

110
29

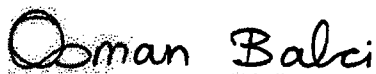
**Regression Analysis of Extended Vectors
to Obtain Coefficients for Use in
Probabilistic Information Retrieval Systems**

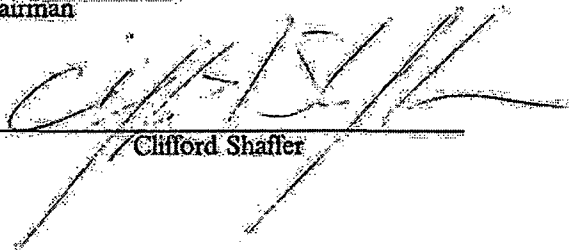
by
Gary L. Nunn

Project submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Master of Science
in
Computer Science and Applications

APPROVED:


Edward A. Fox, Chairman


Osman Balci


Clifford Shaffer

December 7, 1987
Blacksburg, Virginia

EXHIBIT 2027
Facebook, Inc. et al.
v.
Software Rights Archive, LLC
CASE IPR2013-00479

**Regression Analysis of Extended Vectors
to Obtain Coefficients for Use in
Probabilistic Information Retrieval Systems**

by

Gary L. Nunn

Edward A. Fox, Chairman

Computer Science and Applications

(ABSTRACT)

Previous work by Fox has extended the vector space model of information retrieval and its implementation in the SMART system so different types of information about documents can be separately handled as multiple subvectors, each for a different concept type. We hypothesized that relevance of a document could be best predicted if proper coefficients are obtained to reflect the importance of the query-document similarity for each subvector when computing an overall similarity value. Two different research collections, CACM and ISI, each split into halves, were used to generate data for the regression studies to obtain coefficients. Most of the variance in relevance could be accounted for by only four of the subvectors (authors, Computing Review descriptors, links, and terms) for the CACM1 collection. In the ISI1 collection, two of the vectors (terms and cocitations) accounted for most of the variance. Log transformed data and samples of the records gave the best RSQ's; .6654 was the highest RSQ (binary relevance). The regression runs provided coefficients which were used in subsequent feedback runs in SMART. Having ranked relevance did not improve the regression model over binary relevance. The coefficients in the feedback runs with SMART proved to be of limited usefulness since improvements in precision were in the 1-5% range. Although log data and samples of the records gave the best RSQ's, coefficients from log values of all data improved precision the most. The findings of this study support previous work of Fox, that additional information improves retrieval. Regression coefficients improved precision slightly when used as subvector weights. Log transforming the data values for the concept types modestly helped both the regression analyses and the retrieval in SMART.

Acknowledgements

I am grateful to Dr. Edward A. Fox for his patient help and many interesting discussions during the course of this research and my studies. I would to thank Dr. Osman Balci and Dr. Clifford Shaffer for their aid in completing this project. I very much appreciate the many hours of effort that Mr. Whay Lee gave to this study.

I would especially like to thank my wife, Dr. Pamela Garn-Nunn, and our son Bradley, who gave me support and made many sacrifices during this project and my course of study.

Table of Contents

1.0 Introduction	1
1.1 Probabilistic retrieval	1
1.2 Information retrieval in SMART	2
1.3 Research goals	3
2.0 Methods	5
2.1 Division of the collections	5
2.2 Description of the collections and vectors	5
2.3 Descriptive analysis of the data	7
2.4 Obtaining regression coefficients for use in SMART	7
2.5 Testing the usefulness of the coefficients as weights in SMART	9
2.6 Threshold techniques	10
3.0 Results	11
3.1 Description of the data	11
3.2 Linear regressions on the CACM1 collection	19
3.3 Linear regressions on the ISI1 collection	27

3.4	Additional regression techniques	31
3.5	Use of regression coefficients in SMART for CACM1 and ISI1 collections	31
3.6	Use of regression coefficients in SMART for CACM2 and ISI2 collections	32
3.7	Use of threshold techniques	33
4.0	Discussion	40
4.1	Usefulness of concept types as shown by linear regressions	40
4.2	Ranked relevances versus binary relevances	41
4.3	Thresholds as aids in regression	42
4.4	Improvement of retrieval using coefficients	42
4.5	Conclusions and implications for further research	42
	References	44
	Vita	45

List of Illustrations

Figure 1. Histograms of raw versus log terms in sample ISI1 data.	17
Figure 2. Histograms of raw versus log Computing Reviews categories in all CACM1 data.	18
Figure 3. Predicted versus residuals for log sample binary data, CACM1 collection.	24
Figure 4. Predicted versus residuals for log sample ranked data, CACM1 collection.	25
Figure 5. Predicted versus residuals for log sample for the ISI collection.	29

List of Tables

Table 1. Organization of the CACM1 merged data set.	8
Table 2. Descriptive statistics and vector length of raw CACM1 data.	13
Table 3. Descriptive statistics and vector length of raw ISI1 data	14
Table 4. Descriptive statistics of log CACM1 data	15
Table 5. Descriptive statistics of log ISI1 data	16
Table 6. Regression coefficients, ranks and RSQ's for CACM1 ranked relevances.	22
Table 7. Regression coefficients, ranks and RSQ's for CACM1 binary relevances.	23
Table 8. Sums of squares and probability values for CACM1 two-way interactions.	26
Table 9. Regression coefficients, ranks and RSQ's for ISI data.	28
Table 10. Sums of squares and probability values for ISI1 (log sample data) interactions.	30
Table 11. Precision values from base and coefficient runs for the ISI1 collection.	34
Table 12. Precision values from base and coefficient runs for the CACM1	35
Table 13. Precision values from base and coefficient runs for the CACM1	36
Table 14. Precision values from base and coefficient runs for the ISI2 collection.	37
Table 15. Precision values from base and coefficient runs for CACM2 (binary relevance).	38
Table 16. Precision values from base and coefficient runs for CACM2 (ranked relevance).	39

1.0 Introduction

1.1 Probabilistic retrieval

The probabilistic model for information retrieval ((Yu and Salton, 1976 and Robertson and Sparck Jones, 1976) as cited in van Rijsbergen, 1981) assumes that the terms in a query and the terms in a collection of documents are used in an initial retrieval to obtain a sample of the documents. The terms in the sample are then used to estimate the probability that each document in the collection is relevant or not relevant. The collection has usually been indexed by building a vector of terms for each document where the vector consists of binary values (1 for presence - 0 for absence) for all terms in the collection. A document thus is represented as a vector of length n:

`Term(1), Term(2)...Term(n)`

For a given query it is possible to estimate a probability of relevance for each document in the collection by computing an inner product of "term relevance" values for all terms considered:

`Probability(relevance|document) = SUM [term_relevance(i) * Term(i)]`

where:

$$\text{term_relevance}(i) = [r / (R - r)] / [(n - r) / (N - n - R + r)]$$

and:

