Title
COMPARING VECTOR SPACE RETRIEVAL WITH THE RUBRIC EXPERT SYSTEM

Authors
Gey, F.
Chan, W.

Publication Date
1988-11-01

LBL-25837
Preprint

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA, BERKELEY

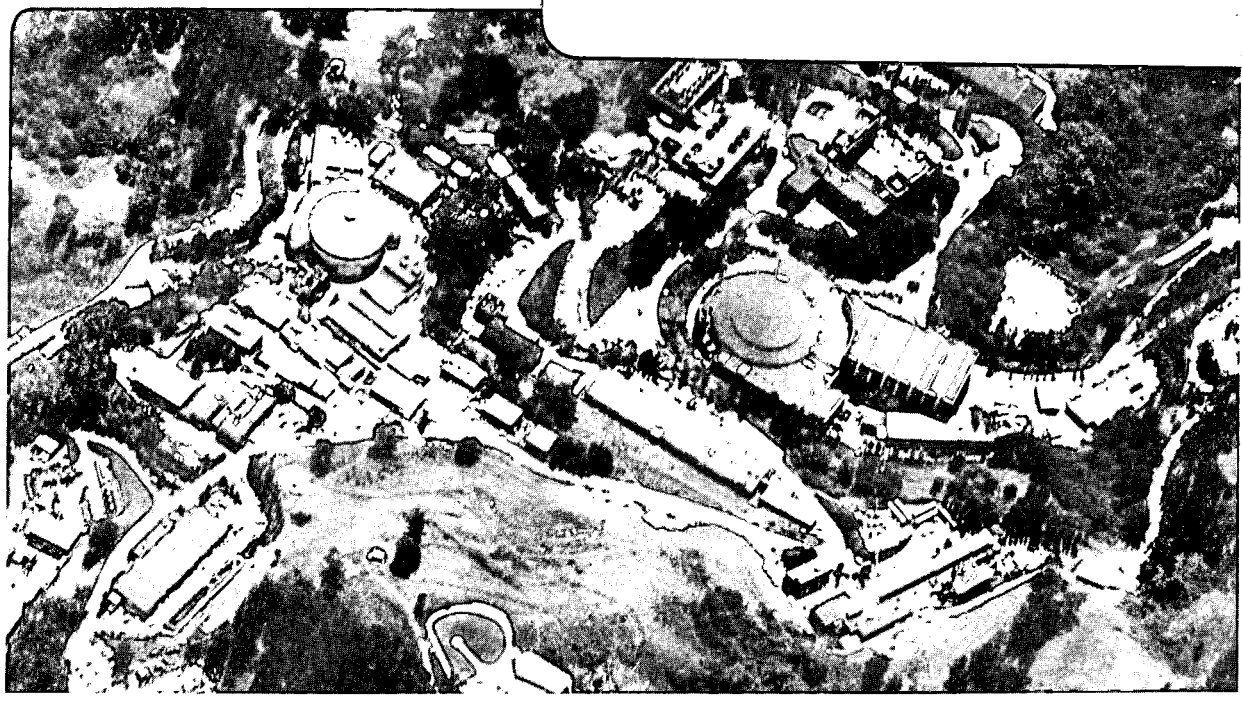
Information and Computing Sciences Division

Submitted to SIGIR Forum

Comparing Vector Space Retrieval with the RUBRIC Expert System

F. Gey and W. Chan

November 1988



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

LBL-25837

Comparing Vector Space Retrieval with the RUBRIC Expert System

Fredric Gey and Wingkei Chan

Information & Computing Sciences Division
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720

November 1988

This work was performed under the auspices of the Department of Energy under Contract DE-AC03-76SF00098. Views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official opinion or policy of the Department of Energy.

Comparing Vector Space Retrieval with the RUBRIC Expert System

by

Fredric Gey and Wingkei Chan*
Computer Science Research Department
Lawrence Berkeley Laboratory
Berkeley, CA 94720

ABSTRACT

RUBRIC is an expert system for full-text information retrieval. The underlying model for RUBRIC's information retrieval process is based upon fuzzy set theory. The RUBRIC developers have compared RUBRIC to the boolean retrieval model, which it subsumes. This study compares RUBRIC to the Vector Space Model for information retrieval, using RUBRIC's own test collection of thirty news articles from the Reuters News Service and their test search for articles which satisfy the information need to find out about "violent acts of terrorism." Results indicate that the vector space model is comparable to RUBRIC for relevant documents, while RUBRIC performs better at retrieving marginally relevant documents.

