document texts.

Harter used a compound poisson model to identify the specialty word. The compound poisson model can be expressed by the following formula:

$$f(k)=h\frac{e^{-m_1}m_1{}^k}{k!}+(1-h)\frac{e^{-m_2}m_2{}^k}{k!}\cdots\cdots(11)$$

$f(k)$ =proportion of documents in the collection containing k occurrences of a word stem.

h =proportion of documents in the collection which belongs to class I documents.

e =base of natural logarithms (2.71828)

m_1 =mean number of occurrences of the word in class I documents.

m_2 =mean number of occurrences of the word in class II documents.

k =number of occurrences of the word(e.g., 0, 1, 2, 3, ···etc.)

This compound poisson model is characterized by three parameters. m_1 and m_2 are the average numbers of occurrences of the word in class I documents and class II documents respectively, and a third parameter h indicates the proportion of documents in the collection which belong to class I documents. These parameters can not be easily obtained from the observed data. Therefore, Harter utilized the method of moments in order to compute the estimates of m_1, m_2, and h. The following gives an example of the computation of two word stem values to demonstrate Harter's term weighting procedures. Table 9 shows frequency distributions for the two word stems.

In Table 9, 1,310 documents have no occurrence of the word stem "albumen." Eighteen documents have one occurrence of the word stem. The document frequency (DOCFREQ) of the word stem is 23. And the total collection frequency (TOTFREQ) of the word stem is 31. The moment estimates of h, m_1, and m_2 for the two word stems listed in Table 9 are shown in Table 10.

Harter's "Z" value is based on the suggestion

made by John Swets and was modified by B. C. Brooks. The "Z" value is a statistical measure consistent with the compound poisson model to separate specialty words from non-specialty words. It is defined by:

$$Z=\frac{m_1-m_2}{(m_1+m_2)^{1/2}} \quad \cdots\cdots(12)$$

where Z=the modified measure of indexing effectiveness of a term.

The "Z" value reflects the "term discrimination power" of a term in the entire collection. For example, if "Z" is large for a given term, the term weighting for that term is large; thus, the given term could be a good index term. With the estimates of m_1, and m_2 obtained from Table 10 for the two sample word stems, their "Z" values can be computed (see Table 11).

TABLE 11. Z VALUE ASSOCIATED WITH EACH OF THE TWO WORD STEMS LISTED IN TABLE 9 AND 10

WORD STEM	Z
albumen	1.1084
abdomin	0.7405

Harter's "B" value of a word stem in a given document d can be obtained from the sum of the "Z" value and the "local" value of the given word stem by:

$$B=Z+p(d\in I/k) \quad \cdots\cdots(13)$$

Where B=the weight value of a term in a given document d

Z=an estimate of the overall effectiveness of a word stem as a potential index term (global value)

$P(d\in I/k)$=the probability that a document belongs to the class I* in which the word is important as a content-bearing term given that it has k occurences (local value)

$p(d\in I/k)$ is defined by:

$$p(d\in I/k)=\frac{he^{-m_1}m_1^{k}}{he^{-m_1}m_1^{k}+(1-h)e^{-m_2}m_2^{k}} \cdots(14)$$

For example, for the word stem "albumen," the "B" values for $k=1, 2, 3$, and 4 are displayed in Table 12.

TABLE 12. THE "B" VALUE ASSOCIATED WITH THE WORD STEM "ALBUMEN" LISTED IN TABLE 9, 10, AND 11

k	WORD STEM	B
1	albumen	1.4223
2	albumen	2.0929
3	albumen	2.1083
4	albumen	2.1084

Harter's model can be looked upon as one in which the importance of a term in representing the contents of a document is balanced against its importance as a discriminator among *documents in the collection*. Harter's "B" value of a word stem necessitated the complicated computation to estimate m_1, m_2, and h. However, Harter has shown the experimental evidence that the "B" value produced better results than weights produced by more simple within-document frequenicies.

Aside from using the three weighting tedhniques devised by Salton, Sparck Jones, and Harter, this present study also made use of two new term weighting techniques. This first is based on a modification of Harter's compound poisson term weighting technique. The second utilizes a combination of the average frequency of a term k in each document of the group of documents containing the term k, and the

* $p(d\in I/k)$ is the probability that a document belongs to the class I given that it has k occurrences of the term. Class I contains a set of documents in which the word is important for *describing the content*. It is assumed that every document d is a member of either class I or class II.

document frequency of the term k. In this section, description of the modified Harter's technique is provided.

4. MODIFIED HARTER'S COMPOUND POISSON WEIGHTING TECHNIQUE

The modified Harter's term weights also consist of the sum of the "global" value and the "local" value of a given term. The "global" value or the "Z" value indicates an estimate of the overall effectiveness of each word stem as an potential index term. The "local" value indicates the probability that a document belongs to the class I documents given that it has a certain given occurrence of a term. Harter used the compound poisson model to identify the specialty words. The compound poisson model defined by Harter is (4):

$$f(k)=h\frac{e^{-m_1}m_1^{k}}{k!}+(1-h)\frac{e^{-m_2}m_2^{k}}{k!}\cdots\cdots(11)$$

There are several other available methods to compute the estimates of m_1, m_2, and h, although Harter utilized the method of moments. Blischke suggested that in dealing with a mixture of two distributions, the method of maximum likelihood (ML) provides iterative solutions rather than exact solutions. And, in general, the solutions are very slow to converge (5). Therefore, for convenience, the use of less efficient moment estimators may be a preferable method for estimating the parameters of a mixture of discrete distributions. For these reasons, Harter used the method of moments for calculating the estimates of m_1, m_2, and h. However, the method of maximum likelihood, which is more efficient than the method of moments, would also be appropriate.

