EXPERIMENTS WITH CITED TITLES FOR AUTOMATIC DOCUMENT INDEXING AND SIMILARITY MEASURE IN A PROBABILISTIC CONTEXT

K. L. Kwok
Computer Science Department, Queens College,
City University of New York, Flushing, NY 11367.

1. INTRODUCTION

Most people will agree that the ultimate goal of an information retrieval (IR) system would be to incorporate as much artificial intelligence techniques as possible so that the system would be able to reply to queries with specific answers (so called fact retrieval). Current capabilities for large scale general purpose retrieval however is still at the stage of document retrieval level, namely retrieving documents that most probably contain answers to a query. This may not necessarily be a transient stage at all, as has been pointed out that document retrieval capability may serve as a pre-processing stage to the more detailed fact retrieval, even if the problems associated with the latter are assumed solved.

In an automatic document indexing and retrieval environment, one generally processes the natural language sentences from a query or from the title and abstract of a document into isolated word stems (after removing stop words), which are then used as representation of the query or document. This is the content term selection stage of the automatic indexing procedure. The next stage may be regarded as making the evidence carried by each term more specific via various term-weighting schemes. Since the words of the original sentences or phrases have been isolated, they lose their meaning in context. They become indicators or symbols of content rather than carriers of content, which is actually what indexing is about. As long as we are using these terms as indicators, then there are other indicators within a document that one can use. It is well-known for scientific documents that their citing or cited documents are highly specific indicators of the source document content. In particular, when two source documents share a number of identical references or citations, the probability of them dealing with the same topical content would be high. When cited documents are used, this sharing is called bibliographic coupling [KESS63]; when citing documents are used it is called co-citation [SMAL73]. In these methods it is inherently assumed that the cited

© 1985 ACM 0-89791-159-8/85/006/0165 $00.75

or citing document identities (these may be coded as Journal name, year, page or unique document numbers assigned by the system) are the indexing terms of the source. They are language independent, no words need to be used. They are in effect coded content symbols or indicators that are both concise and precise, and have been found to be useful in query inaependent circumstances such as document classification, research front identification, impact factors of important papers or journals, author evaluation tools or SDI services [GARF79, SCHI71, SMAL74, PINS76]. In circumstances of answering ad hoc queries their usefulness may be more restricted. Query formulation then means that the user need has to be expressed as a few (at least one) known relevant documents, which is equivalent to having solved the initial search problem. The handling of these document identity codes as indexing terms would also be quite alien to the normal user, and may be more error-prone than the handling of English words. For measuring the similarity between two documents the number of identical document codes is invariably used. If it is larger than a threshold count, the two items are regarded as relevant. This counting measure however has intuitive' but no theoretical justification.

In [KWOK75, KWOK84, KWOK85] we have suggested the use of some characterizing properties of the cited or citing documents instead of the raw identity codes as indexing aids to the source. In particular, if these properties are the normal English words, we would then have restored the more user-oriented vocabulary as indexing terms, yet still retaining the use of the citing/cited relationship between documents. The most readily available and least expensive characterizing property of a document is its title terms, which therefore leads us to the use of the cited or citing title words for source document indexing. Title terms have been studied previously and recently [BUXT77, DIEN84], and they have been found on the whole to be reflective of document content. Moreover, considering the set of title terms accumulated from the approximately ten to thirty relevant and related documents that the author(s) (who is an expert on the topics of the document) cites or is cited by, and also using the source document abstract as a selection filter, the probability of obtaining good relevant document representation would be high. In addition, there is theoretical justification for doing this based on Bayes' decision theory [KWOK85].

Because of the nature of natural language words, their use in our context would invariably lead to a loss in precision compared with using the cited or citing document identity codes directly, although recall would probably be improved. However, since author-provided topical relevance information is available for them, we can relevance weight each of these title terms selected, as in the theory of probabilistic retrieval. This way the loss in precision may be restored. Moreover, a similarity measure based on the probability of relevance between query and document, or between documents may also be derived.

In this paper, we would like to report some trial experimental results using cited title terms only, for the study of indexing and document-document similarities. Section 2 summaries the theoretical formulations, and Section 3 presents the methods used and the results.

2. THEORY

Indexing of a document probably is the most important procedure one takes in IR. In the environment of peer-reviewed scientific papers, the publishing of such a document generally means that the author has done a fair amount of survey of related and relevant articles, part of which naturally appears as references in support of the (source) document. This situation may be viewed as the author(s) having an information need, equivalent to a mental 'query', and the source together with its cited documents may be regarded as 'retrieved' as 'answers' to this 'query'. This model may be employed for deriving the indexing of this mental 'query' from its 'answers', which is then regarded as the indexing of the source document.

