Experimental Comparison of Schemes for Interpreting Boolean Queries

by

Whay C. Lee

Thesis 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

APPROVED:

Edward A. Fox, Chairman

J. Terry Nutter

Lenwood S. Heath

September, 1988

Blacksburg, Virginia

Experimental Comparison of Schemes for Interpreting Boolean Queries

by

Whay C. Lee

Edward A. Fox, Chairman

Computer Science

(ABSTRACT)

The standard interpretation of the logical operators in a Boolean retrieval system is in general too strict. A standard Boolean query rarely comes close to retrieving all and only those documents which are relevant to the user. An *AND* query is often too narrow and an *OR* query is often too broad. The choice of the AND results in retrieving on the left end of a typical average recall-precision graph, while the choice of the OR results in retrieving on the right end, implying a tradeoff between precision and recall. This study basically examines various proposed schemes, the P-norm, Classical Fuzzy-Set, MMM, Paice and TIRS, which provide means to soften the interpretation of the logical operators, and thus to attain both high precision and high recall search performance.

Each of the above schemes has shown great improvement over the standard Boolean scheme in terms of retrieval effectiveness. The differences in retrieval effectiveness between P-norm, Paice and MMM are shown to be relatively small. However, related performance results obtained gives evidence of the ranking: P-norm, Paice, MMM and then TIRS.

This study employs the INNER PRODUCT function for computing the similarity between a document point and a query point in TIRS. There may be other choices of similarity functions for TIRS, but irrespective of the function used, the TIRS approach,

having to deal with associated min-terms rather than the original query, is difficult to realize and involves far greater computational overhead than the other schemes.

The P-norm scheme, being a distance-based approach, has greater intuitive appeal than the Paice or MMM scheme. However, in terms of computational overhead required of each scheme, both the Paice and MMM are superior to P-norm. The Paice and MMM schemes are essentially variations of the classical fuzzy-set scheme. Both perform much better than the classical fuzzy-set scheme in terms of retrieval effectiveness.

Acknowledgements

It has been my privilege to have been associated with Dr. Edward A. Fox, who provided me with the opportunity to undertake graduate research in information storage and retrieval. Apart from being my academic advisor, Dr. Fox has been both a mentor and a very supportive co-worker. His persistence, insight and devotion to research has helped me build my character. I am deeply grateful to him for his consistent tolerance and scholarly supervision of my research work at Virginia Tech.

I sincerely extend my appreciation to Dr. Terry Nutter not only for her willingness to serve on my thesis committee, but also for introducing me to natural language understanding techniques which currently find fruitful applications in information retrieval. I am also grateful to Dr. Lenwood Heath who has been equally willing to serve on my thesis committee. His critical guidance has been and will remain useful to me.

In addition, I would like to thank "Doc" Layne Watson, who possesses inspiring mathematical resourcefulness in scientific research, for agreeing to comment on my study. I finally want to thank my personal friend, B. Chan, from The Johns Hopkins Medical Center, with whom I have had great pleasure discussing techniques for applied statistical analysis.

This research was funded in part by grants from the National Science Foundation (IST-841887 and IRI-8730580), the Virginia Center for Innovative Technology (INF-85-016 and INF-87-012) and by AT&T equipment contributions.

This research has been made possible by the use of the various IR test collections including CISI, CACM and INSPEC. The author hereby acknowledges the effort and kindness of the respective providers.

Table of Contents

Introduction 1

Limitations of Standard Boolean Retrieval 1

Alternative Approaches 3

Retrieval Models 6

Standard Boolean Retrieval 6

P-norm Retrieval 10

Fuzzy Set Retrieval 17

Classical Fuzzy-Set Approach 17

MMM Approach 18

Paice Approach 20

Topological Model 22

Retrieval Experiments 26

IR Test Collections 26

Method of Experiments 27

Similarity Computations 30

P-norm Similarity 30

MMM Similarity 31

Paice Similarity 31

TIRS Similarity 32

A Note On Boolean Query DNF Conversion 34

Complexity of Computations 36
Characteristics of Queries 39

Performance Results 42
Experimental Analysis on CISI Collection 42
P-norm Runs 42
Paice Runs 48
MMM Runs 54
TIRS Runs 59
Prediction Models 60
Discussion 67
The E-measures 73
Experimental Analysis on CACM and INSPEC Collections 78
CACM 78
Collected Results 78
Prediction Models 90
Discussion 96
INSPEC 104
Collected Results 104
Discussion 109