N = number of documents

n = number of documents with term i

R = number of relevant documents

r = number of relevant documents with term i

The Bayesian decision rule can be used to decide whether a document has a high enough probability estimate to be chosen as relevant:

$$\text{Probability}(\text{relevance}|\text{document}) > \text{Probability}(\text{non-relevance}|\text{document})$$

1.2 *Information retrieval in SMART*

The SMART information retrieval system (Salton and McGill, 1983) has options to use the probabilistic model to rank documents that are retrieved as part of a feedback process. Initial retrieval is usually accomplished after computing a cosine similarity (Salton and McGill, 1983) between the terms in a query and the terms in documents. Those documents with the highest similarity have the lowest ranks (also the highest probability of being relevant). The user is presented with a list of the "top ranked" documents. The user can then decide which of the "top ranked" documents are relevant. SMART then can perform a vector feedback search (if desired) by adding any new terms from the relevant documents to the initial query and subtracting those terms from the initial query, which only appeared in the documents that were judged to be nonrelevant. The resultant

feedback query is then used to provide a new ranked list of documents (Salton and McGill, 1983). Alternatively, a probabilistic feedback can be performed.

Fox (1983a) has modified the application of the vector and probabilistic models to utilize additional information. This consists of author, date of publication, bibliographic coupling, bibliographic links, Computing Reviews' categories, and cocitations. SMART has been modified to add the additional information as subvectors (Fox, 1983a) to the document-term vectors of the original vector/probabilistic system. As modified, the collection then consists of extended vectors where the information is separated into subvectors:

Document_identification_number,

**Terms, authors, date of publication, bibliographic coupling,
bibliographic links, Computing Reviews' categories, cocitations**

1.3 Research goals

The goals of this study were to statistically examine the usefulness of the subvectors associated with different concept types (Fox, 1983a) and to determine if coefficients obtained via multiple regressions could be used to further enhance SMART's retrieval using the extended vectors. A less significant goal was to determine if knowledge of relevance as a ranking (from least (1) to most (4) relevant) would facilitate prediction, thus yielding better coefficients. To accomplish these objectives two different research collections were analyzed. These had been loaded into a version of SMART, which has been installed on a VAX-11/785 running UNIX (Fox, 1983b).

The collections, which have been described elsewhere (Fox, 1983b), consist of 3,204 abstracts of documents which appeared (1958 - 1979) in Communications of the Association for Computing Machinery (CACM Collection) and 1,460 abstracts of documents from various sources (1969 - 1977) concerning information science, along with citation data obtained from the Institute for Scientific Information (ISI Collection). The CACM collection has information necessary for all of the above mentioned extended vectors, but the ISI collection has only two additional information components (subvectors), author and cocitations. For both collections, a set of queries with known document-query relevances was used to generate data for the regression studies. Precision (the ratio of relevant documents retrieved to all documents retrieved for that query) averages were the principal measure used for determining the effectiveness of all retrieval runs with SMART.

2.0 Methods

2.1 Division of the collections

Both the CACM and ISI collections were divided into two approximately equally sized sub-collections by randomly selecting documents. The collections were split so that all analyses could be performed on one half of the data and the results obtained could be tested on the other half of the data. The dividing procedure used kept the same proportion of relevant to nonrelevant documents with regard to sets of queries for each collection. In the description and discussion that follow these sub collections are referred to as the CACM1 or CACM2 and ISI1 or ISI2 Collections.

2.2 Description of the collections and vectors

SMART was used to prepare data sets, for both the CACM1 and ISI1 collections, which consisted of query identification number (QID), document identification number (DID), rank in the prob-

abilistic retrieval, and a similarity measure (SIM). However, as it was based on the value of the similarity, rank was not used in this study. A separate data set was obtained for each concept type, which had the appropriate measure of similarity for all QID - DID pairings.

For use in tables and figures the concept types will be represented by the following abbreviations:

AUT	Authors
CRC	Computing Reviews' Category
DTE	Date of Publication
TRM	Terms
BBC	Bibliographic Coupling
LNK	Bibliographic Links
COC	Cocitations

For the CACM1 collection, there were 7 equal length data sets, one for each concept type, and a data set which listed the relevance judgment (ranked from 0 to 4) for each QID - DID pairing. The Statistical Analysis System (SAS, 1985) was used to merge the 7 concept type data sets with the relevance judgment data set by matching QID and DID. This gave a data set, which has QID, DID, and 7 different similarity measures (independent variables) and relevance judgment (dependent variable) for subsequent analysis.

For the ISI1 collection, only three concept types were available. Thus, only three concept type data sets with their appropriate similarity measure were obtained from SMART. Also, only binary relevance judgments were available for the ISI1 collection. SAS was used to merge the three concept type data sets with the relevance data set by matching a query and document. Thus, a data set with

three similarity measures (independent variables) and relevance judgment was constructed. Table 1 on page 8 shows the nature of the matrix that resulted from the merger of the 7 concept type data sets and the relevance data set.

2.3 Descriptive analysis of the data

SAS, the Statistical Analysis System, Version 5 (1985) was used to produce all statistical and graphical results. For all statistical tests of significance a threshold of .05 was used.

Procedures MEANS and UNIVARIATE were used to obtain descriptive statistics for all concept types in the two collections. As the distributions of values for the concept types were quite variable, they were transformed by taking the natural log of the similarity measure plus one. The log transformation was chosen, because the data were highly positively skewed due to the large values that occur in inner product calculations, when there is a good match between a query and document. Log and square root transformations are good choices for positively skewed data, but the log transformation is more effective at bringing the large values closer to the mean. This made the distributions much more symmetrical. The variables for the two collections were summarized in tables and some representative histograms and can be found in section "Description of the data" on page 11.

2.4 Obtaining regression coefficients for use in SMART

Procedure General Linear Model (GLM) was used for most regressions with models specified so that regressions were run without an intercept being calculated. The intercept would have been

Table 1. Organization of the CACMI merged data set.