TABLE OF CONTENTS

This work was performed under the auspices of the Department of Energy under Contract No. De-AC03-76SF00098. Views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official opinion or policy of the Department of Energy.

*Present address: Undergraduate Computer Science program, University of Texas at Austin

Comparing Vector Space Retrieval with the RUBRIC Expert System

by

Fredric Gey and Wingkei Chan
Computer Science Research Department
Lawrence Berkeley Laboratory
Berkeley, CA 94720

1. Introduction

In a series of papers [ToAp 87, ToAp 86, ToAs 85, ToSh 85, ToAs 84, ToSh 83], Richard Tong and others at Advanced Decision Systems Inc. in Mountain View, California have described and developed an expert system for full-text information retrieval. The underlying model for the information retrieval process is based upon fuzzy set theory. The heart of the RUBRIC system is its search for text patterns in documents which form the evidence upon which a relevance ranking of documents to queries is obtained. Much of the system has been tested using a series of thirty news articles from the Reuters News Service. The query most applied to this document set is to search for articles which satisfy the information need to find out about "violent acts of terrorism."

As a summer undergraduate computer science project under the direction of the first author, the second author constructed a simple vector space model Information Retrieval (IR) system incorporating both document and query term weights, and returning a retrieval status value ranking of documents for either the cosine measure or the euclidean distance measure between document and query vectors. The system was tested using the RUBRIC query and document collection to compare the RUBRIC approach with the classical vector space approach to document retrieval.

2. The RUBRIC Conceptual Framework

The standard Information Retrieval framework attaches descriptive *terms* to documents, either with associated binary values (0 or 1 for presence or absence of terms) or associated weighted values (between 0 and 1, interpreted as the probability that users seeking a particular document would be searching under that term). For the RUBRIC framework, a finite set of *concepts* are related to documents to some degree specified by a real number r between 0 and 1 which is called the *relevance* of a particular document to a specific concept. RUBRIC **concepts** are built up in a hierarchical fashion from subconcepts, where the leaf nodes of the hierarchy are text patterns or operators on those patterns. Figure 1 gives such a hierarchy for the "violent acts of terrorism" query.

Leaf nodes of the request hierarchy are weighted combinations of text patterns or text adjacency patterns. RUBRIC scans the documents for these text pattern expressions, and uses a methodology ("calculus") to propagate weight values obtained at leaves up through the hierarchy to obtain a single retrieval status value (RSV) (or "relevance") for each document. RUBRIC then returns a ranked list of documents in descending order of