There are several reasons why one might want to use the maximum likelihood estimator (MLE) to estimate parameters. Although maximum likelihood estimators are not always unbiased and efficient, they are usually the best that one can find because of the following properties.

1. the bias of the MLE tends zero.
2. the standard error of the MLE approaches the smallest possible standard error.
3. the sampling distribution of the MLE approaches normality.

It is because of these properties that many statisticians favor the use of maximum likelihood estimators in many sampling situations. Therefore, we reason that by using the method of ML to estimate m_1, m_2, and h, Harter's results may be improved. Thus in the proposed method of the modified compound poisson weights, the method of maximum likelihood will be utilized to compute the estimates of m_1, m_2, and h in order to compare the retrieval results to these of compound poisson weights using the method of moments in terms of recall and precision ratios.

i) COMPUTIONS OF THE MAXIMUM LIKELIHOOD ESTIMATORS

The probability density function of the compound poisson distribution is (6):

$$f(x;m_1,m_2,h)=h\frac{e^{-m_1}m_1^{x}}{x!}+(1-h)\frac{e^{-m_2}m_2^{x}}{x!}\cdots\cdots\cdots(15)$$

Where h=the proportionality factor $(0\leqq h\leqq 1)$

e=base of natural logarithms(2. 71828)

m_1=component parameter

m_2=conent parameter

$x=0, 1, 2, 3\cdots\cdots$

Then, the likelihood function for the random sample of $x_1, \cdots x_n$, from the compound poisson distribution becomes

$$L(x_1,\cdots,x_n;m_1,m_2,h)=\prod_{i=1}^{n}f(x_i;m_1,m_2,h)$$
$$=\prod_{i=1}^{n}\left(h\frac{e^{-m_1}m_1^{xi}}{xi!}+(1-h)\frac{e^{-m_2}m_2^{xi}}{xi!}\right)\cdots(16)$$

To find the estimates of m_1, m_2, and h, it is easier to work with the logarithm of the likelihood than with the likelihood itself. Since the logarithm is monotonically increasing, the estimates of m_1, m_2, and h that maximize the log-likelihood also maximize the likelihood.

Hence, the log-likelihood function is

$$\begin{aligned}&\mathrm{Log}L(x_1, \cdots, x_n; m_1, m_2, h)\\&=\sum_{i=1}^{n} \mathrm{Log}\, f(x_i; m_1, m_2, h)\\&=\sum_{i=1}^{n} \mathrm{Log}\left(h\frac{e^{-m1}m_1{}^{xi}}{xi!}+(1-h)\frac{e^{-m2}m_2{}^{xi}}{xi!}\right)\end{aligned} \quad \cdots\cdots(17)$$

By computing the partial derivatives of the log-likelihood functions with respect to m_1, m_2, and h, in order to find the location of its maximum, and setting these partial derivatives equal to zero, we obtain:

$$\frac{\partial \mathrm{Log}\, L}{m_1}=\sum_{i=1}^{n}\frac{h(-e^{-m1}m_1{}^{xi}+e^{-m1}x_i m_1{}^{xi-1})}{he^{-m1}m_1{}^{xi}+(1-h)e^{-m2}m_2{}^{xi}}=0 \quad \cdots\cdots(18)$$

$$\frac{\partial \mathrm{Log}\, L}{m_2}=\sum_{i=1}^{n}\frac{(1-h)(-e^{-m2}m_2{}^{xi}+e^{-m2}x_i m_2{}^{xi-1})}{he^{-m1}m_1{}^{xi}+(1-h)e^{-m2}m_2{}^{xi}}=0 \quad \cdots\cdots(19)$$

$$\frac{\partial \mathrm{Log}\, L}{h}=\sum_{i=1}^{n}\frac{e^{-m1}m_1{}^{xi}-e^{-m2}m_1{}^{xi}}{he^{-m1}m_1{}^{xi}+(1-h)e^{-m2}m_2{}^{xi}}=0 \cdots(20)$$

The maximum likelihood estimates of m_1, m_2, and h, can be obtained by solving simultaneously the systems of three nonlinear equations···Eq. (18), Eq. (19), and Eq (20). To solve the equations, a computer program, "MAX", which was developed for the present study, was used to solve the equations based on the secant method. "MAX" is a FORTRAN program which calculates the estimates of m_1, m_2, and h of a compound poisson distribution. "MAX" requires, for each word stem, initial values and frequency distribution as input data. The moment estimates of m_1, m_2, and h were used here as initial values. "MAX" prints out the following:

1) the word stem
2) initial values of m_1, m_2, and h based on the first three sample moments (these values are used as the first approximation in the secant process)
3) itmax···the maximum allowable number of iterations
4) a final values of m_1, m_2, and h based on the secant iteration.
5) fnorm···on output, fnorm is equal to $f(1)^2 \cdots f(n)^2$ at the point x.
6) ier···error parameter. (output)

ILLUSTRATIVE EXAMPLE FOR OUTPUT

word stem =albumen
initial $x(I)$=1.2557 0.0091 0.0114
itmax =100
final $x(I)$=2.1676 0.0125 0.0057
fnorm =0.00
ier =0.00

The maximum likelihood estimates of m_1, m_2, and h for the two word stems listed in Tables 9 (see p. 136) and Table 10 (see p. 136) are shown in Table 13.