In our approach each document x_i of our collection under study (called the source collection of size N) is assumed to be accompanied by a set of c_i relevant cited documents. The set of unique source and cited documents is called the augmented collection of size $N_A <= N + \Sigma_i c_i$, Figure 1. We will assume each source document to have a mental 'query' i behind it, whose content is what the author is trying to write about. The source document x_i and its c_i cited document are then regarded as a relevant answer set to this 'query'. Moreover, if the assumption is also made that no other documents in the augmented collection is relevant to the 'query', we may apply Bayes' decision theory (exactly) as in probabilistic retrieval [ROBE76, RIJS77]. The decision question we ask is as follows: if we were to have a representation of the mental 'query' i, what terms and weightings would it take so that when every document x_j in the collection is compared with this 'query' we would re-form our cluster of source x_i and its cited documents based on the condition $O(R_i/x_j) > 1$, with minimal probability of error? Here $O(R_i/x_j)$ is the odds $P(+R_i/x_j)/P(-R_i/x_j)$ that, given a document x_j, it will have a bigger probability of relevance to i ($+R_i$) than nonrelevance ($-R_i$). The result is that the 'query' i has to be represented by (a) the union of all the title and cited title terms (T u CT), designated as $(y_{i1}..y_{ik}..y_{it})$ with y_{ik}= 1/0 denoting presence/absence of the k-th term in i; and (b) these terms should be weighted by the relevance weights as used in [YU76, ROBE76]:

$$w_{ik} = \log [p_{ik}(1-q_{ik})/(q_{ik}(1-p_{ik})] \qquad (1)$$

where

$p_{ik} = P(y_{ik}=1/+R_i)$, the probability that the k-th
term will occur in the set relevant to i;
$q_{ik} = P(y_{ik}=1/-R_i)$, the probability that the k-th
term will occur in the set not relevant to i;
$1 <= k <= t$ counts the set of t index terms.

The above formula also makes the usual assumption that the index terms are independent in the relevant ($+R_i$) and the nonrelevant sets ($-R_i$).

In real situations, the foregoing assumptions about the cited documents being all relevant to the source, or that they are the only relevant ones in the augmented collection are usually not satisfied. Approximating procedures then have to be devised to pick those terms in (T u

CT) that are useful for indexing, and those among the c_i+1 documents that are relevant and useful for weighting purposes. These procedures are explained in detail in Section 3.

Once a probability-weighted representation for each source document (or in our model, the 'query' behind it) has been obtained, we can use it for computing a similarity measure between each document and a query, or between documents. This investigation only looks at the latter. Various forms of similarity measures can be considered and they are derived in detail in [KWOK85]. The most useful seems to be a symmetric form which also allows all pairs of similarity measures to be directly comparable. Its basis originates from the following arguments. After i has been indexed and weighted, we can investigate if j is relevant to i by asking the following:

If $O(R_i/j) > 1$, then decide j relevant to i.

Here j is considered to be a cluster of c_j+1 items, each being independent and whose probability of relevance to i can be evaluated. This necessarily leads to a nonsysmmetric similarity measure which has the interpretation of the total probability of relevance of j to i. For a symmetric measure an obvious modification would be:

If $O(R_i/j)O(R_j/i) > 1$, then decide i, j mutually relevant;

which takes into account both i and j as focii of evaluation. This however leads to measures with scaling factors linearly dependent on the number of components in i and j. This can be overcome by considering the odds per relevant component (actually log odds per component) by introducing the factors $1/(c_i+1)$ and $1/(c_j+1)$. In addition, to make the measure directly comparable to each other we also introduce the normalizing self odds factors $O(R_i/i)$ and $O(R_j/j)$. The final result is the following decision equation:

$$\text{If } \frac{O(R_i/j)^{1/(c_j+1)}}{O(R_i/i)^{1/(c_i+1)}} \frac{O(R_j/i)^{1/(c_i+1)}}{O(R_j/j)^{1/(c_j+1)}} > \text{Threshold,}$$

then decide i and j relevant.

Here $O(R_i/j)$, for example, is the odds that given j it is relevant to i. The number of components making up j, namely c_j+1, enters explicitely in the decision rule. After taking logarithm and using Bayes' Theorem, we arrive at a symmetric probability-based similarity measure between two source documents as follows:

$$v(i,j) = \Sigma_k \; (t_{jk}/(c_j+1)-t_{ik}/(c_i+1)](w_{ik}-w_{jk}) \qquad (2)$$

Here, t_{jk} are the title frequency of term k in the cluster of relevant documents that define j (i.e. the number of times term k occurs in the c_j+1 titles of cluster j; and likewise for i), and the sum of k is over all terms belonging to both i or j. This is the formula we have used for our experimental results.