Conclusions 111

Bibliography 115

Vita 118

List of Illustrations

Figure 1. Matrix Representation for Document Collection 7

Figure 2. Typical Average Recall-Precision Graph 9

Figure 3. Retrieved Document Sets: Ranked vs. Unranked 12

Figure 4. P-norm Similarity Computations and Equi-similarity Contours 14

Figure 5. The Generalized P-norm Formulation 16

Figure 6. Sample Document and Query from CISI Collection 28

Figure 7a. The P-norm Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 45

Figure 7b. The P-norm Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$.....46

Figure 7c. The P-norm Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$.....47

Figure 8a. The Paice Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 51

Figure 8b. The Paice Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$.....52

Figure 8c. The Paice Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$.....53

Figure 9a. The MMM Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 56

Figure 9b. The MMM Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$.....57

Figure 9c. The MMM Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$.....58

Figure 10. Document 385 from CISI Collection 70

Figure 11. Document 375 from CISI Collection 71

Figure 12. Document 286 from CISI Collection 72

Figure 13a. The P-norm Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 81

Figure 13b. The P-norm Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$ 82

Figure 13c. The P-norm Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$ 83

Figure 14a. The Paice Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 84

Figure 14b. The Paice Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$ 85

Figure 14c. The Paice Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$ 86

Figure 15a. The MMM Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$ 87

Figure 15b. The MMM Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$ 88

Figure 15c. The MMM Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$ 89

Figure 16. Query 24 from CACM Collection 99

Figure 17. Document 1696 from CACM Collection 99

Figure 18. Document 749 from CACM Collection 99

List of Tables

Table 1a. Distributions of #Terms and #Min-terms in CISI Query Set 40

Table 1b. Distributions of #Terms and #Min-terms in CACM Query Set 41

Table 2. Average Precision Values with P-norm Scheme on CISI
for the set of coefficients: 1, 6, 12 and 50.....44

Table 3. Average Precision Values with P-norm Scheme on CISI 45

Table 4. Average Precision Values with Paice Scheme on CISI 51

Table 5. Average Precision Values with MMM Scheme on CISI 56

Table 6a. Stepwise Regression Results on CISI Collection
Summary of SAS Forward Selection Procedure.....63

Table 6b. Selected Prediction Models of Average Precisions on CISI
Best 3, 4, and 5-variable Models Obtained by MAXR.....64

Table 7a. Stepwise Regression Results on CISI Collection
(With Boundary Values Omitted)
Summary of SAS Forward Selection Procedure.....65

Table 7b. Selected Prediction Models of Average Precisions on CISI
(With Boundary Values Omitted)
Best 3, 4, and 5-variable Models Obtained by MAXR.....66

Table 8. Relative Ranks of Schemes
by Average Precision and E-measure on CISI 67

Table 9. Ten Top-ranked Documents Retrieved with Query 35 on CISI 69

Table 10. Relative Ranks of Schemes
by E-measure at β-levels 0.5, 1.0 and 2.0 on CISI.....73

Table 11a. E-measures with P-norm Scheme on CISI at β-level = 0.5.....75

Table 11b. E-measures with Paice Scheme on CISI at β-level = 0.5 75

Table 11c. E-measures with MMM Scheme on CISI at β-level = 0.5.....75

Table 12a. E-measures with P-norm Scheme on CISI at β-level= 1.0.....76

Table 12b. E-measures with Paice Scheme on CISI at β-level= 1.0 76

Table 12c. E-measures with MMM Scheme on CISI at β-level= 1.0 76

Table 13a. E-measures with P-norm Scheme on CISI at β-level = 2.0 77

Table 13b. E-measures with Paice Scheme on CISI at β-level = 2.0 77

Table 13c. E-measures with MMM Scheme on CISI at β-level = 2.0 77

Table 14. Average Precision Values with P-norm Scheme on CACM for the set of coefficients: 1, 6, 12 and 50 80

Table 15. Average Precision Values with P-norm Scheme on CACM 81

Table 16. Average Precision Values with Paice Scheme on CACM 84

Table 17. Average Precision Values with MMM Scheme on CACM 87

Table 18a. Stepwise Regression Results on CACM Collection Summary of SAS Forward Selection Procedure 92

Table 18b. Selected Prediction Models of Average Precisions on CACM Best 3, 4, and 5-variable Models Obtained by MAXR 93