TABLE 13. MAXIMUM LIKELIHOOD ESTIMATES OF m_1, m_2 AND h FOR THE TWO WORD STEMS LISTED IN TABLES 9 AND 10.

word stem	$m1$	$m2$	h
albumen	2.1676	0.0125	0.0057
abdomin	1.8508	0.0254	0.0139

The "Z" value is a statistical measure consistent with the compound poisson model to separate specialty words from non-specialty words. We use the same "Z'" value calculation

as defined by Harter:

$$Z' = \frac{m_1 - m_2}{(m_1 + m_2)^{1/2}} \quad \cdots\cdots(21)$$

Where Z' = the modified measure of indexing effectiveness of a term.

The "Z'" value associated with each of the two word stems listed in Table 13 is shown in Table 14.

TABLE 14. "Z'" VALUE ASSOCIATED WITH EACH OF THE TWO WORD STEMS LISTED IN TABLE 13

word stem	Z'
albumen	1.4596
adbomin	1.3327

Our modification of Harter's "B" value of a word stem in a given document d can be obtained by the sum of its "Z'" value and its "local" value:

$$B' = Z' + p(d \in i/k) \quad \cdots\cdots(22)$$

Where B' = the weight value of a term in a given document d

Z' = an estimate of the overall effectiveness of a word stem as a potential index term using the method of maximum likelihood (global value)

$p(d \in i/k)$ = the probability that a document belongs to the class I in which the word is important as a content-bearing term given that it has k occurrences (local value)

TABLE 15. THE "B" VALUE ASSOCIATED WITH THE WORD STEM "ALBUMEN" USING HARTER'S TECHNIQUE AND THE MODIFIED HARTER'S TECHNIQUE LISTED IN TABLES 13 AND 14

k	WORD STEM	B'	B
1	albumen	1.4223	1.5629
2	albumen	2.0929	2.4119
3	albumen	2.1083	2.4593
4	albumen	2.1084	2.4596

Table 15 gives the comparison of the final composite term weights for $k=1, 2, 3$, and 4 using Harter's original technique and our modified version for the word stem "albumen."

5. PROPOSED TERM SIGNIFICANCE WEIGHTING TECHNIQUE

Finally, this study also utilizes a term weighting technique which combines features of the document frequency of a given term and its average frequency per document over the entire collection. This proposed "term significance" weighting technique is prompted by suggestions made by Luhn and Salton (7)(8):

1) that the broad terms with high document frequency lead to substantial losses in precision, and the narrow terms with low document frequency lead to recall losses. Therefore, the medium document frequency terms in a collection are probably the most effective in terms of document retrieval;

2) that there is some relation between term distribution pattern in a colllection and the average frequencies of terms. In other words, good discriminators, which are clustered together in a group of documents, tend to have high average frequencies. The poor discriminators, which occur evenly in a document collection, have low average frequencies.

The proposed "term significance" weighting··· technique also incorporated two kinds of values··· two "global" values and one "local" value. The "global" value consists of a "resolving power" value and a "term skewness" value. The "resolving power" value is a term borrowed from Luhn, and it indicates an estimate of the overall effectiveness of each word stem as a potential index term based on the document frequency of each word stem. The "term skewness" value represents the word stem importance in the entire collection based on

the average frequency of each word stem. The weight value of a word stem in a given document is called "WEIGHT" which consists of the product of its "global" value and its "local" value. In the present study, the combined "WEIGHT" value will be used to weight each word stem.

This proposed "term significance" weighting technique is very similar to Salton's term discrimination weights in approach. First, both weighting schemes are based on the document frequency of each term and its distribution pattern in a collection. Second, the aim of these two techniques is to give medium document frequency terms more favoured status. However, However, the proposed "term significance" weighting technique differs from Salton's term discrimination weighting techniquen that the proposed "term significaince" weighting technique utilized the document frequency and the average frequency of a given word stem over the entire collection while Salton's term discrmination value involves the complicated computational procedures.

(a) LUHN'S RESOLVING POWER MODEL

Luhn has hypothesized that the medium frequency terms in a document are probably

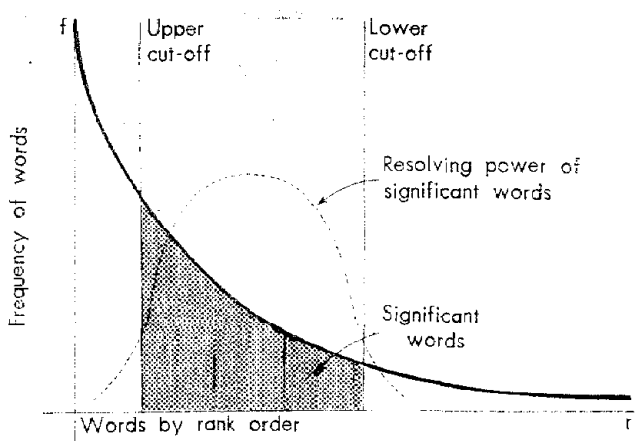

FIGURE 2. A PLOT OF THE HYPERBOLIC CURVE RELATION *F*, THE FREQUENCY OF OCCURRENCE AND *R*, THE RANK ORDER (ADOPTED FROM SCHULTZ (9) PAGE 120.)

the most effective in terms of document retrieval. Figure 2 shows how the "resolving power" of words is related to their frequencies within each document. The relation between the rank of a word in the order of its frequency and its actual frequency is described by a line superimposed on a graph of the resolving power of those words. It can be seen that the highest discriminating power is associated with words in the middle-frequency range.

