ED 146 890 IR 005 176

AUTHOR Kraft, Donald H.
TITLE A Comment on a Threshold Rule Applied to the Retrieval Decision Model. Technical Note.
PUB DATE Nov 75
NOTE 12p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Indexing; *Information Retrieval; Information Science; *Information Systems; *Information Theory; Models; Relevance (Information Retrieval)

ABSTRACT

The retrieval decision problem is considered from the viewpoint of a decision theory approach. A threshold rule based on earlier rules for indexing decisions is considered and analyzed for retrieval decisions as a measure of retrieval performance. The threshold rule is seen as a good descriptive design measure of what a reasonable retrieval system should be able to do. A retrieval mechanism of randomly drawing documents is analyzed to determine the relative strength of the threshold rule. The Neyman-Pearson rule is shown to be a better a priori decision rule for retrieval as it attempts to maximize precision subject to a fixed level of recall, instead of setting a lower limit upon precision, as does the threshold rule. The threshold rule is seen as a necessary, but sufficient, condition for effective retrieval. Finally, a sufficient condition for the threshold rule illustrates the relationship between it and the Neyman-Pearson rule. (Author)

A Comment on a Threshold Rule Applied to the Retrieval Decision Model

Technical Note

Donald H. Kraft

School of Librarianship
University of California, Berkeley
Berkeley, California 94720

November, 1975

Abstract

The retrieval decision problem is considered from the viewpoint of a decision theory approach. A threshold rule based on earlier rules for indexing decisions is considered and analyzed for retrieval decisions as a measure of retrieval performance. The threshold rule is seen as a good descriptive design measure of what a reasonable retrieval system should be able to do. A retrieval mechanism of randomly drawing documents is analyze to determine the relative strength of the threshold rule. The Neyman-Pearson rule is shown to be a better a priori decision rule for retrieval; attempting to maximize precision subject to a fixed level of recall, instead of setting a lower limit upon precision, as does the threshold rule. The threshold rule is seen as a necessary, but not sufficient, condition for effective retrieval. Finally, a sufficient condition for the threshold rule illustrates the relationship between it and the Neyman-Pearson rule.

Alternatives \ States-of-Nature	(G) S_1: Relevant	(1-C) S_2: Not Relevant
A_1: Retrieve	V_1	C_1
A_2: Not Retrieve	C_2	V_2

Figure 1 The Swets Retrieval Decision Matrix

Introduction

Consider the decision of whether or not to retrieve a given document in response to a specific user query for information. This approach to information retrieval systems has been discussed in detail previously [1, 2, 3]. The basic model is illustrated below in Figure 1. Here one can see the alternatives: to retrieve or not to retrieve; the states-of-nature: relevant or nonrelevant; and the evaluations:

V_1 = the value of retrieving a relevant document,

$-C_1$ = the cost of retrieving a nonrelevant document,

$-C_2$ = the cost of not retrieving a relevant document, and

V_2 = the value of not retrieving a nonrelevant document.

We shall assume that value always exceeds cost, i.e.,

$V_1 + C_1 > 0$, $V_2 + C_1 > 0$, $V_1 + C_2 > 0$, and $V_2 + C_2 > 0$; which seems intuitively [2].

Assuming that the case is of decision making under risk, one has G = generality = Pr(relevant), with $G \in (0,1)$. It has been shown [2] that one retrieves a document if and only if $E(A_1) = V_1G + (-C_1)\ (1-G) > E(A_2) = (-C_2)G + V_2\ (1-G)$, or

$$G > (V_2+C_1)/(V_1+C_1+C_2+V_2). \qquad (1)$$

In order to be more generally useful consider a retrieval system that assigns to each document a value of a random variable Z, called a retrieval status value, that attempts to estimate the document's potential relevance to a specific user query [2]. We can now define:

$P(Z) = Pr(\text{relevance}/Z)$,

$f_1(Z) = Pr(Z/\text{relevance})$,

$f_2(Z) = Pr(Z/\text{nonrelevance})$,

$f(Z) = Pr(Z) = f_1(Z)G + f_2(Z)(1-G)$, and

$L(Z) = f_1(Z)/f_2(Z)$ = likelihood ratio.