Q I D	D I D	R N K R E L	PROBABILISTIC "SIMILARITY" FOR EACH CONCEPT TYPE						
			A U T	C R C	D T E	T R M	B B C	L N K	C O C
2	98	0	0.000	0.000	0.0000	0.13	0.000	20.816	0.000
2	655	0	0.000	0.000	0.0000	38.28	0.000	0.000	0.000
2	1138	0	0.000	3.620	0.0000	7.78	87.531	0.000	0.000
2	1179	0	0.000	0.000	0.0000	8.35	73.654	27.754	117.956
2	1314	0	0.000	3.620	0.0000	4.09	6.939	0.000	0.000
2	1426	0	0.000	4.895	0.0000	9.61	0.000	0.000	0.000
2	1429	4	428.254	4.730	20.8158	228.92	143.040	97.140	117.956
2	1435	0	0.000	13.877	0.0000	37.15	0.000	0.000	0.000
2	1541	4	484.816	185.923	27.7544	143.27	0.000	84.316	0.000
3	486	0	0.000	0.000	0.0000	2.41	0.000	0.000	41.250
3	507	0	0.000	0.000	0.0000	63.35	0.000	0.000	0.000
3	561	4	252.947	0.000	10.4079	282.14	87.228	166.526	57.903
3	816	0	0.000	0.000	0.0000	60.52	0.000	0.000	0.000

NOTE: QID = query identification number
 DID = document identification number
 RNKREL = ranked relevance
 Other variables are as previously described.

useless in subsequent runs, which tested the coefficients obtained from regressions. For the CACM collection, two, four, and all variable models were run on raw and log transformed data. In the ISI collection studies, two and all variable models were used. An example of one of the regression equations (two variable model) is as follows:

$$\text{Relevance} = a_1 \times (\text{Similarity_TRM}) + a_2 \times (\text{Similarity_LNK}) + \text{Error}$$

Where a_1 and a_2 are regression coefficients for terms and bibliographic links respectively.

In addition to the linear regressions performed with GLM, logistic regression (Procedure LOGIST) and all possible regressions (Procedure REG) were used on a preliminary data set.

2.5 Testing the usefulness of the coefficients as weights in SMART

Base runs were made in SMART for both halves of both the CACM and ISI collections. These runs used combinations of concept types that corresponded to the two and three variable models for the ISI1 Collections and two, four, and seven variable models for the CACM1 Collection. As no coefficients were used in these runs, they provided equal weights for all concept types for comparison with runs having coefficients. The coefficients obtained in the regression runs were then used as weights in feedback runs to try to improve precision in SMART. The runs using coefficients were compared against base runs, which had no coefficients but did use the same concept types. The coefficients that were developed for CACM1 and ISI1 were also tested on CACM2 and ISI2.

2.6 *Threshold techniques*

Histograms and graphs of the data were examined for possible threshold values for the concept types. Threshold values could be useful, if it could be shown that a value as high or higher than a certain percentile for a concept type gave a high probability that the document was relevant. Accordingly, the 60th, 75th and 90th percentiles were used as thresholds to test their usefulness. Variables were created for each concept type at each of the above percentiles. The corresponding threshold variables were given a value of 1.0 when the concept type value exceed the threshold and 0.0 otherwise. The regressions were then rerun with the additional variables for each threshold. These results of these runs are reported in "Use of threshold techniques" on page 33.

3.0 Results

3.1 *Description of the data*

Descriptive statistics for the CACM1 collection are given in Table 2 on page 13. Vectors for the seven concept types show considerable variation, as they range from zero to four digit numbers. All show relatively high values for CV (the coefficient of variation, from 235.0 to 674.4) and all are highly positively skewed (from 4.22 to 17.3). This is due to the high values that resulted from the inner product calculations when concept types in the query obtained good match with a document. Kurtosis measures were all large (from 8.8 to 381.5) due to peakedness caused by the high proportion of low or zero values in most concept types that were produced when there was a poor match.

As seen in Table 3 on page 14 the variables in the ISI1 collection were similar. The numbers range from 0 to 1188.1 and had high skewness (from 3.4 to 5.8) and high kurtosis (from 17.1 to 45.4). Their CV's were somewhat lower (163.8 to 356.7). Although linear regression and the F test are robust, these data are rather variable and quite different from the kind of examples usually shown in textbooks on regression analysis. Some of the variation is due to the extremes in values, but much of the variability is due to the sparseness of the QID - DID array. Evidence of the sparseness

is seen in Table 2 on page 13 and Table 3 on page 14 regarding length, the column that gives the number of non-zero values for each vector. In fact, five of the seven concept types for CACM1 have non-zero values in from 8.8% (AUT) to 17.9% (BBC) of the data records. In the ISI1 collection one of the three subvectors has only 13% of its values that are non-zero (AUT). An attempt to compensate for the sparseness is discussed under "Linear regressions on the CACM1 collection" on page 19. To try to minimize the effect of the high variability in the data, a natural log transformation was used on all variables and is reported in Table 4 on page 15 and Table 5 on page 16 for the CACM1 and ISI1 collections respectively. As can be seen in these two tables, the transformation does make the distributions considerably more symmetrical and reduces extremes among the various measures.

In the CACM1 collection, skewness (0 is normal) was reduced to more acceptable levels (-.009 to 2.53), kurtosis (0 is normal) was reduced similarly (-.44 to 10.2) and CV (100 is normal) was tighter (52.3 to 265.0). In the ISI1 collection, skewness dropped (-.37 to 2.6), kurtosis declined (-1.4 to 5.1), and CV was lowered (39.5 to 270.2). Thus, in both collections, the distributions of the concept type variables are more evenly matched and closer to normal than for the raw data.

Histograms of raw data and log data further illustrate the effects of the log transformation on the data. Histograms in Figure 1 on page 17 show the impact of the log transformation on TRM in the ISI1 collection (sample data). The large value of 4.8 for skewness in the raw data is shown by the long positive tail. However, the log transformed data show no tail in either direction, as skewness has been reduced to -.154. Not all of the histograms show such dramatic improvement as those seen in Figure 1 on page 17, but all of the concept types do show improved distributions. Figure 2 on page 18, which is of raw and log transformed CRC from the CACM1 collection (all records) is an example of a variable with only modest improvement.

Table 2. Descriptive statistics and vector length of raw CACMI data.

VARIABLE	MEAN	C.V.	MINIMUM VALUE	MAXIMUM VALUE
AUT	8.38	601.6	0	1152.3
CRC	4.24	668.1	0	187.9
DIE	3.41	674.4	0	87.8
TRM	47.69	235.0	0	2389.0
BBC	10.89	313.9	0	433.2
LNK	11.62	486.0	0	1574.3
COC	10.76	603.7	0	1602.0

VARIABLE	SKEWNESS	KURTOSIS	NUMBER NON-ZERO
AUT	14.0	268.3	314
CRC	7.3	78.9	1603
DIE	4.2	24.3	725
TRM	7.6	90.5	3899
BBC	4.7	30.2	819
LNK	12.1	226.3	616
COC	17.3	381.4	566

NOTE: There were 4035 records in this set.

Table 3. Descriptive statistics and vector length of raw ISI1 data

VARIABLE	MEAN	C.V.	MINIMUM VALUE	MAXIMUM VALUE
TRM	45.7	176.6	0	1188.1
AUT	3.8	356.6	0	200.3
COC	33.7	163.7	0	727.5

VARIABLE	SKEWNESS	KURTOSIS	NUMBER NON-ZERO
TRM	5.8	45.4	5449
AUT	5.2	36.4	710
COC	3.4	17.1	3400

NOTE: There were 5456 records in this set.