Luhn suggested that the "resolving power" of significant words reached a peak at a rank order position half way between the two cut-offs and from the peak fell off in either direction, reducing to almost zero at the cut-off points. The concept of the "resolving power" seems reasonable. However, Luhn did not offer a method to compute the weights for each term in a document. He gave no rule to determine the required frequency thresholds. Nevertheless, Luhn's idea to relate term frequency in a document as an indication of subject content in the document has proved to be the basis of most automatic indexing experiments.

Salton also noted that the terms with medium document frequencies are more powerful in discriminating relevant document from the nonrelevant documents than those with high or low document frequencies. He differs from Luhn's original idea in that document frequency rather than term frequency within a single document is utilized in his term discrimination weighting technique. However, there remains the problem of finding a method to determine the exact measure of what constitute "medium" frequencies in the document count or the word count in given text. Goffman was able to identify a "transition" point in a word frequency distribution, in which high frequency words begin to chang from Zipf's first law of high

frequency words to characteristics of low frequency words. (10) Although Goffman's "transition" formula was used in frequency distribution of words in a single text, it is reasonable to speculate that it could be used to identify "medium" document frequency in a document collection. The following section proceeds to show that document frequencies of word stems in a collection also show a Zipfian distribution.

(b) PROPOSED RESOLING POWER WEIGHTING TECHNIQUE

i) GOFFMAN'S TRANSITION POINT

Zipf found that there is a constant relation between the rank of a word in order of frequency and its frequency in a text(11). Zipf's first law for the high frequency words is:

$$R \times F = C \quad \cdots\cdots (23)$$

Where R=rank of a word in the distribution of all words in the text.

F=frequency of occurrence of that word in the text.

C=a constant for a given text.

Zipf's second law for the low frequency words was generalized by A.D. Booth. Booth's modification of Zipf's law is:

$$I_1/I_n = n(n+1)/2 \quad \cdots\cdots (24)$$

Where I_1=number of words occurring only once.

I_n=frequency of occurrence of words appearing n times in the text.

Goffman devised a formula to identify the "transition" position of a word distribution where characteristics of high frequency words is replaced by low frequency words. He further suggested that the "transition" point identifies an area of the distribution where good index terms might appear. His formula is

$$T = \frac{-1+(1+8I_1)^{1/2}}{2} \quad \cdots\cdots (25)$$

Where T=the word frequency at which the transition from high freqency of word occurrence is replaced by low frequency of word occurrence.

I_1=number of words occurring only once.

In the present study, it has been found that there is a constant relationship between document frequency of occurrence of a word stem in a given collection, and the rank of that word stem, when all word stems are ranked in decreasing document frequency of occurrence.

TABLE 16. RANK-HIGH DOCUMENT FREQUENCY WORD STEM DISTRIBUTION

word stem	document frequency (DOCFREQ)	rank (R)	$R \times$DOCFREQ =CONSTANT
diet	1012	1	1012
patient	690	2	1380
element diet	580	3	1740
Vivonex	528	4	2112
us	445	5	2225
feed	401	6	2406
nutrit	353	7	2471
effect	294	8	2352
stud	291	9	2619
dai	257	10	2570
includ	184	20	3680
		total	= 24567
		average	=2233.4

TABLE 17. DISTRIBUTION OF WORD STEMS OF LOW DOCUMENT FREQUENCY

document frequency (DOCFREQ)	number of words (I_n)	experimental I_1/I_n	theoretical $n(n+1)/2$
1	1292	1292/1292= 1.0	1.0
2	445	1292/ 445= 2.9	3.0
3	219	1292/ 219= 5.9	6.0
4	169	1292/ 169= 2.6	10.0
5	119	1292/ 119=10.9	15.0
6	97	1292/ 97=13.3	21.0
7	79	1292/ 79=16.4	28.0
8	66	1292/ 66=19.6	36.0
9	44	1292/ 44=29.4	45.0
10	43	1292/ 43=30.4	55.0
20	19	1292/ 19=68.0	210.0

In other words, the pattern of word stem dispersion in a collection of documents on the same subject can be described by the Zipfian distribution. The distribution of word stems of high document frequency using Zipf's first law is shown in Table 16. The distribution of word stems of low document frequency using Booth's revised second law is illustrated in Table 17.