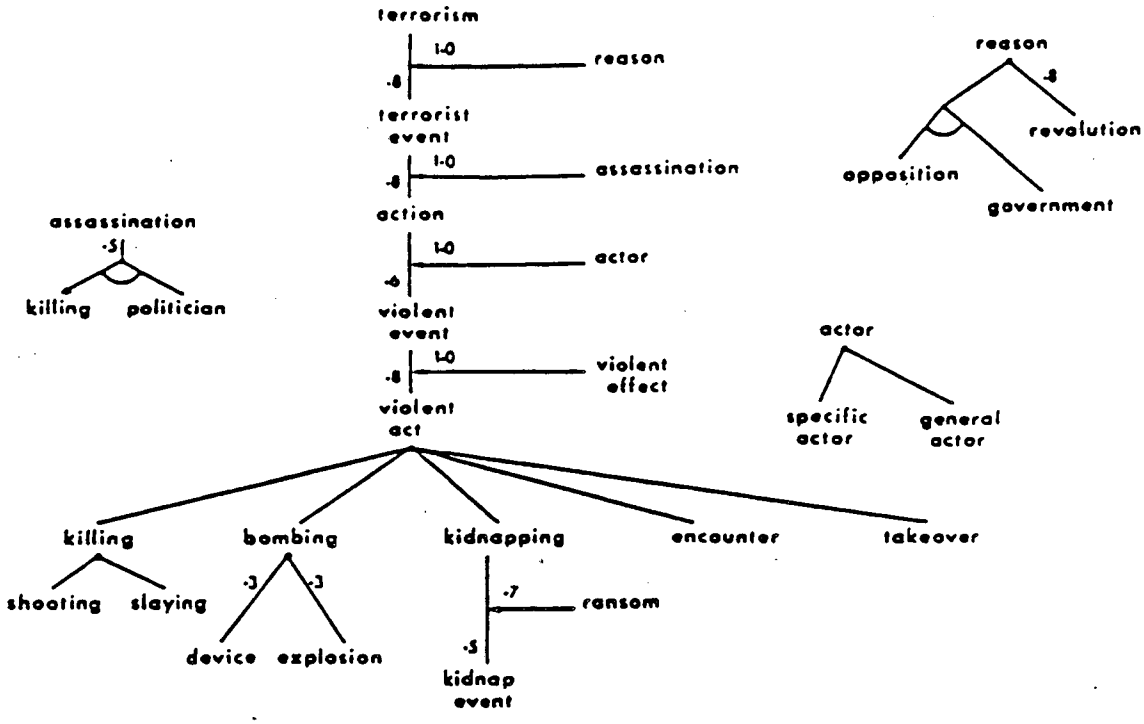


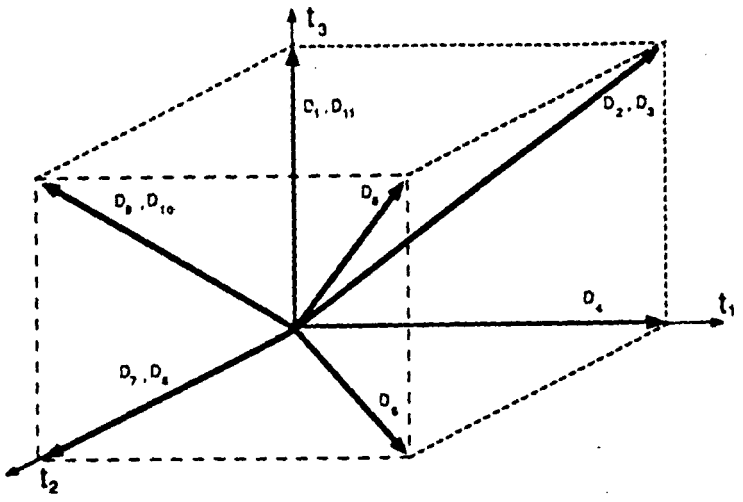
Figure 1: Violent acts of terrorism RUBRIC query tree [from ToAs84]

possible relevance.

3. Vector Space Model (VSM)

In the standard vector space model of information retrieval, terms are attached to documents, and the totality of distinct terms (the vocabulary) forms an ordered set for which the presence or absence of terms for a particular document characterizes that document as a vector in the term space of dimension **m** where m is the total number of distinct index terms in the index vocabulary. In this way the user's information need can be expressed as a *query vector* of terms in the same vector space. The IR system delivers to the user those documents which are "closest" to the query according to some measure of distance between the document and query vectors. This concept is illustrated with the following Figure 2 which shows 11 documents indexed by 3 terms. Using a dot product similarity measure, the weighted query at the bottom yields a retrieval status value for each document with respect to the query.

Certain independence and term orthogonality assumptions are implicit in this model; these have been analyzed extensively in [WoRa86, WoZi87] and will not be reviewed here.



	t_1	t_2	t_3	rsv=Q•D_i
D_1	0	0	1	3
D_2	1	0	1	4
D_3	1	0	1	4
D_4	1	0	0	1
D_5	1	1	1	6
D_6	1	1	0	3
D_7	0	1	0	2
D_8	0	1	0	2
D_9	0	1	1	5
D_{10}	0	1	1	5
D_{11}	0	0	1	3
Q	1	2	3	
	q_1	q_2	q_3	

Figure 2: Vector Space of 11 documents and 3 terms

Two similarity measures were chosen, the well-known *Cosine* measure which computes the cosine of the angle between the document and the query vectors:

$$cosine = \frac{\bar{Q} \bullet \bar{D}}{|Q|\,|\bar{D}|} = \frac{\sum_{i=1}^{m} q_i \times d_i}{\left(\sum_{i=1}^{m} q_i^2\right)^{1/2}\left(\sum_{i=1}^{m} d_i^2\right)^{1/2}}$$

and the *Euclidean distance* measure which computes the euclidean distance between the query and document vectors for each document:

$$Euclid = |\bar{Q} - \bar{D}| = \sqrt{\sum_{i=1}^{m}\left(q_i - d_i\right)^2}$$

4. Reuters Document Collection

Figure 3 on the following page lists the 30 Reuters news stories used in many RUBRIC experiments, and which were utilized for our vector space retrieval comparison with RUBRIC. The RUBRIC developers made a priori relevance judgments which divided the collection into three subsets (relevant, marginally relevant, irrelevant) with respect to the "violent acts of terrorism" information request.