3. EXPERIMENTAL METHODS AND RESULTS

Unlike titles and abstracts, use of cited titles for indexing is uncommon, and databases of documents containing these items are not readily available. One may have a source document collection with cited or citing document codes (for example from ISI), but these may link to documents outside the collection, and therefore the citing or cited titles would not be complete, unless one has access to some large current and retrospective databases that may conceivably contain all possible citing or cited titles of the source collection. Technically, this is quite feasible in a well-designed information system. For our study however, we have manually created two small collections for experimentation. Collection CIS consists of 65 Computer and Information Science (CIS) source documents that were accumulated during classroom projects. They were taken from TODS (Transactions on Database Systems), Data Structure papers of the CACM, and a few others from JACM, JASIS and JDOC. There is no objective judgment, but the topics are quite heterogenous; hence one can judge the topical content without much risk of bias. This collection also has a number of papers that have few (3 to 8) references. Collection MED consists of the 37 medical papers previously studied in [KWOK75]. It contains papers on two topical areas (VW: Platelet Functions and Von Willebrand's Disease; and PY: Effect of Pyrimido-pyrimidine Derivatives on Platelet Functions) obtained by using the clustering procedure of [SCHI71]. Objective relevance judgment as to the correctness of cluster assignment has been obtained for these papers on the scale of four as follows: X (right on the topic), R (related), P (peripherally related) and U (unrelated). Both collections do not have queries. We have used Porter's algorithm [PORT80] as the standard stemming procedure. The collection characteristics are summarised in Figure 2.

The query i behind a source document x_i with c_i cited titles, after indexing will be represented by $(y_{i1}, y_{i2} .. y_{it})$. This is accomplished in two stages: selection and weighting. In theory we take all terms in (T u CT); in reality, because not all cited titles are relevant, selection was done by first forming the two sets of stems: A' = T u A (unique stems from the title and abstract) and CT' = T u CT (unique stems from the title and cited titles). From these the following five disjoint sets were defined:

X : set of overlapping stems (A' n CT');
CT'_n : those left in CT' with frequency >= n, where
n = 2 + TRUNC[(c_i+1)/CITLIM] with CITLIM = 33;
A'_m : those left in A' with frequency >= m, where
m = 3 + TRUNC[ABSTRACT_LEN/ABSLIM] with ABSLIM = 150;
$CT'_{n<}$: remaining stems in CT';
$A'_{m<}$: remaining stems in A'.

The normal thresholds for n, m are 2 and 3 respectively. If the number of references, or the length of abstract equals or exceeds CITLIM or ABSLEN respectively, these are raised by 1. They are chosen fairly arbitrarily and serve to suppress terms that may be included later in IDX of Equation 3 just because of high term usage. We have selected the terms in

$$IDX_{i} = CT'_n \text{ u } X \text{ u } A'_m \qquad (3)$$

as the representation of the document. If this set has more than MXIDX=36 terms, we have also raised the value of either n or m by 1 to limit the final number of unique terms to MXIDX or less. This has the effect of normalizing somewhat the length of representations. It is this investigator's feeling that 30 or so index terms chosen carefully should be quite sufficient for topic descriptions, and that more terms may not be necessarily more effective. It should be noticed that the terms in A'_m are only present in the abstract, and not in any of the cited titles of the document in question. However, they appear with sufficient frequency in the abstract that we felt they should be included. There are however few of them as the result in Fig. 3 shows.

For term weighting we have to determine which of the c_i+1 cited titles are relevant to document content. This is important and the argument for the best method of identifying them is an open question. We have used a rather conservative procedure as follows. First a set of kernel terms were picked from the IDX set obtained earlier. These include all the terms in the set X plus all terms in CT'_n with frequency >= r, where r = 2 + TRUNC[$(c_i+1)/14$]. The purpose of r is to limit the kernel terms to the higher frequency ones as the number of titles increases, so that later when we select the relevant titles, we would have a more restrictive procedure. Our policy is to select those titles that contain more than a threshold of these kernel terms as relevant. However, when these kernel terms are ranked by frequency, they may have a Zipfian behaviour with the lowest ranked term(s) proliferating in many titles. We decided to give a score of 1/2 to these less specific kernel terms and a score of 1 to the rest. A title with z stems would be regarded as relevant if the sum of the score of kernel terms >= 0.4z, or >= 2.5 (for long titles); i.e. if 40% of the title terms are kernel terms, or that it has a weight of at least 2.5 kernel terms. All these arbitrary parameters can be varied to affect the number of relevant titles to be selected. Once the number (RT_i) of relevant titles have been identified, we can proceed to count the occurrence of each index term y_{ik} in the set relevant to i, giving us t_{ik}, the title frequency of term k in i, and the conditional probability p_{ik} is estimated as:

$$p_{ik} = P(y_{ik}=1/+R_i) = (t_{ik}+0.5)/(RT_i+1)$$

For $q_{ik} = P(y_{ik}=1/-R_i)$, we have followed the method suggested in [HARP78] and have to accumulate all the title stems in the augmented collection. If the term y_{ik} appears in n_k titles, then the estimate is:

$$q_{ik} = (n_k-t_{ik}+0.5)/(N_A-RT_i+1)$$

The relevance weight w_{ik} of the term y_{ik} in i is then given by Equation (1). These fundamental probabilities and frequencies also allow us to evaluate the document-document similarity v(i,j) as given in Equation (2). Since the use of cited title terms is uncommon we have divided our experimental investigation into three parts, collection characteristics, indexing term characteristics and similarity measures between documents.

