Information Processing & Management, Vol. 12, pp. 327–331. Pergamon Press 1976. Printed in Great Britain

AN EVALUATION OF GOFFMAN'S INDIRECT RETRIEVAL METHOD

W. B. CROFT and C. J. VAN RIJSBERGEN
University of Cambridge Computer Laboratory, Corn Exchange Street, Cambridge CB2 3QG, England

(*Received* 1 *March* 1976)

Abstract—The indirect retrieval method proposed by Goffman is outlined and some similarities to other retrieval methods are indicated. The method is then evaluated and the results are compared with those obtained on the same document collection with cluster-based retrieval using single-link clustering.

The comparisons show that although the effectiveness of the indirect retrieval method can be comparable to cluster-based retrieval, the efficiency is lower.

1. INTRODUCTION

This paper reports the results of an experiment with a document retrieval method proposed by GOFFMAN[1]. The theoretical basis and retrieval procedure of the method are interesting yet some time has passed since its publication and no significant retrieval tests have been published.

The *indirect retrieval method* is based on a theory of conditional relevance of documents. An answer to a query is defined as a sequence of documents such that the conditional probability of relevance of a document in the sequence, given that the preceding document is relevant, is greater than some threshold value. The sequence is "closed" in that there are no other documents which could extend it at either end.

Goffman defines the sequence of documents in terms of a *communication chain* as follows:

Given d_i, d_j documents from the set D of all documents in the file,
p_{ij} the probability that d_j is relevant, given that d_i is relevant, and
ξ_0 the threshold probability

then we have the following definitions.

(1) If $p_{ij} > \xi_0$ then d_i is said to *converse* with d_j, denoted d_iCd_j.

(2) A sequence of documents d_i to d_j such that $d_iCd_{i+1}C..Cd_{j-1}Cd_j$ is called a *communication chain*; d_i is said to *communicate* with d_j, denoted $d_i\bar{C}d_j$.

(3) An answer to a query is a communication chain for which there is no $d \epsilon D$ with which the last document in the chain communicates and no $d \epsilon D$ which communicates with the first element.

2. THE RETRIEVAL ALGORITHM

The indirect retrieval method proceeds as follows (for a particular query);

(a) Form the matrix M of conditional probabilities of relevance, i.e. p_{ij} for all d_i, d_j in D. In general, $p_{ij} \neq p_{ji}$.

(b) Reduce to zero all elements of M whose values are less than or equal to the threshold probability ξ_0. The value of ξ_0 is given by the user of the retrieval method and its optimum value is not known before searching.

(c) Find either a document known to be relevant to the query or a document likely to be relevant. This is called the *initial document.*

(d) Generate an answer sequence from M using the initial document.

In step (c), if there is no document known to be relevant, an initial document is found by: (i) Partitioning the documents into classes using an *intercommunication* relation which could be defined by: if $d_i\bar{C}d_j$ and $d_j\bar{C}d_i$ then d_iId_j where I is the intercommunication relation. $\bar{C}$ in one case defines a sequence of documents, in the other case it defines a sequence of the same set of

documents but reversed in order. This is not explicitly stated in [1] but follows from the examples given there. (ii) Choose at random a representative for each class. (iii) If the *initial probability* is defined as the probability of relevance of a document given a particular query, then find the class representative with the highest initial probability. (iv) Find the document with the highest initial probability from the class whose representative was selected in (iii).

The intercommunication classes have been used by some people to classify documents (see [2–4]), although Goffman uses them solely to select an initial document.

3. SIMILARITIES TO OTHER METHODS OF RETRIEVAL

This retrieval method was designed as an alternative to the *direct retrieval method* (more often called the serial, linear or full search method) which retrieves documents solely on the basis of their initial probabilities. The indirect retrieval method avoids what, according to Goffman, are the two main defects of the serial search method. The first is an efficiency consideration in that a serial search must, for every query, examine all the initial probabilities. The second defect is that the serial search method does not make use of the knowledge that the relevance of a document is not absolute but may depend on what information is conveyed by other documents in the file being searched. That is, the serial search method assumes that all documents in the file are independent.