Table 4. Descriptive statistics of log CACMI data

VARIABLE	MEAN	C.V.	MINIMUM VALUE	MAXIMUM VALUE
AUT	0.33	351.8	0	7.1
CRC	0.81	138.4	0	5.2
DTE	0.50	218.4	0	4.4
TRM	2.83	52.3	0	7.7
BBC	0.68	214.5	0	6.1
LNK	0.53	257.1	0	7.3
COC	0.50	265.0	0	7.3

VARIABLE	SKEWNESS	KURTOSIS
AUT	3.39	10.20
CRC	1.07	0.05
DTE	1.84	1.75
TRM	-0.00	-0.44
BBC	1.95	2.33
LNK	2.49	4.90
COC	2.53	5.16

NOTE: Number of non-zero values and number of records are given in Table 2 on page 13

Table 5. Descriptive statistics of log ISI1 data

VARIABLE	MEAN	C.V.	MINIMUM VALUE	MAXIMUM VALUE
TRM	3.19	35.2	0	7.1
AUT	0.40	268.5	0	5.3
COC	2.34	74.2	0	6.6

VARIABLE	SKEWNESS	KURTOSIS
TRM	-0.12	0.48
AUT	2.51	4.85
COC	-0.03	-1.30

NOTE: Number of non-zero values and number of records are given in Table 3 on page 14.

3.2 *Linear regressions on the CACM1 collection*

SAS Procedure GLM was used to run full models of all concept types as independent variables and ranked relevance judgment as the dependent variable. These runs were performed with the raw data and log transformed data and are summarized in Table 6 on page 22. However, the proportion of nonrelevant to relevant documents was too high, more than 9 to 1. This problem of unbalanced groups was partly responsible for the low coefficient of determination (RSQ) for the raw data (.387) and for the log data (.396). In order to improve the proportion of relevant versus nonrelevant records, most of the nonrelevant documents were randomly discarded leaving a data set that had equal proportions of relevant versus nonrelevant records (766 total records). This modestly improved the RSQ of the raw data to .445 and considerably improved the RSQ of the log transformed data to .627, as can also be seen in Table 6 on page 22. The sparseness of many of the concept type subvectors is also probably contributing to the relatively low RSQ (see "Description of the data" on page 11), but nothing could be done about that.

Similar runs were made using binary relevance as the dependent variable. The results of these regressions are given in Table 7 on page 23. The same kind of improvement in RSQ that was seen in Table 6 on page 22 was found by discarding most of the nonrelevant records and balancing the relative number of relevant versus nonrelevant documents. However, the improvement between raw data and log data is a little greater, by approximately 1% to 4%. Furthermore, the binary relevance data with the log transformed independent variables gave a better RSQ than the ranked relevance (.6659 versus .6274).

A plot of predicted scores versus residuals for the best log sample model with binary relevance data is displayed in Figure 3 on page 24 and shows a fair degree of closeness for relevant documents (1.0) and considerable spread for nonrelevant documents. A similar plot for ranked relevance data is

shown in Figure 4 on page 25, but here the relevant values are divided into values from 1.0 to 4.0. Again, the relevant documents show less spread than the nonrelevant.

The concept type variables for each regression run were ranked by their Type III Sum of Squares (SAS, 1985), which gives the sum of squares for each variable independently of its order in the regression model. From the rankings, some two and four variable models were chosen and run using the same two dependent variables for all records and for the sample set of records. The coefficients, RSQ's, and rankings are also provided in Table 7 on page 23 and Table 6 on page 22 for binary and ranked relevance data respectively.

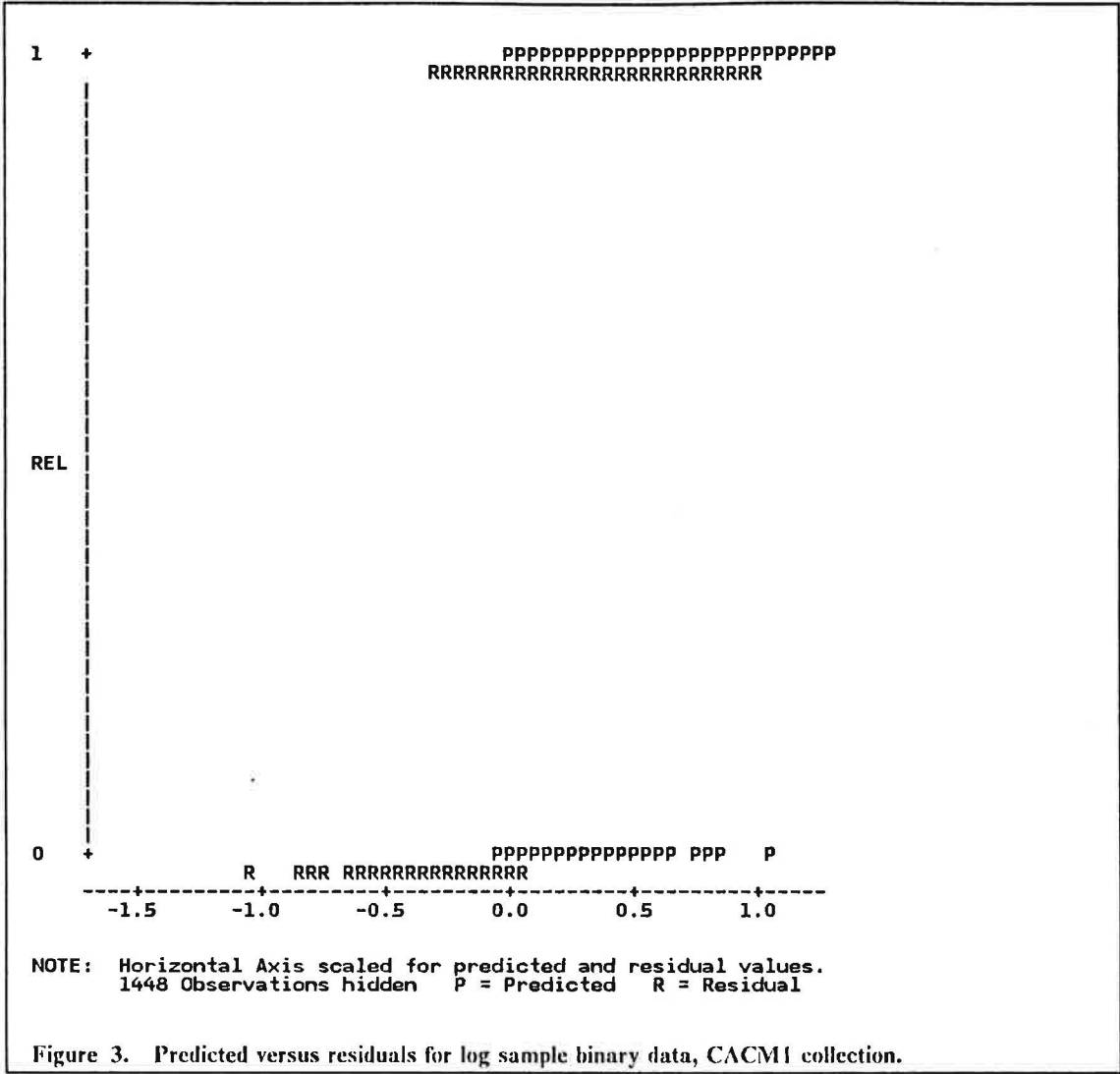
The two variable model (TRM and LNK) using raw data and all records gave 86% of the RSQ of the seven variable model for both dependent variables. For sample raw data, 83% and 86% of the seven variable RSQ was obtained. For log transformed data, all record regressions with the same independent and dependent variables gave 79% and 81% of the original seven variable RSQ (ranked and binary relevances respectively). However, the sample of log data gave 98% of the seven variable RSQ for both ranked and binary relevance data. In fact the two variable model for log transformed independent variables and binary relevance data gave a higher RSQ than any of the other seven variable models. The four variable model (AUT, CRC, TRM, and LNK) gave modest improvement, but clearly most of the variance is accounted for by TRM and LNK.