Table 19a. Stepwise Regression Results on CACM Collection (With Boundary Values Omitted) Summary of SAS Forward Selection Procedure 94

Table 19b. Selected Prediction Models of Average Precisions on CACM (With Boundary Values Omitted) Best 3, 4, and 5-variable Models Obtained by MAXR 95

Table 20. Relative Ranks of Schemes by Average Precision and E-measure on CACM 96

Table 21. Ten Top-ranked Documents Retrieved with Query 24 on CACM 98

Table 22. Relative Ranks of Schemes by E-measure at β-levels 0.5, 1.0 and 2.0 on CACM 100

Table 23a. E-measures with P-norm Scheme on CACM at β-level = 0.5 101

Table 23b. E-measures with Paice Scheme on CACM at β-level = 0.5 101

Table 23c. E-measures with MMM Scheme on CACM at β-level = 0.5 101

Table 24a. E-measures with P-norm Scheme on CACM at β-level= 1.0 102

Table 24b. E-measures with Paice Scheme on CACM at β-level= 1.0 102

Table 24c. E-measures with MMM Scheme on CACM at β-level= 1.0 102

Table 25a. E-measures with P-norm Scheme on CACM at β-level = 2.0 103

Table 25b. E-measures with Paice Scheme on CACM at β-level = 2.0 103

Table 25c. E-measures with MMM Scheme on CACM at β-level = 2.0 103

Table 26a Average Precision Values on INSPEC: P-norm Scheme 105

Table 26b Average Precision Values on INSPEC: Paice Scheme 105

Table 26b Average Precision Values on INSPEC: MMM Scheme 105

Table 27a E-measures at $\beta = 0.5$ on INSPEC: P-normScheme 106

Table 27b E-measures at $\beta = 0.5$ on INSPEC: Paice Scheme 106

Table 27c E-measures at $\beta = 0.5$ on INSPEC: MMM Scheme 106

Table 28a E-measures at $\beta = 1.0$ on INSPEC: P-norm Scheme 107

Table 28b E-measures at $\beta = 1.0$ on INSPEC: Paice Scheme 107

Table 28c E-measures at $\beta = 1.0$ on INSPEC: MMM Scheme 107

Table 29a E-measures at $\beta = 2.0$ on INSPEC: P-norm Scheme 108

Table 29b E-measures at $\beta = 2.0$ on INSPEC: Paice Scheme 108

Table 29c E-measures at $\beta = 2.0$ on INSPEC: MMM Scheme 108

Table 30 Relative Ranks of Schemes by Average Precision and E-measure on INSPEC 109

Table 31 Relative Ranks of Schemes by E-measure at β-levels 0.5 and 2.0 on INSPEC 109

Chapter 1

Introduction

Limitations of Standard Boolean Retrieval

The limitations of traditional Boolean retrieval systems are well-known. Among the most critical limitations are those annotated by Bookstein [BOOKS 85]:

- Boolean logic may produce counterintuitive results that are technically correct but are often disturbing to many users. For example, consider a request (A or B or C ... or Z). A Boolean retrieval system responds identically to a document indexed by a single one of these terms as it does to a document indexed by all terms. Similarly, for a request (A and B and C ... and Z), a document indexed by all but one of the query terms is deemed just as useless as a document not indexed by any of them.

- There is no provision for assigning importance factors or weights to index terms for both the documents and the queries; each index term is thus considered as important as any other index terms. In parallel with this limitation is the unavailability of a ranking mechanism for the retrieved documents according to the degree of relevance to the query, though such is believed by many to be desirable.

- The conceptual model that underlies Boolean retrieval does not in theory recognize uncertainty and incompleteness often intrinsic to both indexing and retrieval. There is no systematic scheme for a user to modify a request in a subsequent search attempt in response to feedback about the quality of the initial set of items retrieved. Also, index terms that describe a collection of documents are generally fixed; they cannot be updated easily in response to feedback on how well they are performing.

- It is difficult for a general user to formulate a near-optimal, if not perfectly optimal, Boolean query. Depending upon the assignment frequency of the query terms and the actual term combinations used in the formulation, too many or too few of the documents in the collection may be retrieved. Also, the correct Boolean format for many searches of user interest is usually lengthy or awkward; a clear example is the formulation in a search for documents indexed by, say, any three of ten given terms.