A simple way of checking whether a set of data confirm with the Zipf's first law is to plot the log of document frequency (DOCFREQ) against the log of rand (R), and then to check whether the log-log graph expresses the expected formulation as a linearity. According to the original application, each different word stem is assigned a unique rank. In the given test collection, the highest document frequency (1012) will be assigned the rank one, the secondhighest document frequency (690) will be given the rank two, etc. when two word stems have same document frequency 201, two unique ranks of the document frequency 201 will be arbitrarily assigned 17 and 18. Therefore, the average rank (AVGR) of the document frequency 201 will be $(17+18)/2=17.5$. The time series processor (TSP), which is a program for econometric estimation, was used to plot the logarithm of average rank (AVGR) on the x axis against the logarithm of document frequency (DOCFREQ) on the y axis. The graph shown in Figure 3 can be explained in the power low relationship formula:

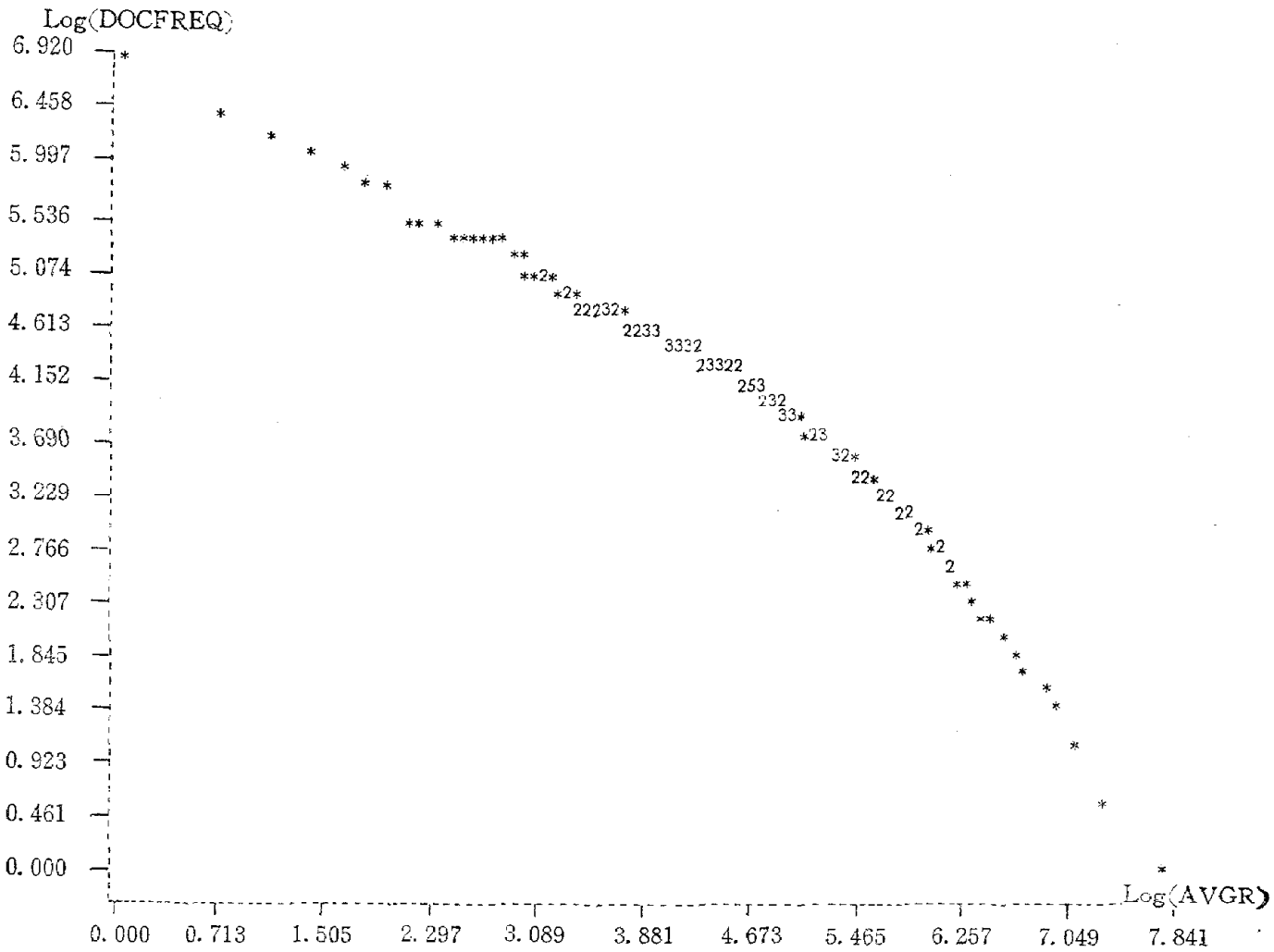

FIGURE 3. A PLOT OF THE LOG OF AVERAGE RANK (AVGR) FOR THE LOG OF DOCUMENT FREQUENCY (DOCFREQ)

$$DOCFREQ \times R^{-b} = C \quad \cdots\cdots\cdots\cdots\cdots\cdots (26)$$

Where DOCFEQ=the number of documents of a given word stem in the collection

R=rank of the word stem in the distribution of word stems in the collection.

b=the slope of the straight line (in this case, b is -0.78)

C=a constant for a given collection.