All possible two-way interactions were tested (using proc GLM in SAS) on ranked relevance and binary relevance data. Several were found to be significant at the .05 level. Table 8 on page 26 shows interactions for binary and ranked relevances for log sample data, the best models. For example, there is a significant interaction between AUT and TRM using the reduced sample log data for ranked relevance. This makes sense as authors may write more than one article in the same subject area using common terms and few articles in other areas with different terms. The amount of variance explained by this interaction was relatively small, less than 3% of that of TRM. However, this interaction accounts for more variance than the three lowest ranking concept types, which did not add much predictive ability to the seven variable model (see Table 6 on page 22)

either. Other interactions seemed reasonable, but also did not account for much variance. In fact, all of the interactions together only raised RSQ from .6274 to .6432. Additionally, some interactions had negative coefficients, which indicated inverse relationships with the other variables. Furthermore, as SMART is not currently programmed to use interactions, they were not used on subsequent regression runs. A subset (based on the largest two-way interactions and the concept types) of the possible three-way interactions was tested with none being significant at the .05 level.

Table 6. Regression coefficients, ranks, and RSQ's for CACM ranked relevances.								
	Raw data				Log data			
Model variables	Rank	All data coefficients	Rank	Sample coefficients	Rank	All data coefficients	Rank	Sample coefficients
AUT	2	.0037*	3	.0027*	1	.2619*	4	.1300*
CRC	5	.0069*	4	.0120*	3	.0653*	3	.1645*
DTE	6	.0060*	6	.0096*	5	.0552*	5	-.0808
TRM	1	.0021*	1	.0022*	4	.0262*	1	.2724*
BBC	4	.0027*	2	.0070*	7	.0014	7	.0120
LNK	3	.0031*	5	.0025*	2	.1788*	2	.1893*
COC	7	.0006	7	.0009	6	.0135	6	.0176
Model RSQ		.3869		.4452		.3959		.6274
TRM	1	.0033*	1	.00415*	1	.0630*	1	.3287*
LNK	2	.0050*	2	.0049*	2	.3063*	2	.2719*
Model RSQ		.3362		.3853		.3120		.6152
AUT	3	.0039*	3	.0030*	1	.2740*	4	.1003*
CRC	4	.0077*	4	.0147*	4	.0646*	3	.1633*
TRM	1	.0074*	1	.0029*	3	.0327*	1	.2699*
LNK	2	.0042*	2	.0041*	2	.1403*	2	.1938*
Model RSQ		.3753		.4192		.3922		.6260
NOTE: * = significant at the .05 level								

	Raw data				Log data			
Model variables	Rank	All data coefficients	Rank	Sample coefficients	Rank	All data coefficients	Rank	Sample coefficients
AUT	2	.00093*	5	.00061*	1	.0740*	5	.0193
CRC	4	.00270*	3	.00509	3	.0228*	3	.0604*
DTE	6	.00192*	6	.00300	5	.0167*	4	-.0249
TRM	1	.00060*	2	.00059*	4	.0084*	1	.0876*
BBC	5	.00098*	1	.00243*	7	-.00066*	7	-.00167*
LNK	3	.00094*	4	.00074*	2	.05729*	2	.0667*
COC	7	.00021*	7	.00029	6	.00655	6	.0111
Model RSQ		.3721		.4386		.4064		.6659
TRM	1	.00096*	1	.00121*	2	.01997*	1	.10446*
LNK	2	.00151*	2	.00147*	1	.09478*	2	.08194*
Model RSQ		.3212		.3678		.3294		.6538
AUT	3	.00099*	4	.00071*	1	.0777*	4	.01007
CRC	4	.00312*	3	.00596*	4	.0222*	3	.05829*
TRM	1	.0070*	1	.00084*	3	.01047*	1	.08679*
LNK	2	.00128*	2	.00128*	2	.06138*	2	.06903*
Model RSQ		.3577		.4056		.4024		.6640
NOTE: * = significant at the .05 level								