Alternative Approaches

As an alternative to the standard Boolean model, Salton, Fox and Wu [SALT 83a] proposed the P-norm model which allows weighted terms to be incorporated into both the documents and the queries, and also retrieved documents to be ranked in strict similarity rule with the input query. Extensive experiments have demonstrated the capability of such an approach in improving retrieval effectiveness. Other schemes such as that of Paice [PAICE 84], Cater & Kraft [CATER 87], and Tong & Shapiro [TONG 85], which subsequently appeared in the literature, have also been suggested as being useful in remedying some of the limitations of the standard Boolean model. Since these methods have not been carefully compared with one another, it is important here to consider how they relate and to identify key theoretical and performance characteristics of each.

In similar vein to Salton et al., Paice argued strongly in favor of the use of 'soft' logical operators. While the P-norm approach is distance-based, Paice's method expands on fuzzy versions of the conventional set operations. Another scheme similar to that of Paice and classical fuzzy-sets, which we called Mixed Max and Min (MMM), is also considered in this study. Tong and others, however, investigated the effects of different representations of uncertainty in RUBRIC, an interactive rule-based expert retrieval system whose implementation draws heavily upon fuzzy-set theory. Fuzzy-set schemes or the like have been shown to give significantly better retrieval performance over traditional Boolean retrieval scheme, since they allow users to construct descriptors of documents that reflect more specific needs and interests.

In his doctoral work, Cater [CATER 86] attempted to come up with a unifying model of information retrieval (IR) which will encompass all the standard models -- the Boolean, vector space, fuzzy-set theoretic, probabilistic, and hierarchical models. He and Kraft constructed the TIRS model based on the topological paradigm. He claimed that the P-norm scheme of similarity computation is inadequate, and that their topological model is able to retain all of the advantages of the P-norm model, while not sharing its other apparent weaknesses. However, his criticisms and claims are subject to debate.

This paper reports a comparative study of the different approaches taken by the above research teams. An outline of each of these approaches is first provided, and then experimental results are collected for comparing their retrieval effectiveness. The complexity and computational overhead of these approaches are also considered. For details of each scheme, however, the reader is referred to the publications of the individual authors.

Fox and Sharan [FOXE 86] performed a preliminary study of the P-norm and MMM schemes for soft Boolean operator interpretations in information retrievel on a test collection. The findings of Fox and Sharan suggest that fuzzy set membership functions are very valuable for the construction of information retrieval systems, the logical interpretations of *AND* and *OR* are too strict for Boolean queries, and also that the P-norm retrieval scheme is generally more effective than the MMM scheme. The current work provides a more complete study of various schemes for softening logical interpretations of Boolean queries, including the P-norm, Paice, MMM and TIRS. The earlier findings are confirmed and new insights are obtained to give a better

understanding of how standard Boolean retrieval method can be improved upon for high performance searches.

Chapter 2

Retrieval Models

Standard Boolean Retrieval

Consider a standard Boolean retrieval system with a collection of N documents. A user wishes to select a subset of the documents relevant to a query Q. Let us denote such a subset by D_Q^R. In the case that N is large, one cannot avoid the use of a computer to automate such a process. To allow automatic matching on the computer, the collection must be represented in some form that essentially captures data in a structure, such as the N by M matrix shown in Figure 1. Here, M is the number of index terms chosen to characterize the N documents.

$$
\begin{array}{cc}
 & \textit{Index Terms} \\
 & \begin{array}{cccc} T_1 & T_2 & \ldots & T_M \end{array} \\
\textit{Documents} & \begin{bmatrix}
D_1 & d_{11} & d_{12} & \ldots & d_{1M} \\
D_2 & d_{21} & d_{22} & \ldots & d_{2M} \\
\vdots & \vdots & \vdots & & \vdots \\
D_N & d_{N1} & d_{N2} & \ldots & d_{NM}
\end{bmatrix}
\end{array}
$$