	1	Overview story about the war in Chad and its effects
	2	Overview story of the situation in Poland and Solidarity
**	3	Short on car bomb in London
*	4	US deports Palestinian terrorist to Israel
*	5	FBI takes Reagan's Secret code card after assassination attempt
	6	Political effects of attack on Angola's oil refinery
*	7	Chilean secret service agent brought nerve gas into US
**	8	Follow-up to London bombing story
**	9	More on London bombing
	10	Boxing match - WBC Featherweight champion
	11	Earthquake in Pakistan
	12	Cyclone in India
	13	Soviet reaction to Polish crisis
	14	Reaction of Soviet bloc countries to Polish crisis
	15	Spanish army officers placed under house arrest
	16	Story on Iraq-Iran conflict
	17	Accidental chain of explosions at Army arms dump in Zimbabwe
	18	General interest story about Napolean and Waterloo
**	19	Bomb explosions in two Yugoslav restaurants
	20	Accidental explosion in apartment building in NE Italy
**	21	Italian couple freed by kidnappers after ransom paid
**	22	Bomb explosion in central Tehran street
**	23	Part story about murdered Italian industrialist
	24	Iranian leftists executed by firing squad in Tehran
	25	Shell exploded and killed bomb disposal experts in E. Beirut
**	26	Mayor and seven others kidnapped and shot in Guatemala
*	27	Lawyers for Sadat's assassins argue against charges
	28	Violence caused by Haitian refugees in Miami detention center
**	29	Iranian Parliament member assassinated in Tehran
	30	Brazilian athlete recovering from auto accident

** Relevant Stories
* Marginally-relevant Stories

Figure 3: REUTERS News stories

An examination of document number 21 shows certain significant words (highlighted) which might indicate a relationship to the information need "violent acts of terrorism.":

> *Reggio Calabria, Italy, Oct 12, Reuter* -- **Kidnappers** freed an Italian student and his fiancee today after their families paid a 430,000 dollars **ransom.** Carlo Speziale, 21, and Maria Antonietta Raschella, 20, were released during the night in Calabria's Aspromonte mountains after walking blindfolded for three hours across the rugged countryside, police said. The two students were snatched by gunmen on July 25 in an archeological park while vacationing in Calabria. About 30 people have been **kidnapped** in Italy this year.

5. Choosing the Right Vocabulary

In the classical vector space model, vectors are constructed for bibliographic reference document surrogates, which are indexed according to a (usually) fixed index vocabulary. To apply the VSM to full text, the central problem is choice of appropriate vocabulary which specifies the vector space. In boolean full-text systems, all words except common words on a "stop list" are used to form the vocabulary. Thus, the simplest way to select the vocabulary words is to have no selection at all, i.e. to put all the words in the documents into the vocabulary. This method is inefficient because the size -- the number of words in the vocabulary (the number of terms in a vector) -- equals the number of distinct words in all documents. Moreover, constructing vectors based on all words may yield inaccurate results due to the low resolving power of some words.

In our project, we wrote software to generate a frequency distribution of word frequencies in the document collection. It was our intention to utilize the results of H.P. Luhn on resolving power of significant words to identify the vector space vocabulary. Luhn's work [LUHN 58] showed the relationship of word resolving power and word frequency to be a bell-shaped curve (Figure 4); words with a medium word frequency have the highest resolving power.