Consequently, it can be seen that the distribution pattern of word stems in a collection of documents on the same subject can be discribed by Zipfian distribution. It is reasonable to speculate that the "transition" region may contain good word stems for indexing purposes for the collection. Thus, in the proposed "resolving power" weighting technique, the "transition" formula will be used to find the most significant words in a collection. In other words, the words whose document frequencies are at the "transition" point will be assigned the larger weighting values.

ii) LOG-NORMAL DISTRIBUTION

In the following section, a description is provided for a weighting procedure. It will assign larger values to the word stems whose document frequencies are at Goffman's "transition" point. It will give smaller values to the terms whose document frequencies are higher or lower than the "transition" point, reducing to almost zero at the highest and lowest document frequency terms in the collection.

In the test collection, low DOCFREQ values appear sequentially (e.g., DOCFREQ is 1, 2, 3…). But, high DOCFREQ values do not appear sequentially (e.g., DOCFREQ is…580, 690, 1012). The entire collection has only 142 different kinds of DOCFREQ, while the range of DOCFREQ values is from 1 to 1012 (the lowest DOCFREQ is 1 and the highest DOCFREQ is 1012). Consequently, if we assign the moderately medium document frequency whose value is derived from Goffman's "transition" point formula the largest weighting value, and give the document frequencies whose values are higher or lower than "transition" point relatively smaller weighting values, then we may draw the relation between each document frequency and its weighting value approximately by a righted skewed distribution. The nonnormal distribution such as the righted distribution and the lefted skewed distribution can be explained using mathematical transformation on the sample data. It is known that logarithmic transformation makes a skewed distribution more normal. Thus log-normal distribution equation will be used to compute the "resolving power" value of each word stem. The Goffman formula which identifies the "transition" point between high document frequency and low document frequency is used to compute the mean (peak) which is $\bar{x}$ in the following Eq. (27) The "resolving power" of a word stem which has x document frequency, as computed by log-normal distribution equation, is as follows:

$$\begin{aligned} REPOValue_x &= [1/(\sqrt{2\pi}\cdot s)] \exp[-0.5((\ln(x) \\ &\quad -\ln(\bar{x}))/s)^2] \quad \cdots\cdots\cdots\cdots\cdots\cdots (27) \end{aligned}$$

Where $REPOValue_x$=the resolving power value for the word stems exhibiting DOCFREQ x

s=standard deviation

π=3.14

x=document frequency.

$\bar{x}$=mean

The following gives an example of the computation of the word stem "albumen" value to demonstrate the proposed "resolving power"

weighting procedure. The document frequency (DOCFREQ) of the word stem "albumen" is 23. In order to make use of the log-normal distribution equation described in the Eq. (27) for the word stem "albumen" weighting, we need to compute the standard deviation(s) and the mean ($\bar{x}$) of the Eq. (27) In this case, the standard deviation(s) is 1.15361, and the mean($\bar{x}$) is 50.34. Thus the "resolving power" value of the word stem "albumen" exhibiting document frequency 23 can be computed by:

$$\text{REPOValue}_{23} = [1/(\sqrt{2\pi}\cdot 1.15361)] \exp[-0.5((\ln(23) - \ln(50.34)/1.15361)^2] = 0.274$$

Figure 4 displays the plot giving the "resolving power" value for the logarithm of each DOCFREQ indicating a ±2 standard deviations. These standard deviations can be used to determine the upper cut-off and the lower cut-off points, for the term transformation or deletion respectively. In the given example, the standard deviation(s)=1.15361, and the mean($\bar{x}$)=50.34. Thus, the logarithm of the mean value will be ln(50.34)=3.9188. With a 95% confidence limit, i.e., $\ln(\bar{x}) \pm 2s$. $\ln(50.34) - 2\times 1.15361 = 1.61158$. Since ln(5)=1.61158, DOCFREQ 5 can be used for the lower cutoff value. $\ln(50.34) + 2\times 1.15361 = 6.22602$. Since ln(506)=6.22602, DOCFREQ 506 can be used for the upper cutoff value. For all these reasons, it is reasonable to use the "resolving power" value