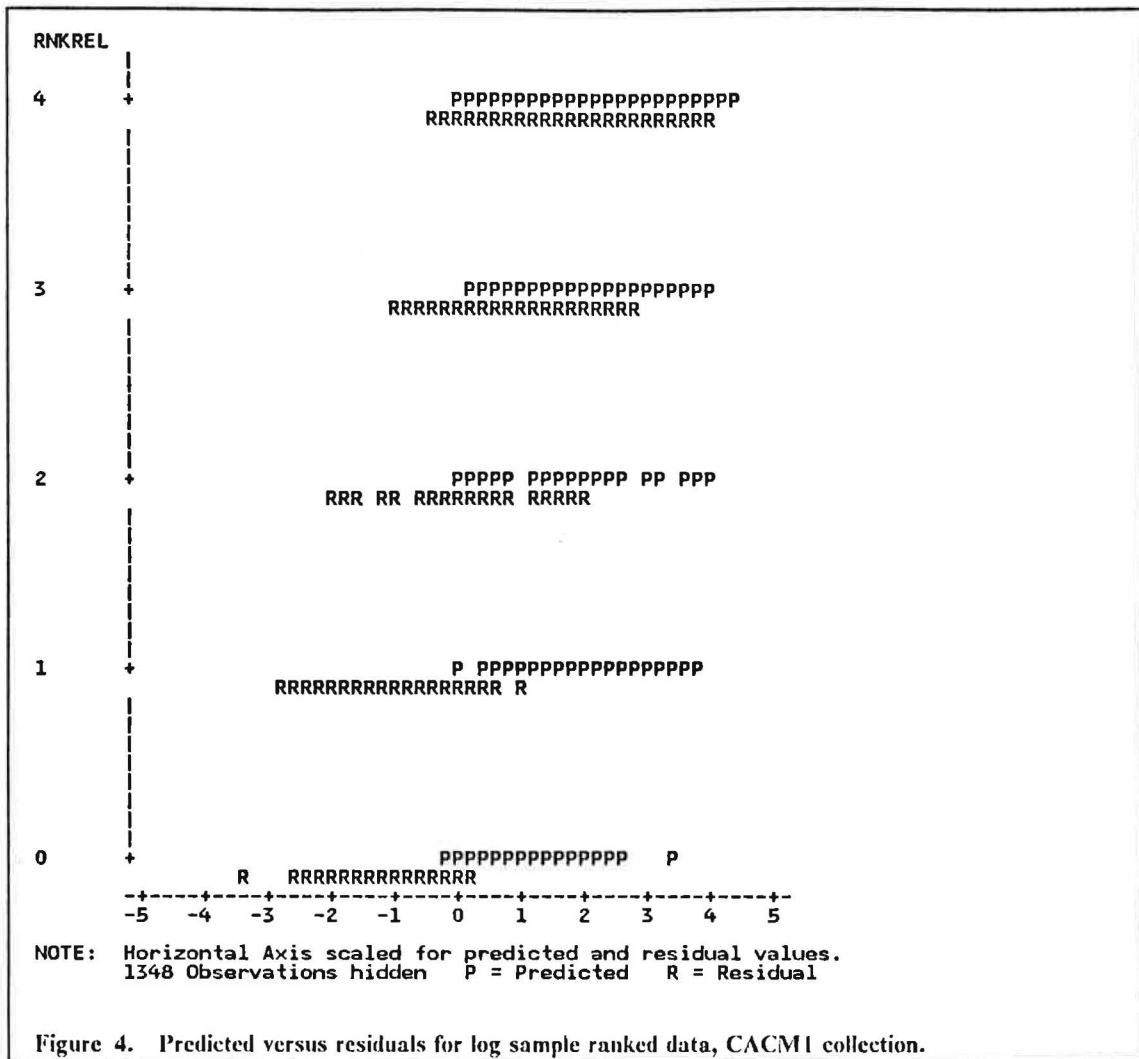


Table 8. Sums of squares and probability values for CACMI two-way interactions.

RANKED RELEVANCE DATA

SOURCE	SUM OF SQS	F VALUE	PR > F
AUT	1733.250	852.55	0.0001
CRC	69.646	231.01	0.0001
DTE	8.562	4.21	0.0405
TRM	378.762	186.30	0.0001
BBC	8.829	4.34	0.0375
LNK	53.929	26.53	0.0001
COC	0.621	0.31	0.5806
AUT*CRC	16.005	7.87	0.0051
AUT*TRM	10.277	5.06	0.0248
AUT*BBC	10.966	5.39	0.0205

BINARY RELEVANCE DATA

SOURCE	SUM OF SQS	F VALUE	R > F
AUT	156.031	942.08	0.0001
CRC	3.271	321.64	0.0001
DTE	1.081	6.53	0.0108
TRM	39.924	241.05	0.0001
BBC	0.649	3.92	0.0481
LNK	7.164	43.25	0.0001
COC	0.247	1.49	0.2220
AUT*CRC	2.318	14.00	0.0002

3.3 *Linear regressions on the ISI1 collection*

Regressions on the ISI1 collection followed the same general pattern as those for the CACM1 collection, but there were only three concept types (TRM, AUT, and COC) and only binary relevance data. Full models on raw data gave (see Table 9 on page 28) RSQ's of .2356 and .2812 for all records and a sample respectively. Full models of log transformed data gave RSQ's of .3162 (all records) and .481 (sample). A plot of predicted scores versus residuals for the best log sample model is displayed in Figure 5 on page 29 and shows a fair degree of closeness for relevant documents (1.0) and considerable spread for nonrelevant documents. As shown in Table 9 on page 28, the variables TRM and COC were ranked one and two by Type III Sums of Squares. Thus, they were used in two model runs which had RSQ's of 96% to 99.5% of the full model RSQ's.

All possible two-way interactions were done and all were significant at the .05 level, but they accounted for little variance (see Table 10 on page 30, which has interactions for the best model, log sample data). For example, the regression for log sample improved from .481 to .4978. However, the coefficient for AUT became negative and one of the interactions (TRM with LNK) was also negative. As for the CACM1 collection, coefficients for the interaction terms were not used with SMART.

Table 9. Regression coefficients, ranks, and RSQ's for ISI data.								
	Raw data				Log data			
Model variables	Rank	All data coefficients	Rank	Sample coefficients	Rank	All data coefficients	Rank	Sample coefficients
TRM	1	.0017*	2	.0017*	1	.0614*	1	.1092*
AUT	3	.0035*	3	.002834*	2	.0608*	3	.0281*
COC	2	.0018*	1	.00283*	3	.0256*	2	.0373*
Model RSQ		.2812		.2812		.3162		.4810
TRM	1	.0020*	1	.0020*	1	.0667*	1	.1128*
COC	2	.019*	2	.0029*	2	.0314*	2	.0400*
Model RSQ		.2300		.2784		.3036		.4790
NOTE: * = significant at the .05 level								