Another retrieval method designed as an alternative to the serial search method is *cluster-based retrieval* [5]. In this method, documents are grouped or clustered on the basis of a dissimilarity measure between documents. During retrieval the clusters are treated as single units, i.e. the method retrieves clusters of documents using the initial probabilities of cluster representatives. Cluster-based retrieval is then more efficient than the serial search method and, because the clusters are constructed using relationships between documents, this retrieval method also avoids the two defects (mentioned above) of the serial search.

3.1 *Relationship to the single-link clustering method*

There are other similarities between Goffman's method and cluster-based retrieval using *single-link clustering* [5]. In [1] Goffman used the following estimate for the conditional probability of relevance (we will assume that documents are represented by sets of index terms);

$$p_{ij} = \frac{|d_i \cap d_j|}{|d_i|}$$

where $|d_i \cap d_j|$ is the number of common index terms in the sets representing documents d_i and d_j and $|d_i|$ is the number of index terms in the set representing document d_i. This is an asymmetric similarity measure between documents. In [5] Van Rijsbergen uses the following dissimilarity measure for single-link clustering, called the *Normalised Symmetric Difference*;

$$NSD_{ij} = 1 - 2\frac{|d_i \cap d_j|}{|d_i| + |d_j|}.$$

For a given pair of documents d_i and d_j the following relationship exists;

$$NSD_{ij} = 1 - \frac{1}{\frac{1}{2}[(1/p_{ij}) + (1/p_{ji})]}.$$

That is, the symmetric dissimilarity value NSD_{ij} can be expressed as a function of the average of the reciprocals of the two asymmetric similarity values.

If both p_{ij} and p_{ji} are greater than a certain threshold probability value T, then it can be shown that NSD_{ij} will be less than $1 - T$, although the converse is not necessarily true. This means that, at threshold level T, documents in the same intercommunication class will also be in the same single-link class. The classes will not in general be identical because for some cases the NSD_{ij} value will be less than $1 - T$ even though one of p_{ij}, p_{ji} may be less than T.

This relationship only holds for the particular estimate of the p_{ij} values used in [1].

4. THE EXPERIMENT

The indirect retrieval method was tested using the Cranfield 200 document collection[6] which consists of 200 documents and 42 queries with relevance judgements. The programmed retrieval algorithm used the estimates given in [1] for p_{ij} and the initial probability. Each retrieval run of the set of queries used a fixed threshold probability. Goffman assumes that a different ξ_0 is used for each query but as there is no method of calculating the best ξ_0 for a given query, this was not feasible. It is also unreasonable from a computational viewpoint because for each threshold probability the intercommunication classes must be recalculated.

The results of the indirect retrieval method are compared with Van Rijsbergen's results[7] obtained with a single-link clustering of the Cranfield 200 collection. The single-link clustering is hierarchic, that is, it can be represented as a tree structure where the leaves are documents and the nodes represent clusters containing the documents which can be accessed from that node. A number of different *search strategies* which reflect different user needs can be implemented using the tree structure (hereafter called the *single-link hierarchy*). The three search strategies used for comparisons in the next section are;

(a) *a narrow search*—this search follows a "narrow" path (i.e. examines only a few nodes) through the hierarchy from the "root" downwards. It is precision oriented.

(b) A *broad search*—this follows a "broad" path through the hierarchy and is recall-oriented.

(c) A *bottom-up search*—a search which uses a known relevant document to start at the document level of the hierarchy and searches upwards.

4.1 *The evaluation method*

The measure used for the effectiveness of the retrieval method is as follows (VAN RIJSBERGEN[5] describes and justifies this measure); Given the precision P and recall R for the set of documents retrieved by a query, the measure (E) is a weighted combination of P and R

$$E = \frac{1}{\alpha(1/P) + (1-\alpha)(1/R)} \quad 0 \leq \alpha \leq 1$$

where α is a parameter giving the relative weight a user may attach to recall and precision. For convenience α has been transformed by $\alpha = 1/(\beta^2 + 1)$ so that the values of β range from 0 to ∞. $\beta = 1$ corresponds to attaching equal importance to recall and precision, and $\beta = \frac{1}{2}$ or 2 corresponds to attaching half or twice as much importance to recall as to precision respectively. Also, note that the smaller E the more effective the retrieval. The actual E values for a particular run will be summarised by plotting them as a cumulative frequency distribution. The higher and more skewed to the left the distribution is, the more effective the retrieval.