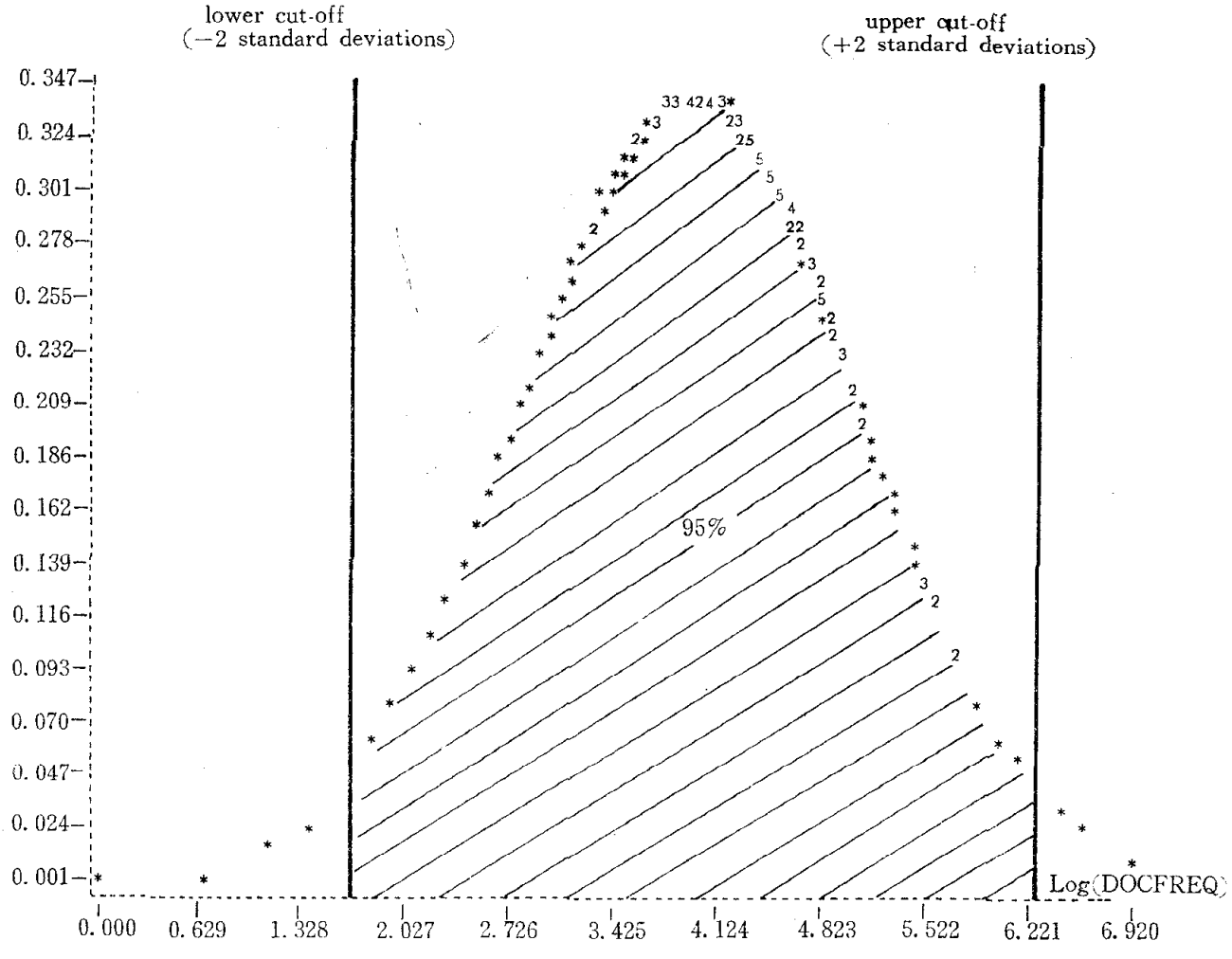

FIGURE 4. A PLOT OF THE RESOLVING POWER VALUE FOR THE LOG OF EACH DOCFREQ INCICATING ±2 STANDARD DEVIATIONS.

as the first "global" value.

(C) PROPOSED SKEWNESS WEIGHTING TECHNIQUE

The second "global" value is computed by the average frequency of the word stems per document in the collection. Salton has observed that there is some relation between term distribution pattern in a collection and the average frequency of the term per document in the entire collection. In other words, the good discriminators, which are clustered together in a group of documents, tend to have high average frequencies. The poor discriminators, which occur evenly in the entire collection, have low average frequencies. In our experimental collection, the first ten word stems with the highest average frequencies are similar to the best ten word stems based on Harter's "Z" value. Table 20 shows the ten best word stems using Harter's "Z" value and a measure based on average frequency.

In the proposed "term skewness" weighting technique, the average frequency of each word stem will be utilized to measure its term skewness value. In computing the proposed "term significance" value consisting of the product of the "resolving power" value and the "term skewness" value, if we use directly the average frequency value of each word stem as the "term skewness" value, then the "term skewness" value, which is usually much larger than the "resolving power" value, dominates the product. Therefore, some normalization is necessary to utilize the average frequency value of each word stem for the "term skewness" value. Therefore, the following term skewness value equation will be utilized to measure the "term skewness" value.

$$SKEValue_k = \ln(AVEFREQ_k) + 1 \quad \cdots\cdots(28)$$

Where $SKEValue_k$ = the skewness value of word stem k

$AVEFREQ_k$ = average frequency of word stem k

In our previous example, the word stem "albumen" has 23 DOCFREQ and 31 TOTEREQ, thus its average frequency(AVEFREQ) will be 31/23=1.35. The "term skewness" value of the word stem "albumen" can be computed by:

$$SKEValue_{albumen} = \ln(1.35) + 1 = 1.300$$

A composite of the two "global" values…the "resolving power" value (REPOValue) and the "term skewness" value (SKEValue) forms our "term significance" value:

TABLE 20. TEN WORD STEMS USING TWO WORD STEM SIGNIFICANCE MEASURES

	compound poisson(moment)		average frequency(AVEFREQ)			
R	best word	weight	best word	TOTFREQ	DOCFREQ	AVEFREQ
1	decarboxylas	2.45	decarboxylas	7	1	7.0
2	glucos	2.44	color	6	1	6.0
3	color	2.34	sweat	6	1	6.0
4	sweat	2.34	flatu	5	1	5.0
5	leucin	2.24	s-185	5	1	5.0
6	uric	2.17	tunnel	4	1	4.0
7	duoden	2.07	chenodexychol	4	1	4.0
8	orithin	2.07	fructokinas	4	1	4.0
9	flatu	2.07	hear	4	1	4.0
10	s-185	2.00	mutagen	4	1	4.0

$$SIGValue_k = (REPOValue_x + 1)^* \times SKEValue_k \cdots (29)$$

Where $SIGValue_k$ = the significance value of word stem k

$REPOValue_x$ = the resolving power value for the word stem k exhibiting DOCFREQ x

$SKEValue_k$ = the skewness value of word stem k.