Figure 1. Matrix Representation for a Document Collection

The *ith* document denoted by D_i can thus be viewed as a vector $(d_{i1}, d_{i2}, \ldots, d_{iM})$, each of whose elements (d_{ij}) takes a value of 0 or 1, depending on whether the corresponding *jth* term is assigned to the document or not. Boolean queries can easily be understood in the context of this representation. For simplicity, consider queries with only two terms as follows:

$$Q_{\text{and}} = (T_\alpha \ AND \ T_\beta)$$

$$Q_{\text{or}} = (T_\alpha \ OR \ T_\beta)$$

The standard Boolean interpretations for the above queries are respectively the logical *AND* and *OR* of the two vectors T_α and T_β, the first of which gives their intersection and the second, their union, as the retrieved set of documents. Formally, the retrieved set of documents are represented as follows:

$$D^{Bool}_{Q_{\text{and}}} = \{d_i \mid d_{i\alpha} \wedge d_{i\beta}\}$$

$$D^{Bool}_{Q_{\text{or}}} = \{d_i \mid d_{i\alpha} \vee d_{i\beta}\}$$

For high *precision*[1] retrieval, especially in the case where N is large, an AND query is recommended since often

$$|D_{Q_{\text{and}}}^{Bool}| \ll min(|T_\alpha|, |T_\beta|).$$

However, for casual or high *recall*[2] search, an OR query is advised. This is particularly so when the terms T_α and T_β are nearly synonymous or very closely related, since

$$max(|T_\alpha|, |T_\beta|) \leq |D_{Q_{\text{or}}}^{Bool}| \leq |T_\alpha| + |T_\beta|$$

and if the two terms co-occur frequently, the size of the retrieved set is close to the lower bound.

It is important to note that a Boolean query rarely comes close to retrieving all and only those documents which are relevant. An AND query is often too narrow, and conversely, an OR query is often too broad. As indicated by many previous experimental studies [SALT 83b], a typical average recall-precision graph is of the form in Figure 2, for which there is, on average, a trade-off between the two measures of retrieval effectiveness.

---

[1] $Precision = \frac{\text{the number of relevant retrieved documents}}{\text{the number of retrieved documents}}$

[2] $Recall = \frac{\text{the number of relevant retrieved documents}}{\text{the number of relevant documents}}$

Because of the strictness of the standard Boolean retrieval, it is almost impossible for a typical user to construct optimal queries which will provide both high precision and high recall searches, resulting in a higher average recall-precision curve. Modern advanced retrieval techniques based on P-norm or fuzzy-set concepts, to some extent, enable the user to shift the recall-precision curve upward by letting the computer interpret the query with a better 'understanding' of what the user means by it.

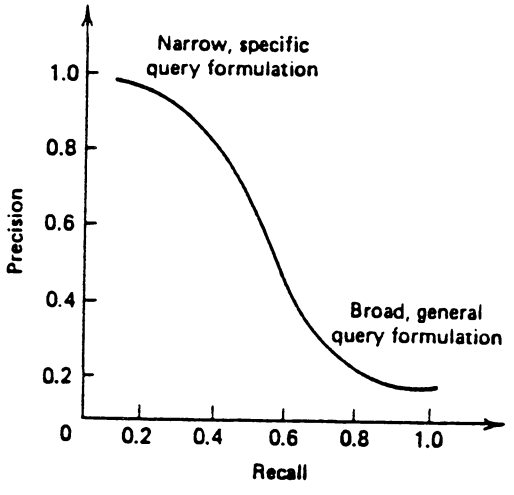


Figure 2. Typical Average Recall-Precision Graph

P-norm Retrieval

The P-norm approach is based on a distance measure. It includes the following assumptions:

- Indexing is a fuzzy process, and one should allow each term weight, d_{ij} to vary between 0 and 1.

- A query should define a fuzzy set so that documents can be presented to users in order of decreased degree of relevance.

- Strict logical interpretation of *AND* and *OR* is inappropriate, since linguistic relationships frequently do not correspond to statistical reality for retrieval.

In this model, non-binary term weights can be used to reflect relative term importance for terms assigned to documents and queries. Several weighting schemes have been proposed; a key consideration, however, is that terms which occur often in a document are more likely to characterize it than terms that occur less often. The use of non-binary weighting schemes is also important from another perspective. Paice in his study [PAICE 84] pointed out that one limiting function in IR systems relates to the more or less tentative nature of the choice of terms to denote particular concepts, which results from

- the existence and possible use of alternative terms that can lead to potential loss of recall, and

- the possible use of a particular term with other meanings in other contexts that can lead to a potential loss of precision.

Thesauri have commonly been used to control the assigment of index terms to documents. However, linguistic research is now underway for means to allow users to formulate more 'precise' queries. Incorporating unsuitable search terms into the queries will often lead to low retrieval performance. A non-binary weighting scheme, in some manner, enables us to register the tentativeness of a term relative to other terms in the document collection without risk of much performance degradation but with the hope of significant performance improvement.