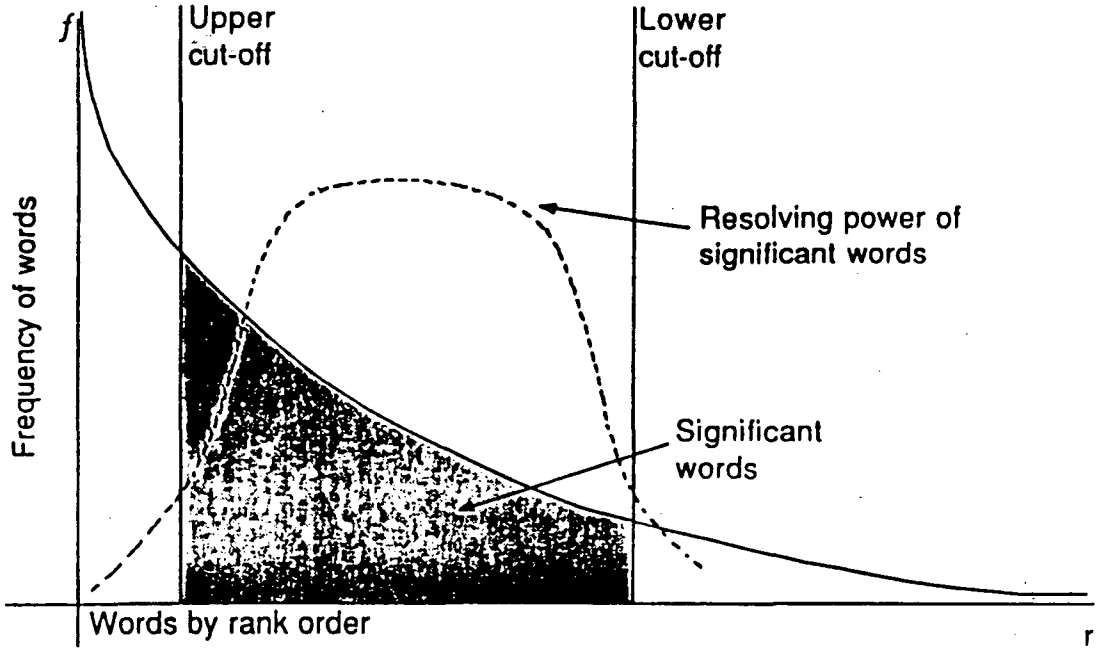


Figure 4: Luhn's Curve: Resolving Power of Significant Words (from [LUHN 58])

Our statistical word frequency results (see Figure 5) showed that 2.2% (56 words) of the 2573 distinct words accounted for more than 40% of the total word occurances in the Reuters document collection. Furthermore, 85.7% (2216 words) of the distinct words in the collection occurred less than four times in the entire collection. These word distribution results are similar to those encountered by Maron [MARO 61] in his classic experiments which used the computer for automatic indexing of a collection of 240 technical abstracts from the Proceedings of the IRE.

We chose to discard those words which occurred more than 17 times in the collection or less than 4 times in the collection. The remaining 310 words, which accounted for 24.87 percent of the total word occurrances in the 30 Reuters documents, became the term vocabulary for vector space representation. Moreover, we weighted each document-term pair by the frequency of occurrence of that word within the document. Nothing was done about the effect of different suffix endings of the same word stem. We will return to the matter of stemming below.

WORD	OCCURRENCE COUNT	PERCENT OF OCCURRENCE	CULMULATIVE PERCENT OF OCCURRENCE	CULMULATIVE PERCENT OF DISTINCT WORDS
THE	614	7.2235	7.2235	0.0389
OF	282	3.3176	10.5412	0.0777
IN	238	2.8000	13.3412	0.1166
AND	205	2.4118	15.7529	0.1555
.	.	.	.	.
REUTER	27	0.3176	34.0588	1.2048
.	.	.	.	.
GOVERNMENT	25	0.2941	34.9647	1.3214
PEOPLE	25	0.2941	35.2588	1.3603
TODAY	25	0.2941	35.5529	1.3991
.	.	.	.	.
THREE	18	0.2118	40.0706	2.1376
UP	18	0.2118	40.2824	2.1764
KILLED	17	0.2000	40.4824	2.2153
ARE	16	0.1882	40.6706	2.2542
GENERAL	16	0.1882	40.8588	2.2930
.	.	.	.	.
BOMB	10	0.1176	47.4118	4.0420
.	.	.	.	.
SHOT	9	0.1059	49.6000	4.8193
.	.	.	.	.
EXPLODED	7	0.0824	52.2824	5.9464
.	.	.	.	.
ATTACK	6	0.0706	55.0235	7.2678
.	.	.	.	.
BOMBING	6	0.0706	55.2353	7.3844
.	.	.	.	.
MURDER	5	0.0588	60.1176	10.2604
.	.	.	.	.
DEVASTATED	4	0.0471	62.1529	11.7769
.	.	.	.	.
EXPLOSION	4	0.0471	62.4353	12.0093
.	.	.	.	.
GUERRILLAS	4	0.0471	62.6706	12.2037
.	.	.	.	.
WHAT	4	0.0471	65.0704	14.1858
WITHOUT	4	0.0471	65.1174	14.2246
100	3	0.0353	65.1527	14.2635
.	.	.	.	.
HOPE	3	0.0353	67.8703	17.2561
HOPED	3	0.0353	67.9409	17.2950
.	.	.	.	.
KIDNAPPED	3	0.0353	68.5762	18.0334
.	.	.	.	.
KIDNAPPER	1	0.0118	90.8229	69.6852
.	.	.	.	.