3.1 COLLECTION CHARACTERISTICS

Figure 2 shows the characteristics of the collections. The augmented collections are formed by union of all sources and their cited documents, with each represented by its title content terms only. It can be seen that medical literature is comparatively more suitable for our approach than CIS papers. Each medical source paper not only cites more (average 23.1 documents versus 16.7), each title also averages longer (6.9 terms versus 5.1). Out of the 65 CIS papers, 19 have 3 to 8 references (mainly CACM papers). In contrast only 1 out of the 37 medical papers has 8 references only. The tight topical content of the medical collection as against the loose CIS collection can be observed from the drop of a total of 838 to 572 unique references. In Collection CIS this drops modestly from 1075 to 934.

3.2 INDEX TERM CHARACTERISTICS

Figure 3 shows the term selection statistics for the two collections. All numbers are averaged over the collection sizes. It also gives a comparison betweeen our method of indexing (IDX) and the conventional title and abstract (A'= T u A). As can be seen the size of IDX is much more compact, about 40% of that of A'; yet about 70% of IDX overlaps with A'. It is our conjecture that most of the content bearing terms characterizing the documents are already in IDX, and that the set $A'_{m<}$, though large, may not be important for content description. (See also comparison with manual terms later). As stated before, the terms in the set A'_{m} which are the higher frequency terms in the abstract that do not overlap with the cited title terms, are few in number.

We have also gathered the manual indexing terms (MAN) for ACM papers in Collection CIS (only 20 are done so far). These consist of the unique terms of the Computing Review category headings, the general terms and the additional keywords and phrases that normally accompany them. Figure 4a shows how the overlaps between the set of manual indexing terms with IDX as well as with A' are to be counted. In Figure 4b, the averages are tabulated for those available. As can be seen the overlap counts of IDX with MAN are very similar to that of A' with MAN. It must be indicated that the IDX indexing achieves this with a much smaller set of terms compared with A'. If we had used the usual overlap ratio C/(A+B-C), where C is the number of elements in the intersection of set A with set B, the result would be overwhelmingly in favour of IDX. The results in this table may give us a feeling of the quality of indexing with cited title terms.

In Figure 5 we have also plotted the number of (T u CT) terms as well as IDX terms, both against c_i+1. (T u CT) contains all (untruncated) unique content terms obtained by merging all title and cited titles. A plot of this against c_i+1 may show how new terms are introduced within a set of related and relevant titles. If these titles were unrelated we would expect a linear graph (at least for lower values of c_i+1 before significant overlap occurs) with slope about 5.1 and 6.9 for Collection CIS and MED respectively. These are their average unique terms per title, see Fig. 2. In effect the experimental points obey a linear least square fit $y = a + bx$ (with distribution at each point assumed to be Poisson with variance equal to observed count) very well. The slopes of 2.31 (CIS) and 3.37 (MED) are much smaller than for unrelated items, and may serve as a measure of the relatedness of the titles.

The other graph on Fig. 5 is a plot of IDX vs c_i+1. IDX as defined in Equation 3 contains effects of terms from abstracts and truncation procedures. A linear fit is also applied as before, but only for points that are not truncated (i.e. excluded are cases where c_i+1 => CITLIM, or IDX => MXIDX; these points are shown as Δ in the graph). The fit also seems to be well obeyed. Based on our selection process, new terms seems to appear at the rate of less than 1 per title.

For each document x_i we have to select out of the c_i+1 titles those considered relevant for weighting purposes. The procedure outlined earlier was followed and the result tabulated in Figure 6. As can be seen our procedure is conservative; less than 75% of the titles are considered relevant for weighting purposes.