In the given example, the "term signicance" value for the word stem "albumen" can be computed by:

$$SIGValue_{albumen} = (REPOValue_{23} + 1) \times SKEValue_{albumen}$$
$$= (0.274 + 1) \times 1.300$$
$$= 1.6562$$

The proposed "WEIGHT" value of a word stem can be obtained from the product of the "term significance" value (SIGValue) consisting of the product of the "resolving power" value (REPOValue) and the "term skewness" value (SKEValue), and the proportion of frequency of word stem in a given document (FREQ):

$$WEIGHT_{ik} = ((REPOValue_x + 1) \times SKEValue_k) \times FREQ_{ik} \cdots (30)$$

Where $WEIGHT_{ik}$ = weight value of word stem k in a given document i.

$REPOValue_x$ = the resolving power value for the word stem k exhibiting DOCFREQ x.

$SKEValue_k$ = the skewness value of the word stem k

$FREQ_{ik}$ = proportion of frequency of the word stem k in a given document i.

In the given example, the within-document frequency of the word stem "albumen" in Document 1 is 5, and total number of word tokens in Document 1 is 45: i.e., including all occurrences of all word stems, therefore, the "WEIGHT" value of "albumen" in Document 1 is

$$WEIGHT_{albumen} = ((0.274 + 1) \times 1.30) \times 5/45$$
$$= 1.6562 \times 0.1111$$
$$= 0.1840$$

CHAPTER III
SUMMARY AND CONCLUSIONS, AND IMPLICATIONS

1. SUMMARY AND CONCLUSIONS

The present study has two main objectives. The first objective is to devise a new term weighting technique which can be used to weight the significance value of each word stem in a collection. The second objective is to evaluate the retrieval performance of the proposed "term significance" weighting technique, together with four other term weighting techniques. In order to evaluate the five term weighting techniques in terms of recall and precision ratios, a test collection of documents was used. The test collection on the subject of "enteral hyperalimentation" consists of the title and abstracts of 1,333 documents, together with 22 questions. These 22 questions from users were posed to an operational information system, the Information Service Center at Norwich-Eaton between January 1984 and May 1984. Sets of relevant documents were obtained from retrieval results based on the test collection manually indexed by information specialists.

Based on the results of these experiments:

* SKEValue is larger than REPOValue; therefore, it dominates the multiplication. Thus, "1" has been added to the REPOValue.

(1) Sparck Jones's inverse document frequency weighting technique was the most efficient even though the technique is relatively simple. It produced the best recall and precision values for the 22 questions posed to the test collection.

(2) The proposed "term significance" weighting technique produced better results than the other three more complex weighting techniques which require complicated computations.

(3) The retrieval results of Harter's weighting technique based on the method of moments and modified Harter's weighting technique based on the method of maximum likelihood are about the same, although the method of maximum likelihood is the best one for estimating the parameters.

(4) Salton's term discrimination weighting technique produced the worst recall and precision values.

2. IMPLICATIONS

The present experiments showed that the inverse document frequency weighting and the proposed term significance weighting techniques, which involve relatively simple calculations than the other complex weighting techniques, produced better recall and precision ratios. The relatively high performance of the inverse document frequency weighting and proposed term significance weighting techniques is the most significant result of this dissertation. The weighting values of both simple techniques can be produced by computer processing using relatively little computer time. Therefore, they may be feasible to apply in an operational information system.

BIBLIOGRAPHY

1. G. Salton, C.S. Yand and C.T. Yu, "A theory of term importance in automatic text analysis," *J. of the ASIS 26* (1975):33-44.
2. K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. of Docu. 28* (1972): 11-21.
3. S.P. Harter, "A probablistic approach to automatic keyword indexing: part 1. on the distribution of specialty words in a technical literature," *J. of the ASIS 26* (July-August 1975): 197-206.
4. _____, "A probabilistic approach to automatic keyword indexing: part II. An algorithm for probabilistic indexing," *J. of the ASIS 26* (Sep.-Oct. 1975): 208-88.
5. W.R. Blischke, "Mixtures of descrete distributions," (in) "*Classical and contagious discrete distributions,*" Proceedings of the International Symposium (McGill univ.), Montreal, Canada, August 15-20, 1963, Oxford: Pergamon Press, 1965.
6. A.C. Cohen, "Estimation in mixtures of poisson and mixtures of exponential distributions," *NASA TM X-53245*, 1963.
7. H.P. Luhn, "The automatic creation of literature abstracts", *IBM J. of Rese. and Deve. 2* (1958): 159-165.
8. G. Salton, A. Wong, and C.T. Yu, "Automatic indexing using term discrimination and term precision measurements, *Info. Proc. and Mana. 12* (1976): 43-51.
9. C.K. Schultz, *H.P. Luhn: Pioneer of information science-selected works.* London: Macmillan, 1968.
10. M.L. Pao, "Automatic text analysis based on transition phenomena of word occurrences," *J. of The ASIS 29* (May 1978): 121-124.
11. G. Zipf, *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley Press, Inc., 1949.