Experimental studies [SALT 83b] have shown that a weighting technique such as that using 'term frequency' and 'inverse document frequency', with the *jth* term in the *ith* document being given the weight

$$d_{ij} = tf_{ij} \times idf_j$$

can lead to better retrieval than the ordinary binary scheme.

In this study, we have adopted the following weighting scheme:

$$d_{ij} = (0.5 + 0.5 \times \frac{tf_{ij}}{max\_tf}) \times \frac{\log(N/f_j)}{\log(N)}$$

where

N = the total number of documents in the collection

max_tf = the freq. of the most widely occurring term in the document

tf_{ij} = the number of occurrences of term j in document i

f_j = the number of documents containing term j.

Note that d_{ij} is set to 0 if tf_{ij} is equal to 0.

Regarding the issue of having a ranked versus an unranked set of retrieved documents (as shown in Figure 3), it is not difficult to see that there are two clear advantages in the first. In an interactive environment, documents can be displayed in rank order, so that the 'best' retrieved item is seen before others by the user. In a batch environment, the ranked output list can be cut off at a certain level of query-document similarity, thus controlling the amount of output.

A. Fuzzy Set

Document	*Similarity*
ID001	0.989
ID742	0.843
ID253	0.632
:	:
ID004	0.104
ID106	0.632
:	:
ID695	0.104

B. Boolean Set

Document	*Similarity*
ID001	1.0
:	:
ID223	1.0
ID104	0.0
ID105	0.0
:	:
:	:
ID997	0.0

Figure 3. Retrieved Document Sets: Ranked vs. Unranked

Another problem with the conventional Boolean model is that it is much too strict. With respect to the Boolean query

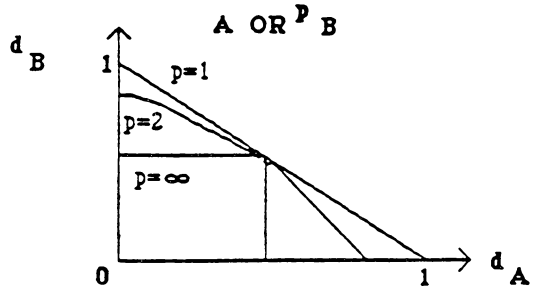
FIND A AND B AND C AND D AND E,

a document is retrieved if and only if it contains a match with each of the search terms, *A* through *E*. As Paice [PAICE 84, p. 36] explicitly puts it, "irrespective of the user's view of the correctness of this search formulation, however, the vagaries of index-term

assignment and of search-term selection are such that many relevant items might contain only three or four of the stated terms." The P-norm model associates a parameter with the logical operators in a query formulation so as to soften their interpretation, and thus avoid missing out too many relevant items in a search like the above. As the p value changes from ∞ to 1, the logical operators are interpreted more and more loosely. When $p = \infty$, in the case where both the document and query terms carry binary weights, the P-norm model essentially reduces to the standard Boolean model. When $p = 1$, the P-norm model becomes a version of the vector processing model, in which there is no distinction between compulsory phrase bonding (using *AND*) and alternative synonym specification (using *OR*).

Figure 4 shows the P-norm similarity computations and the equi-similarity contours for the two-term queries (*A OR B*) and (*A AND B*), with respect to a document *D* which has d_A and d_B as the term weights for the two search terms. The axes indicate the membership function values for the document for term *A* and term *B*. The contours X, Y, Z are defined for constant p values at levels 1, 2 and ∞, respectively, to connect documents with equal levels of similarity.

$$SIM(A\ OR^p\ B, D) = \left\{ \frac{d_A^p + d_B^p}{2} \right\}^{1/p} = 2^{-1/p} |D|_p$$



$$SIM(A\ AND^p\ B, D) = 1 - \left\{ \frac{(1-d_A)^p + (1-d_B)^p}{2} \right\}^{1/p} = 1 - 2^{-1/p} |1-D|_p$$

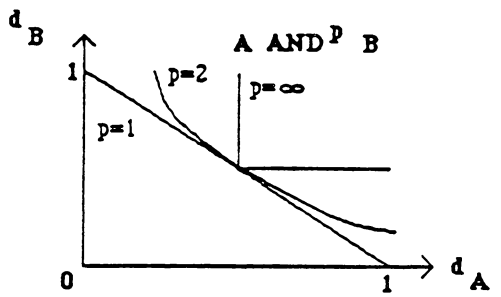


Figure 4. P-norm Similarity Computations and Equi-similarity Contours

For *OR* queries, similarity follows the intuition that one would like to be as far as possible from having no terms present in any document retrieved. Thus, similarity is a normalized distance from the origin, (0,0), in the subspace defined by the query terms. For *AND* queries, however, the point (1,1) represents the situation when both terms are fully present in a document, and is thus the most desirable location. Retrieved documents should therefore be ranked in order of increasing distance from it.