3.3 DOCUMENT-DOCUMENT SIMILARITY MEASURE

For Collection MED a symmetric similarity measure based on Equation (2) of Section 2 is also calculated between every pair of documents. Because objective relevance judgment as to correctness of group assignment is available, we have shown in Fig. 7 the distribution of similarity measures for pairs of documents both chosen from the same group [R-R curves], and for pairs chosen from different groups [N-R curves]. For this purpose, 5 papers in Group VW (leaving 15) and 3 papers in Group PY (leaving 14) were removed from consideration because they were judged to be U (unrelated) ot P (peripherally related). This results in 196 R-R and 210 N-R measures. For comparison, the cosine similarity meausres were also plotted. Fig. 7a gives the distribution for cosine based on indexing using A'=(A u T) (the normal title and abstract). Fig. 7b also uses cosine, but indexing is based on our IDX. Fig. 7c uses IDX and our similarity measure v(i,j). There seems to be a significant improvement in the separation of the R-R and N-R curves when IDX (Fig. 7b) is used instead of A' (Fig.7a). A' generally uses a diverse set of vocabulary as well as longer abstracts; they lead to small cosine values even for R-R measures. IDX on the other hand uses a much tighter vocabulary and also the sizes of representation are smaller; they lead to more higher cosines. This effect is however more pronounced in the R-R than N-R, thereby leading to better separation of the curves. Fig. 7c shows the result of our similarity measure (Equation 2) on IDX terms. It is interesting to see two approximately normal distributions, but the average separation does not seem to be as pronounced as for the cosine of Fig. 7b, but better than for Fig. 7a. If cut-off thresholds are chosen as shown in the figures to differentiate similarity from nonsimilarity, then the number of wrong assignments are about the same in all 3 cases. Results for Collection CIS are not yet available.

4. CONCLUSION

The use of cited title terms for indexing and similarity measures in a probabilistic context has been investigated using two small collections, and the results have to be viewed with this restriction in mind. For general conclusions, much larger databases and much more experimentation have to be done. Our purpose is to see whether the models and formulae that we proposed are reasonable or not. The results seem to show that our theory and approach are correct. There seems to be merit in using our IDX indexing procedure and employing cited titles, compared with normal indexing from title and abstract. Our similarity measure v(i,j) (Equation 2) also gives reasonable results, but it does not seem to perform better than the cosine, as we had hoped. However, our probabilistic approach has more room for improvement, both in theoretical development and in approximation procedures. It seems that larger scale of experimentation is warranted.

ACKNOWLEDGEMENT

The author would like to thank the students in his Information Retrieval classes, especially Mr. Shiaw-Tarn Shu, who have helped in creating the databases and part of the programming.

REFERENCES

[BUXT77] BUXTON, A.B. & MEADOWS, A.J. "Variation in the information content of titles of research papers with time and discipline." Journal of Documentation. 33:46-52; 1977.

[DIEN84] DIENER, R.A.V. "Informational dynamics of journal article titles." Journal of the ASIS. 35:222-227; 1984.

[GARF79] GARFIELD, E. Citation Indexing - its theory and application in science, technology, and humanities. Wiley: N.Y.; 1979.

[HARP78] HARPER, D.J. & van RIJSBERGEN, C.J. "An evaluation of feedback in document retrieval using co-occurrence data." Journal of Documentation. 34:189-216; 1978.

[KESS63] KESSLER, M.M. "Bibliographic coupling between scientific papers." American Documentation. 14:10-25; 1963.

[KWOK75] KWOK, K.L. "The use of title and cited titles as document representation for automatic classification." Information Processing and Management. 11:201-206; 1975.

[KWOK84] KWOK, K.L. "A document-document similarity measure based on cited titles and probability, and its application to relevance feedback retrieval." in Research & Development in Information Retrieval, edited by C.J. van Rijsbergen. Cambridge: Cambridge Universty Press; 1984.

[KWOK85] KWOK, K.L. To be published in Journal of ASIS (1985).

[PINS76] PINSKI, G. & NARIN, F. "Citation influence for journal aggregates of scientific publications." Information

Processing & Management. 12:297-312; 1976.

[PORT80] PORTER, M.F. "An algorithm for suffix stripping." Program. 14:130-137; 1980.

[RIJS77] van RIJSBERGEN, C.J. "A theoretical basis for the use of co-occurrence data in information retrieval." Journal of Documentation. 33:106-119; 1977.

[ROBE76] ROBERTSON, S.E. & SPARCK JONES, K. "Relevance weighting of search terms." Journal of the ASIS. 27:129-146; 1976.

[SCHI71] SCHIMINOVICH, S. "Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm." Information Storage and Retrieval. 10:267-278; 1971.

[SMAL73] SMALL, H. "Co-citation in te scientific literature: a new measure of the relationship between two documents." Journal of the ASIS. 24:265-269; 1973.

[YU76] Yu, C.T. & Salton, G. "Precision weighting - an effective automatic indexing method." Journal of ACM. 23:76-88; 1976.