Figure 5: Word distribution in Reuters collection

6. Results

Both the RUBRIC system and the vector space model return a ranking of documents in response to query. In our experiment, we evaluate the retrieval performance of both systems by comparing the ranking of the same collection of documents on the query "violent act of terrorism". From a cursory examination of the 310 word vocabulary, 10 words were chosen to represent the "violent acts of terrorism" query, and given 3 distinct weights between 1 and 10 as follows:

ATTACK 3
BOMB 7
BOMBING 7
DEVASTATED 3
EXPLODED 7
EXPLOSION 7
GUERRILLAS 5
KILLED 5
MURDER 5
SHOT 5

Running this query yields the ranking shown in Figure 6 on the following page.

6.1. Cosine Similarity Measure Results

Processing the query using the cosine measure, the resulted ranking is remarkably close to the RUBRIC system. The RUBRIC developers put the 30 Reuters documents into three categories -- relevant, marginally-relevant, and irrelevant. Out of the nine RUBRIC relevant documents, eight of them are ranked among the nine highest-ranked documents in our vector space model. Only one document, document number 21, which was considered as relevant in the RUBRIC system is given a low rank in our model.

Why did document 21 rank so low? The reason emerges if we examine the tail of the word distribution curve (refer back to Figure 5) Notice the words *kidnapped* and *kidnappers* appear 3 times and once respectively. These words do not appear in the vector space vocabulary and consequently were not put in the query. If **word stemming** were applied, "kidnap" would rise above the threshold and enter the vocabulary. RUBRIC uses word stemming as an integral part of the system. Indeed, if we artificially add the word "kidnapped" to the vocabulary and to the query with weight 5, the fate of document 21 changes dramatically, rising from rank 29 to rank 13, for the cosine measure.

Recent research at the National Library of Medicine has questioned the effectiveness of word stemming [HARM 87], but whatever its deleterious side effects, our study shows word stemming to be essential to proper vocabulary selection in a vector space representation of full-text documents.

Rank	Rubric Expert System	Vector Space Similarity	
		Cosine Measure	Euclidean Measure
1	** 8	** 8	** 8
2	** 19	** 3	** 3
3	** 9	** 19	** 19
4	** 22	** 22	** 22
5	** 21	** 29	** 9
6	** 23	** 9	** 26
7	** 3	24	20
8	* 4	** 26	24
9	** 26	** 23	** 23
10	** 29	25	25
11	* 27	20	12
12	* 5	* 4	11
13	* 7	12	** 29
14	2	* 7	** 21
15	24	6	17
16	25	11	6
17	6	17	15
18	14	2	* 27
19	17	* 27	30
20	20	* 5	* 7
21	1	10	* 4
22	10	1	* 5
23	11	14	14
24	13	16	28
25	15	18	16
26	16	28	10
27	18	30	18
28	12	13	13
29	28	** 21	1
30	30	15	2

Figure 6: Document Ranking for "violent acts of terrorism' query

** shows RUBRIC relevence

* shows RUBRIC marginally relevance

6.2. Euclidean Similarity Measure Results

Ranking results for the Euclidean similarity measure are not nearly so sanguine. To understand why, we must delve deeper into the quantitative values which determine the document ranking. Figure 7 plots retrieval status values, normalized to the interval [0,1] for the three rankings. Following the example of Tong [ToAs84], the relevant documents are arranged to the left, followed by marginally relevant, with non-relevant documents on the right.