5. RESULTS

Table 1 gives the average E values ($\beta = \frac{1}{2}$, 1, 2) for the indirect retrieval method using three different threshold probability values. Figure 1 compares the cumulative distributions for $\beta = 1$. The number of intercommunication classes generated for each threshold value is also given.

It can be seen that the choice of threshold probability is critical for good performance. The best results were obtained for $\xi_0 = 40$ but at this level the number of classes is very high with a resultant loss in efficiency.

Figure 2 compares, using precision-oriented evaluation, the narrow search strategy used on the single-link hierarchy with the $\xi_0 = 40$ indirect retrieval method.

Figure 3 compares, using recall-oriented evaluation, the broad search strategy used on the single-link hierarchy with the $\xi_0 = 40$ indirect retrieval method.

Figure 4 compares the bottom-up search strategy used on the single-link hierarchy with the $\xi_0 = 40$ indirect retrieval method where, for both methods, the relevant document with the highest initial probability was used as an initial document.

Table 2 summarises the results using average E values and a column headed "overhead". The figures in this column refer to the number of query-class representative similarity calculations performed, expressed as a percentage of the number of documents in the file (the figures for the cluster-based retrieval method using single-link have been taken from CROFT[8] which used a

Table 1. Results of indirect search method

Threshold probability	Average E			No. classes
ξ_0	$\beta = \frac{1}{2}$	1	2	
30	0.85	0.82	0.78	33
35	0.72	0.72	0.71	76
40	0.65	0.67	0.67	121

Table 2. Comparison with Van Rijsbergen's results

Retrieval method	Average E	Overhead
	($\beta = \frac{1}{2}$)	
Indirect	0.65	62% (plus matrix searching)
Cluster-based (narrow search)	0.64	5%
	($\beta = 2$)	
Indirect	0.67	62% (plus matrix searching)
Cluster-based (broad search)	0.66	<20%
	($\beta = 1$)	
Indirect (using relevant document)	0.44	0% (plus matrix searching)
Cluster-based (bottom-up search)	0.43	<1%

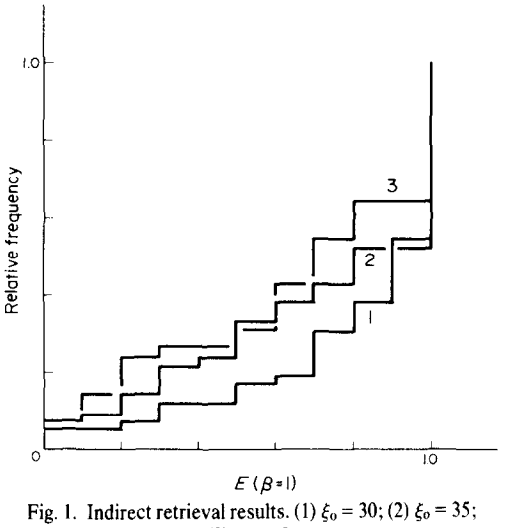

Fig. 1. Indirect retrieval results. (1) $\xi_0 = 30$; (2) $\xi_0 = 35$; (3) $\xi_0 = 40$.

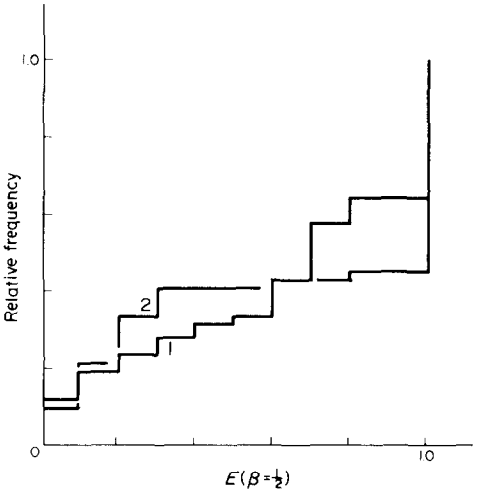

Fig. 2. Indirect(1) vs narrow search(2).

different document collection, but this should still be a good estimate for the types of search strategies used here). This percentage is an approximation of the overhead incurred using these retrieval methods after the initialization has been done. In the case of the indirect retrieval method, the initialization involves forming the matrix M and constructing the intercommunication classes; for cluster-based retrieval the initialization involves forming the half-matrix of dissimilarity values and constructing the single-link hierarchy. In both cases, the processing time required for the initialization is of order n^2, where n is the number of documents in the file.