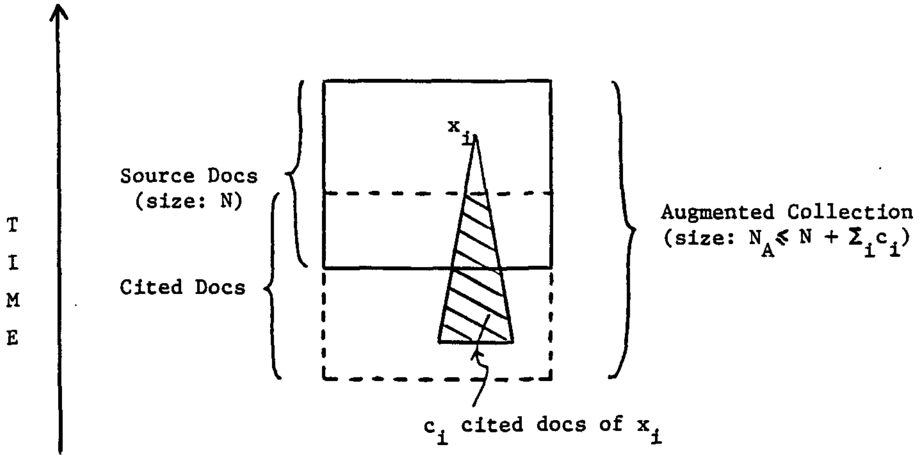

T
I
M
E
Source Docs
(size: N)
Cited Docs
x_i
Augmented Collection
(size: $N_A \leqslant N + \Sigma_i c_i$)
c_i cited docs of x_i

Figure 1: Source and Augmented Collections

Coll.	Source			Augmented			
	# docs N	$\Sigma_i(c_i+1)$	Av(c_i+1) per doc	# unique docs N_A	# title trms/doc	# unique terms	# docs per term
CIS	65	1075	16.5	934	5.1	969	4.9
MED	37	838	22.6	572	6.9	871	4.5

Figure 2: Collection Statistics

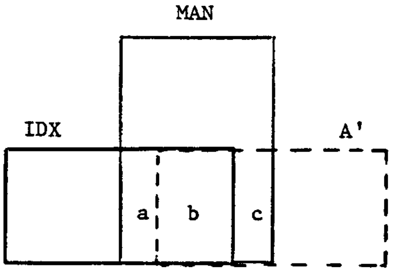
Coll.	$CT'_{n<}$	CT'_n	X	A'_m	$A'_{m<}$	IDX				A'				MAN
						tot docs	#terms per doc	tot terms	# docs per term	tot docs	#terms per doc	tot terms	# docs per term	
CIS	28.2	5.7	13.1	1.02	31.7	65	19.8	482	2.67	65	45.8	1089	2.74	18.4
MED	58.5	9.8	17.9	0.89	27.8	37	28.6	329	3.21	37	46.6	783	2.20	N.A.

CT'

A'