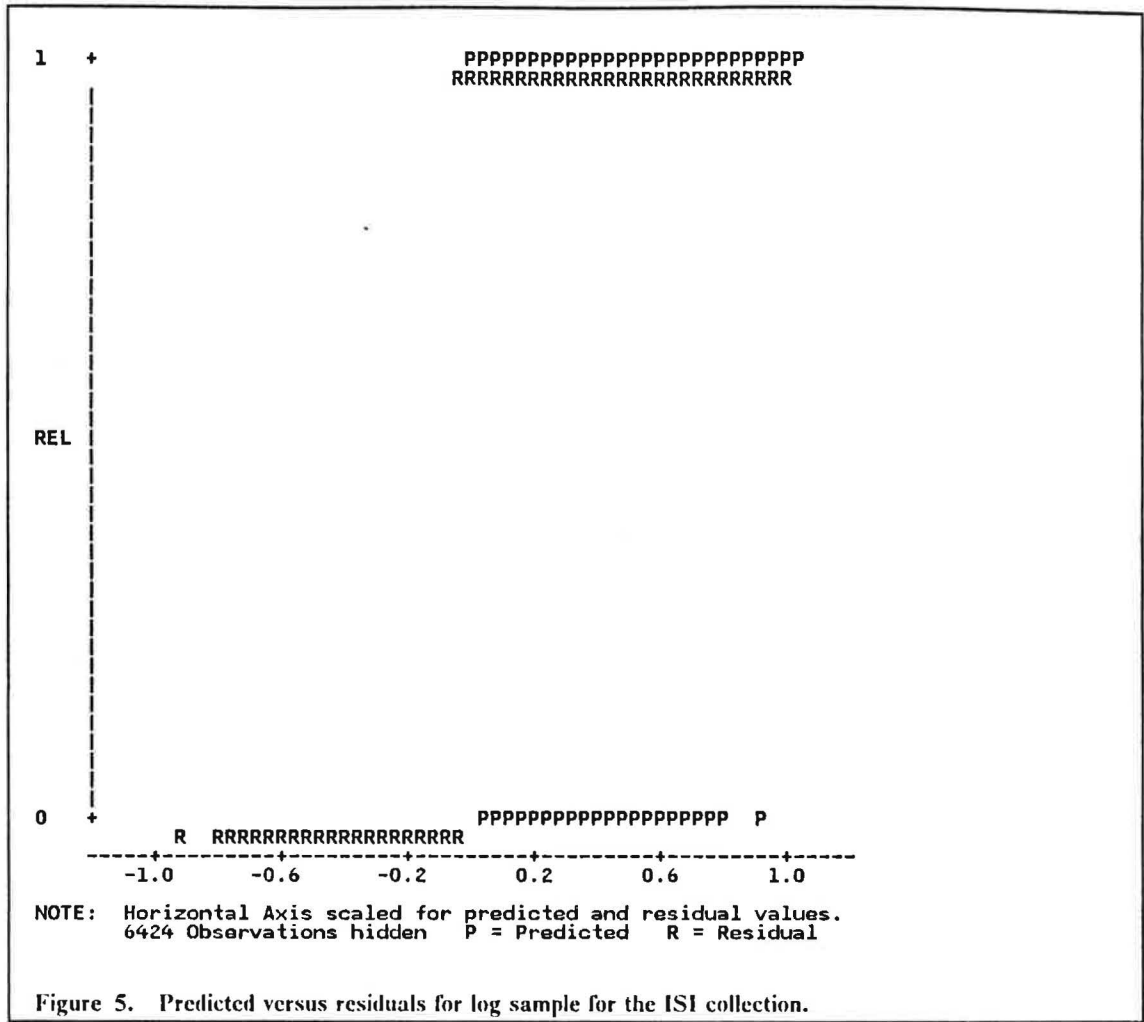


Table 10. Sums of squares and probability values for IS11 (log sample data) interactions.

SOURCE	SUM of SQS	F VALUE	PR > F
TRM	777.610	3126.57	0.0001
AUT	4.869	19.58	0.0001
COC	11.332	45.57	0.0001
TRM*AUT	7.240	29.11	0.0001
TRM*COC	9.867	39.67	0.0001
AUT*COC	2.774	11.16	0.0008

3.4 Additional regression techniques

Other regression techniques, including all possible regressions (using procedure REG in SAS) and logistic regression (using procedure LOGIST in SAS) were tried. In these runs different coefficients were obtained, but the same four predictor variables were most important in the models (TRM, AUT, LNK, and CRC). They did not give improved RSQ, but rather were much worse with RSQ's of approximately half that of the models described earlier. Additionally, the logistic regression program gave an intercept, which was not desirable for use in SMART.

3.5 Use of regression coefficients in SMART for CACM1 and ISI1 collections

Table 11 on page 34, Table 12 on page 35, and Table 13 on page 36 show the average precision values for base runs and precision values for the coefficient runs for CACM1 and ISI1 collections. Base runs are retrievals, which used the concept type values, but which give equal weights to the various concept types. In the coefficient runs the various concept type values are weighted by the regression coefficients.

Some minor improvement was found in most runs. For example, in the ISI1 collection, precision with coefficients from the sample log regressions showed a 2.9% improvement for the two concept type run and a .7% improvement for the three concept type run. Also for ISI1 runs, both raw data and log data in the two and three concept type runs showed very small improvement (.4% - 1.1%), while coefficients from a sample of raw data were worse (.4% - 1.2%). The best precision was obtained using coefficients developed from log data with all records included.

In the CACM1 runs, coefficients from log data and all records with binary relevance gave the best improvement over the base runs (1% for the seven concept type model to 5% for the two concept type model), but the matching sample log runs were all worse (3.3% - 9.4%). In general, all of runs using all data for both binary and ranked relevance measures in the CACM1 runs yielded improved precision values (1% - 5%) with the exception of binary relevance raw data all, which was worse by nearly 15% (see Table 7 on page 23).

3.6 Use of regression coefficients in SMART for CACM2 and ISI2 collections

Table 14 on page 37, Table 15 on page 38, and Table 16 on page 39 show the average precision values for base runs and precision values for the coefficient runs for CACM2 and ISI2 collections. The base runs in CACM2 and ISI2 of each collection are not as high as those of CACM1 and ISI1. That was expected, because the relevance weights were computed over a different document sample. In the CACM2 base runs, when the other concept types were added to terms, consistent improvement in average precision was found. However, in the ISI2 base runs, when the other concept types were added to terms, average precision was consistently lower (see Table 14 on page 37). For the runs with coefficients that were developed using the CACM1 and ISI1 collections, precision improved progressively as concept types were added to terms. However, all of the runs which used coefficients had lower precision values than the corresponding base run. The best CACM2 precision value for a run with coefficients used coefficients that had been obtained from all raw data with ranked relevance in CACM1 (.1987), while the best CACM2 base run was for log data with all concept types in the model (.2251). Though disappointing, these results are not surprising - term relevance weights and coefficients for concept types were derived through a feedback sampling process on a different half of the collection.

3.7 Use of threshold techniques

Histograms for all of the concept type values were examined for possible threshold values. A threshold would be a value for a concept type, which when reached or exceeded would give a very high probability that that document would be relevant, without regard to the values of the other concept types. The 60th, 75th and 90th percentiles for each concept type were tested as thresholds and variables were created for each concept type at each level. The corresponding threshold variables were given a value of 1.0 when the concept type value exceeded the threshold and 0.0 otherwise. Regressions were then rerun on the CACMI data with the additional variables for each threshold. With the additional variables RSQ's went down. For example, using all records, ranked relevance, and raw data at the 90th percentile threshold, RSQ dropped from .3869 to .3344. Other runs showed similar drops, but were not reported here.

Table 11. Precision values from base and coefficient runs for the IS11 collection.						
Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.3220	.3309				
TRM, COC	.3485	.3239	.3499	.3474	.3521	.3588
ALL	.3573	.3356	.3558	.3531	.3589	.3597

Table 12. Precision values from base and coefficient runs for the CACMI collection (binary relevance).

Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.4813	.4775				
TRM, LNK	.5714	.5584	.5806	.5753	.6006	.5466
AUT, CRC, TRM, LNK	.5855	.5965	.6182	.6121	.6309	.5663
ALL	.6315	.6066	.5383	.5975	.6384	.5770

Table 13. Precision values from base and coefficient runs for the CACM1 collection (ranked relevance).						
Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.4813	.4775				
TRM, LNK	.5714	.5584	.5783	.5746	.6007	.5510
AUT, CRC, TRM, LNK	.5855	.5965	.6127	.6146	.6285	.5720
ALL	.6315	.6066	.6298	.5962	.6340	.5779

Table 14. Precision values from base and coefficient runs for the IS12 collection.

Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.1444	.0984				
TRM, COC	.1249	.0854	.1256	.1190	.0966	.09833
ALL	.1265	.0904	.1257	.1190	.1035	.0999

Table 15. Precision values from base and coefficient runs for CACM2 (binary relevance).						
Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.1889	.1807				
TRM, LNK	.1736	.1799	.1548	.1687	.1237	.1801
AUT, CRC, TRM, LNK	.2125	.2037	.1927	.1770	.1721	.1936
ALL	.2185	.2251	.1502	.1547	.1725	.1961

Table 16. Precision values from base and coefficient runs for CACM2 (ranked relevance).

Model	Base Runs		Raw Data		Log Data	
	Raw	Log	All	Sample	All	Sample
TRM	.1889	.1807				
TRM, LNK	.1736	.1799	.1548	.1694	.1229	.1800
AUT, CRC, TRM, LNK	.2125	.2037	.1930	.1794	.1709	.1940
ALL	.2185	.2251	.1987	.1502	.1692	.1979

4.0 Discussion

4.1 Usefulness of concept types as shown by linear regressions

The first goal of this study was to statistically examine the usefulness of the concept types (extended vectors) in probabilistic retrieval. As had been shown by Fox (1983a), the extended vectors (base runs) appreciably improve feedback retrievals over those with just terms; in the CACM1 collection, as divided here, precision is increased from .4813 to .6315. Additional information (concept types) also improved precision for base runs in the ISI1 collection, but by more modest amounts (.3220 to .3556).

Descriptive statistics revealed that the concept type data values were highly variable and natural log transformations were used to partially alleviate problems due to the variability of the data. The descriptive analyses also pointed out potential problems due to sparseness of some of the vectors when they were compared to others. Although these data were not normally distributed, linear regression seemed robust enough to be of some use in their analysis. It was believed that consist-

ency and repeatability of results across different runs and different data sets were of more importance than meeting all of the assumptions of the linear regression technique. A high degree of consistency was found, as shown by Table 6 on page 22, Table 7 on page 23, and Table 9 on page 28. This is especially evident from the rank columns in these tables as the same concept types consistently achieved the highest rankings. Consistency was also shown in that the same highest ranking concept types were also consistently statistically significant. The same kind of consistency was also shown by the logistic regression and all possible regressions techniques that were also tried.

The regression analyses of this study were used to obtain measures of the importance of each concept type in predicting whether a document would be relevant or not. The relative importance of each concept type was summarized in Tables 6, 7, and 9 as their ranks for each run. The various models that were run also showed that most of the variance in relevance could be accounted for by only four of the vectors (AUT, CRC, LNK, and TRM) for the CACMI collection. In the ISI1 collection, two of the vectors (TRM and COC) accounted for most of the variance. The regression runs did provide numerous coefficients, which were used in subsequent feedback runs in SMART. Log transformed data and samples of the records gave the best RSQ's; .6654 was the highest RSQ (binary relevance).

4.2 Ranked relevances versus binary relevances

A lesser goal of this study was to test whether more knowledge of relevance would result in better prediction. It is surprising that having relevance ranked from 0 to 4 did not really improve the regression model over binary relevance alone. Perhaps relevant documents (ranked from 1 to 4) have more in common with each other than they have differences from nonrelevant (ranked 0).

4.3 Thresholds as aids in regression

The investigation of thresholds did not yield anything useful for these collections. It was thought that a high value for a single concept type would virtually assure that a document would be in the relevant group. However, no significant improvement was obtained in any of the threshold runs and some runs even showed poorer RSQ.

4.4 Improvement of retrieval using coefficients

The second major goal of this project was to test whether coefficients obtained by linear regression would prove useful as weights for extended vectors in probabilistic retrieval. The coefficients that were used in the feedback runs with SMART proved to be of limited usefulness here as improvements in precision were limited to the 1% to 5% range. Although log data and samples of the records gave the best RSQ's, coefficients from log values of all data improved precision the most.

4.5 Conclusions and implications for further research

The findings of this study support previous work of Fox (1983), which showed that additional information improves retrieval as measured by precision and illustrated in the base runs. Regression coefficients were of some usefulness, when used as subvector weights in improving precision. It was also found that terms, authors, bibliographic links and Computing Reviews' abstracts accounted for the most variance in predicting relevance. Log transforming the data values for the

concept types modestly helped both the regression analyses and the retrieval in SMART. Binary relevance seemed to be better than ranked relevance in this study.

The author of this study would like to suggest that further research might be pursued along two paths. The first path would be to obtain a larger collection of documents, of a more general nature. The purpose of this collection would be to try to better characterize the properties of the concept types and to try to develop coefficients that could be generalized to other collections. The second path would be to try to develop a simulated document collection to try to learn more about the capabilities and limitations of the probabilistic model.

References

- (Fox, 1983a) Fox, Edward A. Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. Cornell University Ph.D. Thesis, University Microfilms, Ann Arbor, Michigan, Aug 1983.
- (Fox, 1983b) Fox, Edward A. Some Considerations for Implementing the SMART Information Retrieval System Under UNIX. Technical Report 83-560, Cornell University Department of Computer Science, Ithaca, New York, July 1983.
- (Fox, 1983c) Fox, Edward A. Characterization of Two New Experimental collections in Computer and Information Science Containing Textual and Bibliographic Concepts. Technical Report 83-561, Cornell University Department of Computer Science, Ithaca, New York, September 1983.
- (Robertson and Sparck Jones, 1976) Robertson, S.E. and K. Sparck Jones. Relevance Weighting of Search Terms, Journal of the ASIS, Vol. 27, No. 3, May-June 1976, pp. 129-146.
- (Salton and McGill, 1983) Salton, Gerald and Michael J McGill. Introduction to Modern Information Retrieval, McGraw-Hill, inc. New York, New York, 1983.
- (SAS, 1985) SAS Institute, Inc., The SAS Users' Guide: Basics, 1985 Edition, SAS Institute, Inc., Cary, North Carolina, 1985.
- (SAS, 1985) SAS Institute, Inc., The SAS Users' Guide: Statistics, 1985 Edition, SAS Institute, Inc., Cary, North Carolina, 1985.
- (van Rijsbergen, 1981) van Rijsbergen, C. J. Information Retrieval: Second Edition, Butterworths, Boston, Mass., 1981.
- (Yu and Salton, 1976) Yu, C.T. and Gerald Salton, Precision Weighting- an Effective Automatic Indexing Method, Journal of the ACM, Vol. 23, No. 1, January 1976, pp. 76-88.

Vita

GARY L. NUNN

Date of birth:

October 2, 1947

Address:

35 Fieldale Drive

Radford, Virginia 24141

Education:

Ph.D.- In Zoology with a major concentration in Statistics,
Southern Illinois University at Carbondale,
1975-1978, requirements met Dec 1982

M.A.- In Ecology with a minor in Botany, Southern
Illinois University at Carbondale, 1975

B.A.- In Philosophy, Ohio University, 1969