6. DISCUSSION

The effectiveness of the indirect retrieval method was similar to that of cluster-based retrieval using various search strategies on a single-link hierarchy. Also, because the results do not depend

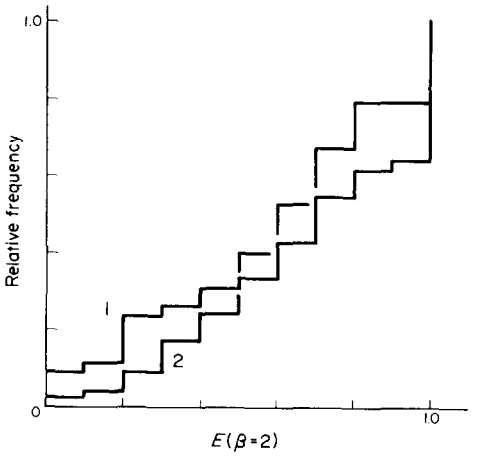

Fig. 3. Indirect(1) vs broad search(2).

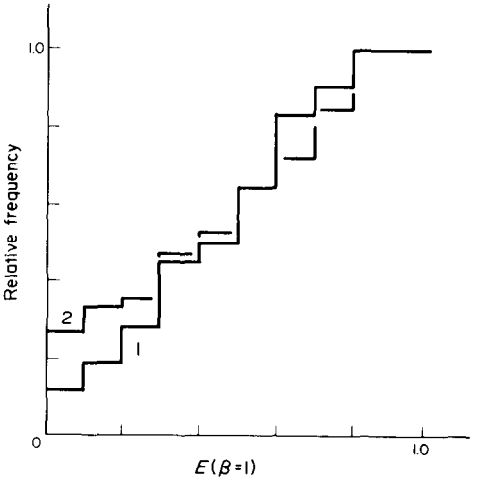

Fig. 4. Indirect(1) vs bottom-up search(2).

on different search strategies, the user of the indirect search method does not have to specify an interest in recall or precision. However, difficulties arise when practical usage of the indirect method is considered. Firstly, there is the matter of giving a value for the threshold probability. Secondly, and most importantly, there is the problem of efficiency. The overhead shown in Table 2 would be prohibitive for large files of documents. Even if a lower threshold probability is used, the overhead does not decrease drastically (for instance, the overhead for $\xi_0 = 30$ is 42%) but the effectiveness drops considerably. The overhead is reasonable only when relevant documents are available. Forming the answer sequence involves examining elements of the matrix M. Thus the entire matrix must be stored somewhere. This again is a prohibitive restriction for large files. For cluster-based retrieval using the single-link hierarchy, the matrix of dissimilarity values is never required to be stored. During construction of the single-link hierarchy, the dissimilarity values can be used as they are generated and then discarded.

Acknowledgements—The authors wish to thank Dr. KAREN SPARCK JONES for suggesting the experiment. The work was conducted while W. B. CROFT was in receipt of a C.S.I.R.O. Postgraduate Studentship and C. J. VAN RIJSBERGEN was holding a Royal Society Scientific Information Research Fellowship.

REFERENCES

[1] W. GOFFMAN, An indirect method of information retrieval. *Inform. Stor. Retr.* 1969, **4**, 361–373.
[2] J. C. BAUGHMAN, A structural analysis of the literature of sociology. *Library Q.* 1974, **44**, 293–308.
[3] J. C. DONOHUE, A bibliometric analysis of certain information science literature. *JASIS* 1972, **23**, 313–317.
[4] A. H. SHIMKO, An experiment with semantics and Goffman's indirect method. *Inform. Stor. Retr.* 1974, **10**, 387–392.
[5] C. J. VAN RIJSBERGEN, *Information Retrieval.* Butterworths, London (1975).
[6] C. W. CLEVERDON, J. MILLS and M. KEEN, *Factors Determining the Performance of Indexing Systems.* ASLIB Cranfield Project, Cranfield (1966).
[7] C. J. VAN RIJSBERGEN, Automatic information structuring and retrieval. Ph.D. Thesis, University of Cambridge (1972).
[8] W. B. CROFT, Document clustering. M.Sc. Thesis, Monash University, Melbourne (1975).