IDX

Figure 3: Statistics of Indexing Terms

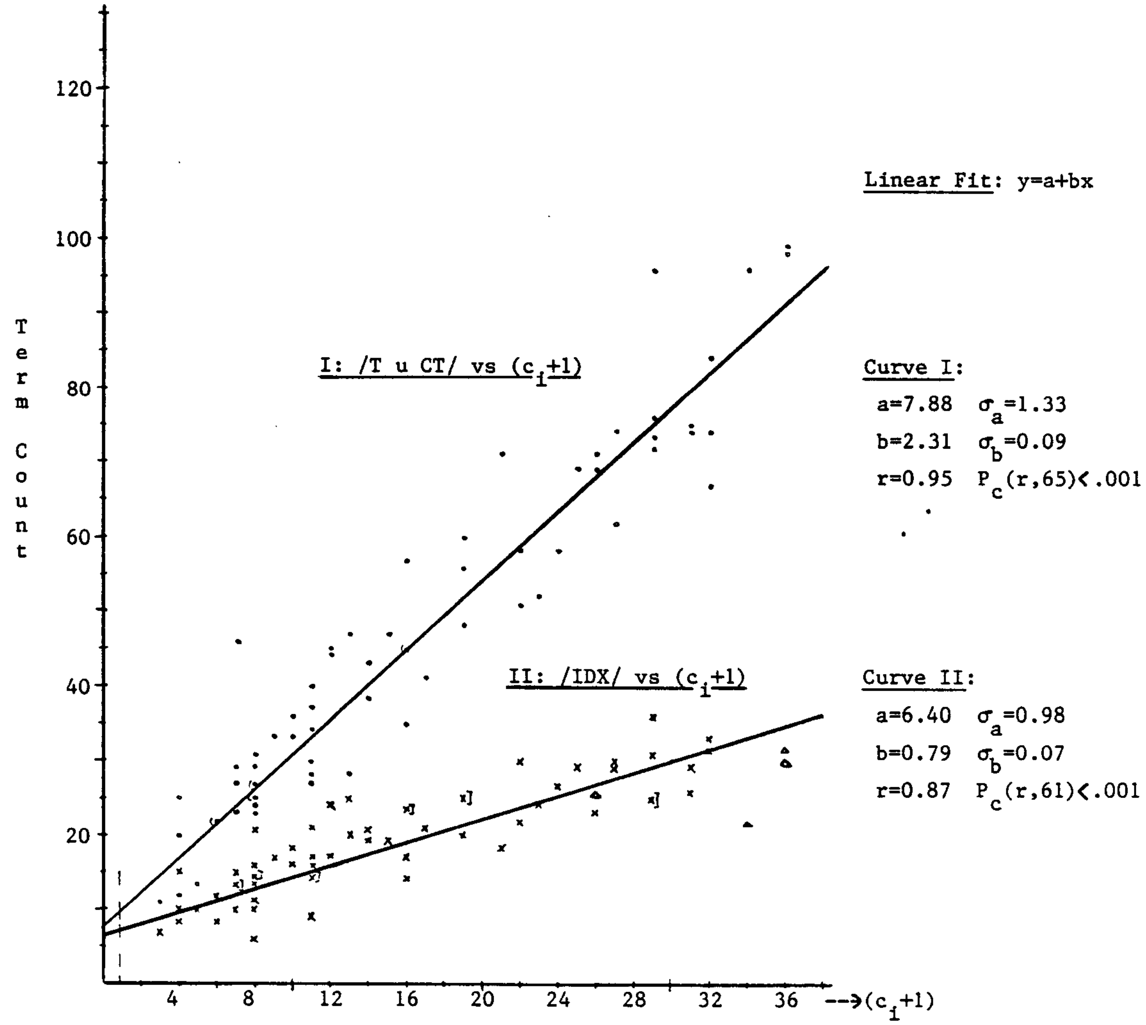


(a)

Coll.	Average				
	a	b	c	$\frac{a + b}{MAN}$	$\frac{b + c}{MAN}$
CIS (20)	1.5	6.55	1.25	0.47	0.46
MED			N.	A.	

(b)