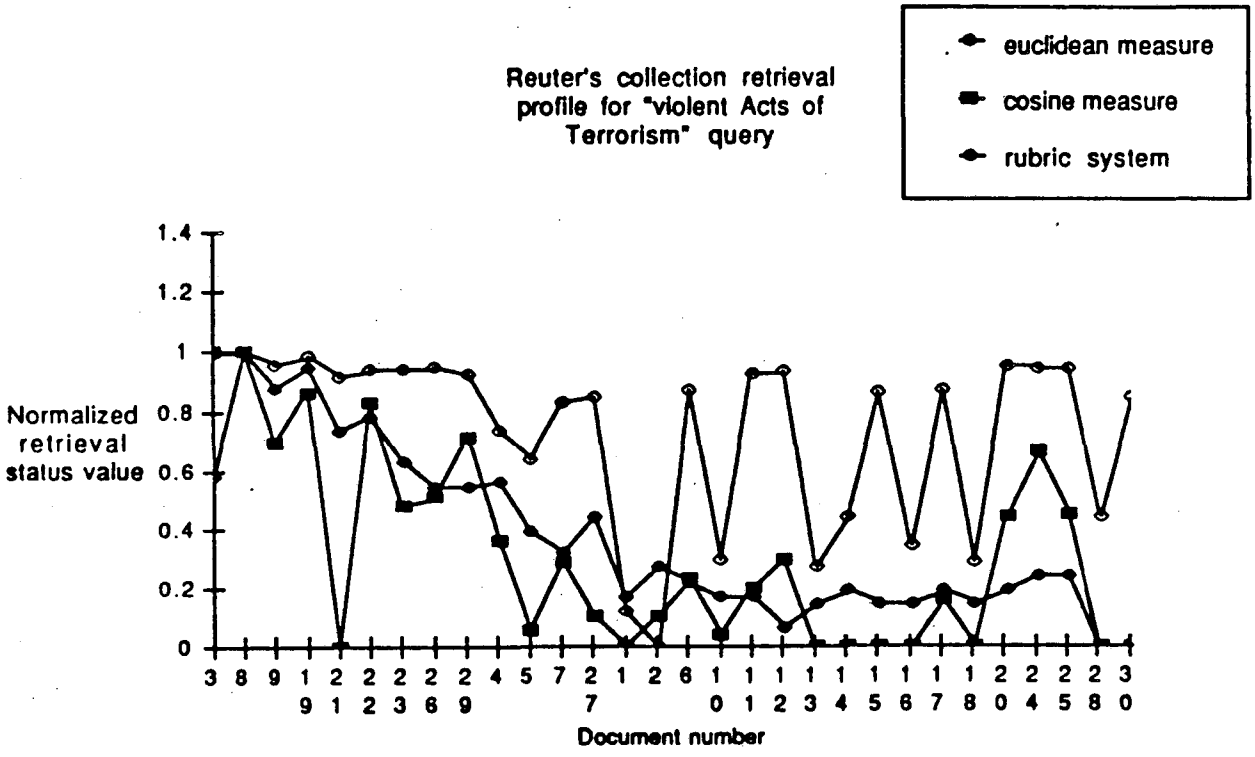


Figure 7: Individual Values for the RUBRIC query

Note that the Euclidean rank values offer only marginal discrimination between relevant and irrelevant documents. Indeed the distribution of RSVs seems to be bi-modal, with clusters of high and low values. The reason is well-known in the literature: the Euclidean measure, unlike the cosine measure, is not normalized by either query or document vector lengths (see [VanR 79], p. 39 for a conceptual discussion of normalization). Indeed we find Euclidean RSVs dominated by the distribution of terms in the documents. Figure 8 shows the total term occurance distribution together with a scatterplot of Euclidean RSV by term occurance.

The scatterplot shows (within the limits of this small collection) a straight line relationship between RSV and term occurance. The larger the term occurance, the smaller the normalized RSV, and consequently (for raw RSV) the greater the distance between the query and document vectors.