It can be shown [2] that the decision rule dictating retrieval can now be restated as:

$P(Z) > (V_2+C_1)/(V_1+C_1+C_2+V_2)$, or

$$L(Z) > (V_2+C_1)\ (1-G)/\ (V_1+C_2)G = K. \qquad (2)$$

This allows for the definition of

R = range of Z for retrieval $= \{Z \mid L(Z) > K\}$.

Evaluation Mechanism

The retrieval system can be evaluated in terms of its effectiveness in retrieving relevant documents. Consider the traditional performance measures:

Re = recall = Pr(retrieve|relevance) $= \sum_{Z \in R} f_1(Z)$,

Fa = fallout = Pr(retrieve|nonrelevance) $= \sum_{Z \in R} f_2(Z)$, and

Pn = precision = Pr(relevance|retrieve)

$= ReG/Pr(\text{retrieve}) = ReG/(ReG+Fa(1-G))$.

The Neyman-Pearson criterion [2] is

$\text{Max } (1-Fa) \quad \text{s.t. } (1-Re) = \alpha$,

which is equivalent to maximizing precision subject to a fixed level $(1-\alpha)$ of recall.

The Threshold Rule

Maron [4] has suggested a rule for indexing decisions that an analog may have applicability here for retrieval decisions. In our terms for the retrieval decision problem, we have as the analog of Maron's rule, the statement:

> The act of retrieval should add information about the relevance of each document. This translates to the rule that given a set of users with similar queries for information that would be shown a given document, that document should be retrieved if the proportion of the users that would find the document relevant exceeds the probability that any document selected at random would be relevant to the query.

In terms of the retrieval decision model, the rule implies that

$$\Pr(\text{relevance} \mid \text{retrieve}) = Pn > \Pr(\text{relevance}) = G. \qquad (3)$$

This formula can be manipulated algebraically to yield:

$$Re = \Pr(\text{retrieve} \mid \text{relevance}) > \Pr(\text{retrieve}) = ReG+Fa(1-G), \qquad (3a)$$

$$Pn = \Pr(\text{relevance} \mid \text{retrieve}) > \Pr(\text{relevance} \mid \text{not retrieve}) \qquad (3b)$$

$$= (1-Re)G/(1-ReG-Fa(1-G)), \text{ and}$$

$$Re = \sum_{Z \in R} F_1(Z) > Fa = \sum_{Z \in R} f_2(Z). \qquad (3c)$$

Note that this threshold rule does not seem so much an a priori decision rule to determine whether or not to retrieve a given document in response to a specific user query. Rather, is is an after-the-fact descriptive design measure to evaluate the overall retrieval mechanism, based on a very rigorous and common sense approach.

Random Retrieval

Consider the situation of randomly drawing a subset of the documents in the collection as an act of retrieval. This seemingly poor method of retrieval should provide a baseline for comparing retrieval decision rules. Let

N = number of documents in the collection,

R = number of relevant documents in the collection,

n = number of documents randomly retrieved, and

r = number of relevant documents randomly retrieved.

We shall assume that N, R, and n are fixed and known. This, incidentally, is at variance with Cooper [5], who assumes that N, R, and r are fixed, in a model that considers minimizing (n-r), the search length. It is noteworthy that this is equivalent to maximizing precision for a fixed level of recall [2]. For our situation, we have:

$G = R/N$,

$Re = r/R$,

$Fa = (n-r)/(N-R)$, and

$Pn = r/n$.

In the situation of a random retrieval mechanism, r becomes a random variable having a hypergeometric probability distribution; i.e.,

$$Pr(r) = \binom{R}{r}\binom{N-R}{n-r} \Big/ \binom{N}{n}, \quad r = 0,1,2,\ldots,\min(n,R).$$

This means that

$$u = E(r) = nR/N = nG. \qquad (4)$$

The threshold rule is now written as:

$$Pn = r/n > G = R/N, \text{ or}$$

$$r > nG = u. \qquad (5)$$

Since the hypergeometric distribution is symmetric around its mean, the probability that a random draw will satisfy the threshold rule, which is the equivalent to the probability that a hypergeometric random variable exceeds its mean, its approximately equal to one-half. Thus, the threshold rule is not an especially strong discriminating decision rule for retrieval, in that even a random draw of documents can satisfy it nearly half the time. This implies that the threshold rule is a minimum requirement for retrieval systems, a necessary but not sufficient condition for effective retrieval.