Figure 4: Comparison with Manual Indexing



Linear Fit: y=a+bx

Curve I:
a=7.88 σ_a=1.33
b=2.31 σ_b=0.09
r=0.95 $P_c(r,65)<.001$

Curve II:
a=6.40 σ_a=0.98
b=0.79 σ_b=0.07
r=0.87 $P_c(r,61)<.001$

Figure 5a: Variation of Unique Terms versus (c_i+1) - Collection CIS

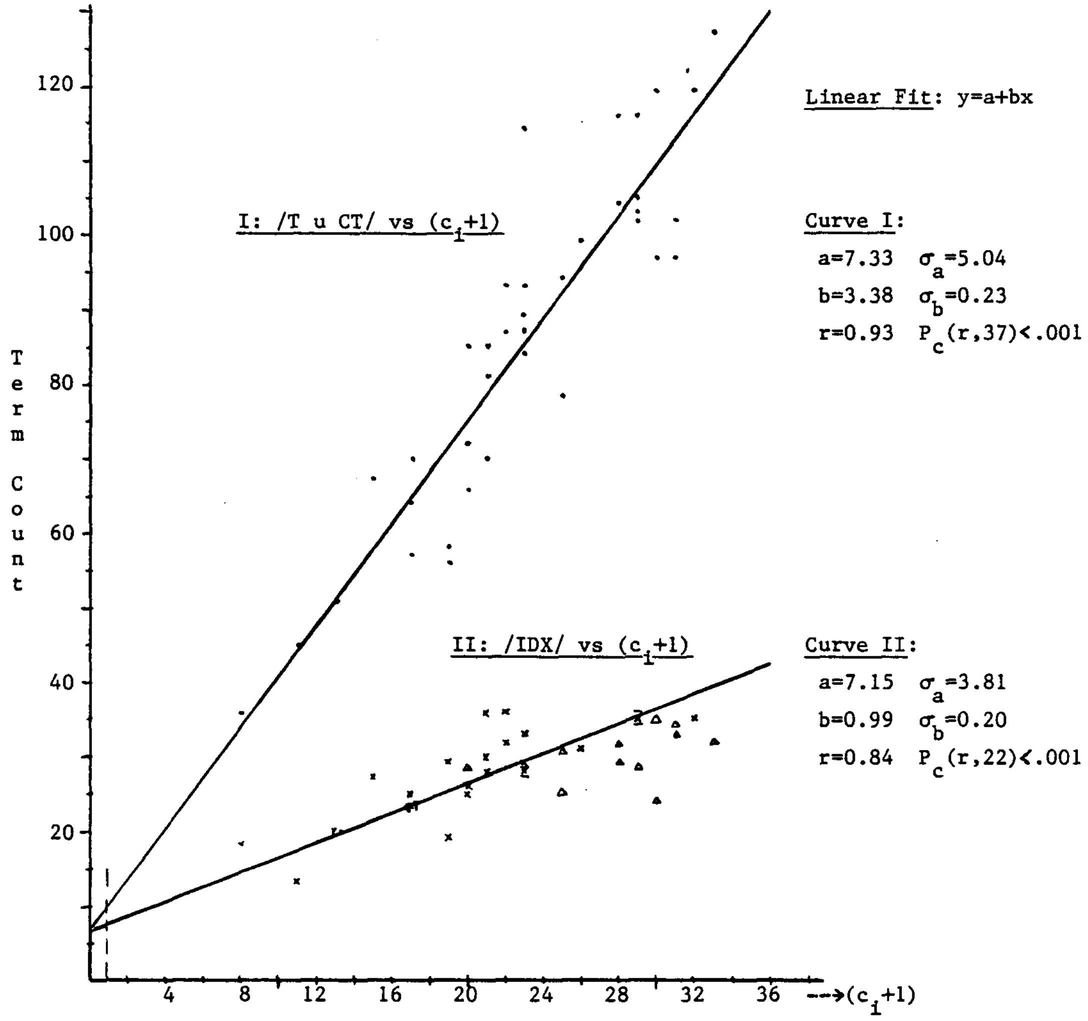


Figure 5b: Variation of Unique Terms versus (c_i+1) - Collection MED

Coll.	Relevant Titles	
	Av. # selected RT_i	Av. % selected $RT_i/(c_i+1)$
CIS	11.0	67
MED	16.3	71

Figure 6: Statistics of Relevant Titles Selected

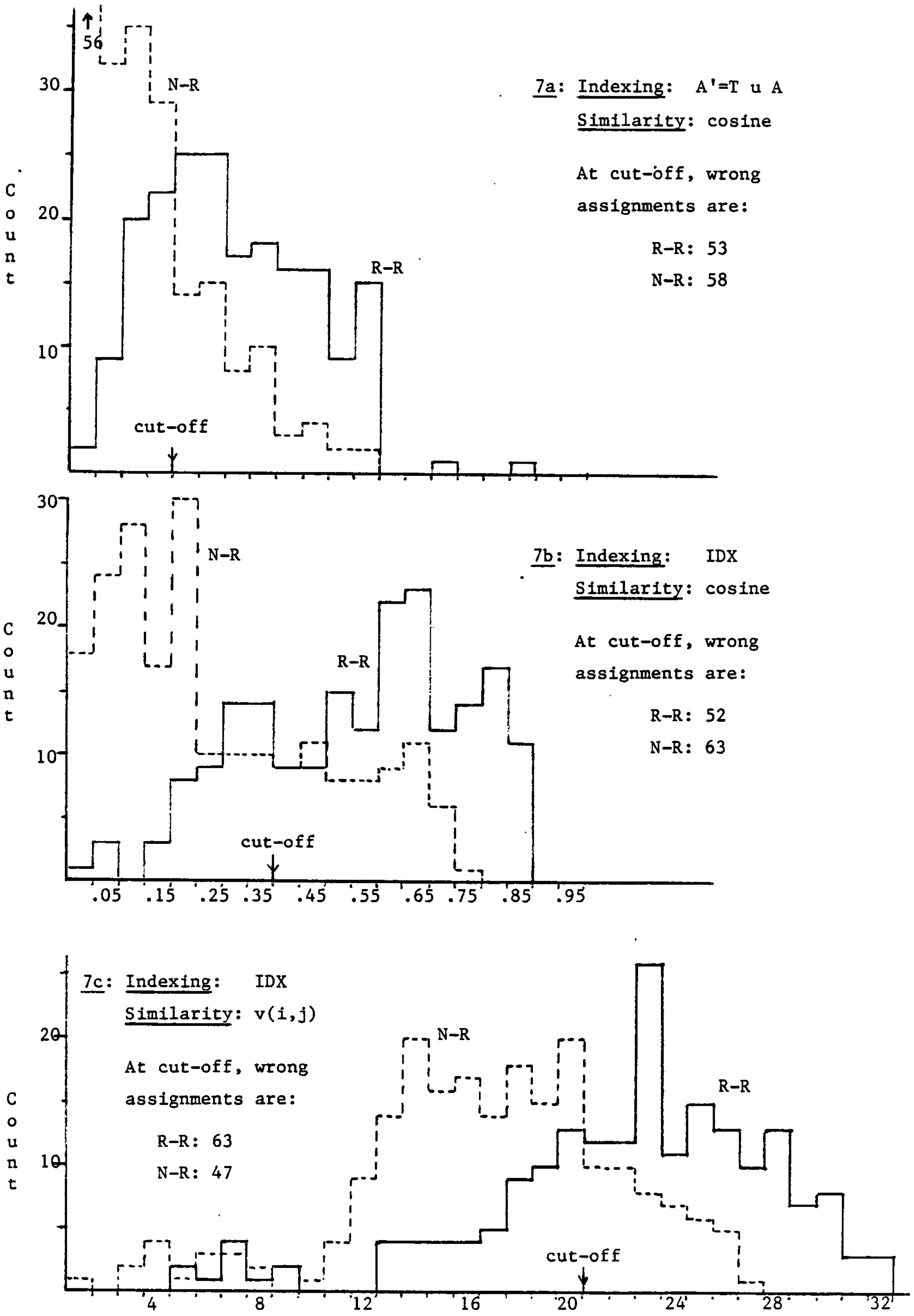


Figure 7: Distribution of Similarity Measures for R-R and N-R Pairs