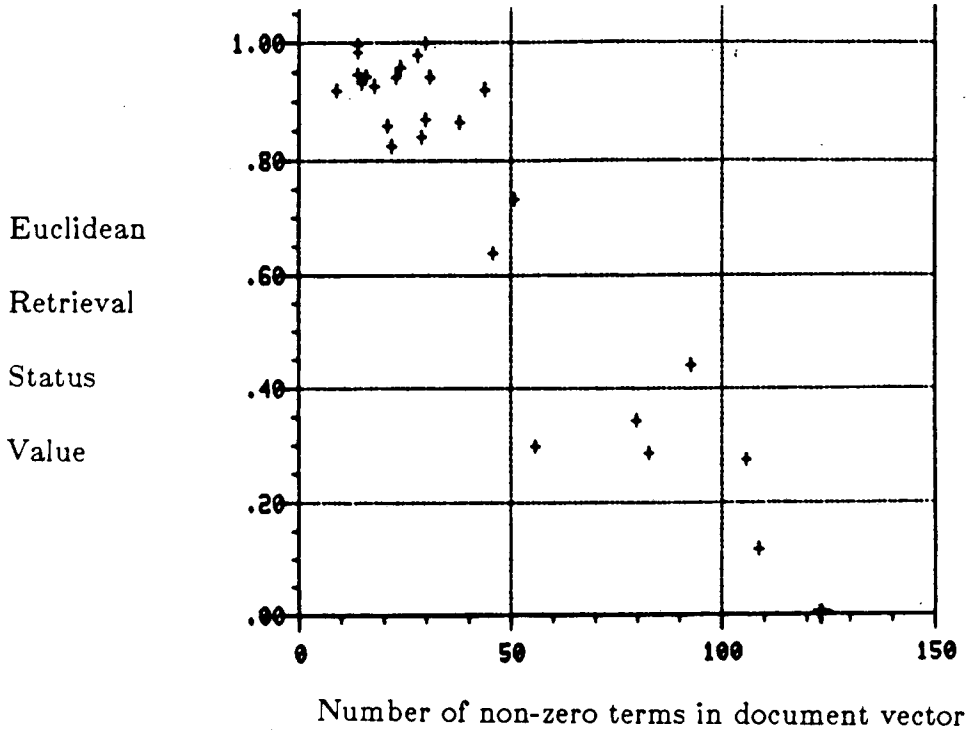


Figure 8: Euclidean retrieval status value versus term occurance

7. Summary and Conclusions

This project has compared the vector space model retrieval to the RUBRIC expert system. In the process we have rediscovered some fundamental truths about information retrieval. From our study we can conclude the following:

- Normalized vector space measures of similarity are essential to achieving unbiased retrieval results.
- Word stemming is important in constructing a vector space representation for full-text document collections.
- The vector space model using the cosine measure yields comparable results to the RUBRIC expert system for relevant documents. RUBRIC seems to perform better on retrieving marginally relevant documents. Since improvement at the margin is what IR is all about, RUBRIC makes a contribution to advancement of the field.

8. Acknowledgments

We would like to acknowledge Richard Tong's wholehearted contribution to the spirit of scientific inquiry in sending us the Reuter's collection and the RUBRIC document ranking values in ASCII machine- readable format, and for quickly answering questions about the RUBRIC results. Our thanks to our colleague at Lawrence Berkeley Laboratory, John L. McCarthy, who suggested the scatterplot for the Euclidean measure.

9. References

HARM 87 Harman, Donna, "A Failure Analysis on the Limitations of Suffixing in an Online Environment," *Proceedings of the Tenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA June 3-5, 1987, pp 102-107.

LUHN 58 Luhn, H.P., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, April, 1958. reprinted in *Key Papers in Information Science*, American Society for Information Science, Washington, D.C., 1971.

MARO 61 Maron, M.E., "Automatic Indexing: An Experimental Inquiry," *Journal of the ACM, v. 8, 1961, pp. 404-417* reprinted in *Key Papers in Information Science*, American Society for Information Science, Washington, D.C., 1971.

ToAp 87 Tong, Richard M., Lee A. Applebaum, Victor N. Askman, James F. Cunningham, "Conceptual Information Retrieval Using RUBRIC," *Proceedings of the Tenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA June 3-5, 1987, pp 247-253.

ToAp 86 Tong, Richard M., Lee A. Applebaum, Victor N. Askman, James F. Cunningham, "RUBRIC III, An Object-Oriented Expert System for Information Retrieval," *Proceedings of the Second Annual International IEEE Symposium on Expert Systems in Government*, McLean, VA October, 1986, pp 106-115.

ToAs 85 Tong, Richard M., Victor N. Askman, James F. Cunningham, Carl J Tollander "RUBRIC An Environment for Full Text Information Retrieval," *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Montreal, Quebec, CANADA, June 5-7, 1985, pp 243-250.

ToSh 85 Tong, Richard M., Daniel G Shapiro, "Experimental Investigations of Uncertainty in a Rule-Based System for Information Retrieval *International Journal of Man-Machine Studies (1985) v. 22, pp265-282.*

ToAs 84 Tong, Richard M., Victor N. Askman, James F. Cunningham "RUBRIC An Artificial Intelligence Approach to Information Retrieval, *Proceedings of the First International Workshop on Expert Database Systems*, Kiawah Island, South Carolina, October 1984.

ToSh 83 Tong, Richard M, Daniel G. Shapiro, Jeffrey S. Dean and Brian P. McCune, "A Comparison of Uncertainty Calculi in an Expert System for Information Retrieval, *Proc. International Joint Conference on Artificial Intelligence*, Karlsruhe, W. Germany, August 1983.

VANR 79 Van Rijsbergen, C. J. *Information Retrieval (Second Edition)*, Butterworths, London/Boston, 1979

WoRa 86 Wong, S.K.M, V.V. Raghavan, "A Critical Analysis of Vector Space Model for Information Retrieval," *Journal of the American Society for Information Science*, v. 37, 5, 279-287 (1986)

WoZi 87 Wong, S.K.M, W. Ziarko, V.V. Raghavan, P.C.N. Wong, "On Modeling of Information Retrieval Concepts in Vector Spaces," *ACM Transactions on Database Systems*, v. 12, 2 (June 1987), pp. 299-321.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720