Suppose we assume that $n \leq R$. Further, suppose that for each of the n documents retrieved, there is a constant probability p of the document being relevant. Then, r has a binomial probability distribution, i.e.,

$$Pr(r) = \binom{n}{r} p^{r} (1-p)^{n-r}, \quad r = 0,1,2,\ldots,n; \text{ and}$$

$$E_b(r) = np.$$

For the system to be more effective than a mere random draw, we would require

$$E_b(r) = np > u = nG, \text{ or}$$

$$p > G. \qquad (6)$$

Note that

$$p = Pr(\text{relevance} \mid \text{retrieve}) = Pn, \qquad (7)$$

so that formula (6) is equivalent to formula (3), the threshold rule. The threshold rule implies that a retrieval system should be able to generate a relevant document with greater frequency than a random draw, which is eminently reasonable.

The Neyman-Pearson Rule

The Neyman-Pearson lemma of statistical decision theory can be used [2] to show that the Neyman-Pearson criterion, discussed above, can be optimized by retrieving all documents with a retrieval status value Z in the range

$$R = \{Z \mid L(Z) > K\}. \qquad (8)$$

K is now determined by the formula

$$Re = \sum_{Z \in R} f_1(Z) = 1-\alpha. \qquad (9)$$

Let us recall that the threshold rule can be stated as

$$Re = \sum_{Z \in R} f_1(Z) > Fa = \sum_{Z \in R} f_2(Z). \qquad (3c)$$

We can now derive a sufficient, but not necessary, condition for the threshold rule. If

$$L(Z) = f_1(Z)/f_2(Z) > K \geq 1 \quad \forall Z \in R, \qquad (10)$$

then the threshold rule will hold. Equating the constant K of formula (8) to that of formula (2) yields:

$$K = (V_2+C_1)(1-G)/(V_1=C_2)G \geq 1, \text{ or}$$

$$G \leq (V_2+C_1)/(V_1+C_1+C_2+V_2). \qquad (11)$$

Note that formula (11) is just the opposite result of formula (1). What this says is that for small enough G, the Neyman-Pearson rule will satisfy the threshold rule. This seems reasonable, in light of formula (3). For large G, the situation is dependent upon the value of α and the specific values of $f_1(Z)$ and $f_2(Z)$ for the various values of $Z \in R$. In point of fact, it would be a most perverse retrieval system that would satisfy the Neyman-Pearson rule but not allow the threshold rule to be satisfied for a "reasonable" level of recall. The real difference here is that the Neyman-Pearson rule maximizes precision subject to a fixed level of recall, while the threshold rule imposes a minimum level on precision.

Summary and Conclusions

A threshold rule has been analyzed as a useful tool to evaluate information retrieval systems. It provides a good descriptive design measure of what a reasonable retrieval mechanism should be able to accomplish. Yet, the threshold rule is not seen as an especially strong discriminating rule for a priori retrieval decisions, since the rule can be satisfied nearly half the time by a random draw.

The threshold rule can be analyzed in terms of the retrieval decision model, derived from the assumptions for the binomial probability distribution case of the random draw. When compared to the Neyman-Pearson rule, the threshold rule is not seen as having the ability to compute the extent to which precision can be maximized. The Neyman-Pearson rule is shown to be a better a priori decision rule for retrieval. Moreover, the Neyman-Pearson rule is a sufficient condition for the threshold rule, for appropriate values of generality. Thus the threshold rule provides a necessary, but not sufficient, measure of the least a retrieval system should be able to achieve.

Bibliography

1. Swets, J.A., "Information Retrieval Systems", Science, vol. 241, 1963, pp. 245-50.

2. Kraft, D.H. and A. Bookstein, "Evaluation of Information Retrieval Systems: A Decision Theory Approach," technical paper, School of Librarianship, University of California, Berkeley, submitted to Journal of the American Society for Information Science, 1975.

3. Kraft, D.H., "A Decision Theory of the Information Retrieval Situation: An Operations Research Approach", Journal of the American Society for Information Science, vol. 24, 1973, pp. 368-76.

4. Maron, M.E., personal communication, School of Librarianship, University of California, Berkeley, 1975.

5. Cooper, W.S., "Expected Search Length: A Single-Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems", American Documentation, vol. 19, 1968, pp. 30-41.

6. Cooper, W.S., personal communication, School of Librarianship, University of California, Berkeley, 1975.