It is noted that when $p = \infty$ and all the query terms carry equal weights, the P-norm similarity computations [SALT 83a, p. 1024] in theory simplify to the classical fuzzy-set formula as shown below:

$$SIM\,(A\ OR^{p=\infty}\ B, D) = max\,(d_A,\ d_B)$$

$$SIM\,(A\ AND^{p=\infty}\ B, D) = min\,(d_A,\ d_B)$$

Furthermore, P-norm operators can be generalized to handle many terms in a clause instead of just two, and to handle user specified relative weights on each of the query clauses or terms. To complete the definition of the P-norm model, the similarity computation for a *NOT* query is given as

$$SIM\,(NOT\ <\text{expression}>\,, D) = 1 - SIM\,(<\text{expression}>\,, D)$$

A more elaborate example which involves more than two query terms will be given in a later section. The generalized P-norm formulation is given in Figure 5, and a discussion of the classical fuzzy-set formulation is found in the following section.

Consider a set of index terms $A_1, A_2, \ldots, A_n$, and let d_{A_i} represent the weight of term A_i in some document $D = (d_{A_1}, d_{A_2}, \ldots, d_{A_n})$, where $0 \leq d_{A_i} \leq 1$. The generalized Boolean OR and AND queries are written as

$$Q_{OR,p} = (A_1, a_1)\, OR^p\, (A_2, a_2)\, OR^p \ \ldots\ OR^p\, (A_n, a_n)$$

$$Q_{AND,p} = (A_1, a_1)\, AND^p\, (A_2, a_2)\, AND^p \ \ldots\ AND^p\, (A_n, a_n)$$

The similarities between the document D and $Q_{OR,p}$, $Q_{AND,p}$ are then defined as

$$SIM\,(Q_{OR,p}, D) = \left\{ \frac{a_1^p d_{A_1}^p + a_2^p d_{A_2}^p + \cdots + a_n^p d_{A_n}^p}{a_1^p + a_2^p + \cdots + a_n^p} \right\}^{1/p}$$

$$SIM\,(Q_{AND,p}, D) = 1 - \left\{ \frac{a_1^p (1 - d_{A_1})^p + a_2^p (1 - d_{A_2})^p + \cdots + a_n^p (1 - d_{A_n})^p}{a_1^p + a_2^p + \cdots + a_n^p} \right\}^{1/p}$$

Figure 5. The Generalized P-norm Formulation

Fuzzy Set Retrieval

Classical Fuzzy-Set Approach

Zadeh [ZADEH 65] developed the concept of a fuzzy set to allow for partial set membership. Within the context of information retrieval, a fuzzy set of documents is associated with each index term. As in the P-norm model, for each given term, a membership function is defined that indicates to what degree each document is characterized by that term. In fuzzy-set systems, the retrieved set of documents is itself a fuzzy set, whose membership function is derived from membership functions associated with the index terms by means of a set of manipulation rules on fuzzy sets.

The most fundamental set of manipulation rules on fuzzy sets is given below. If a document D is in the set A to a degree d_A and is in the set B to a degree d_B, then the membership functions for the union, intersection and complement are defined as follows:

$$d_{(A \cup B)} = max\,(d_A, d_B)$$

$$d_{(A \cap B)} = min\,(d_A, d_B)$$

$$d_{\bar{A}} = 1 - d_A$$

Consider a simple example with documents D_1 and D_2 indexed as follows by term T_α and term T_β: $D_1 = \{(T_\alpha, 0.5), (T_\beta, 0.8)\}$; $D_2 = \{(T_\alpha, 0.9), (T_\beta, 0.1)\}$. It then follows that, associated with the index terms are the sets: $T_\alpha = \{(D_1, 0.5), (D_2, 0.9)\}$ and $T_\beta = \{(D_1, 0.8), (D_2, 0.1)\}$. The set of documents retrieved by the query, $(T_\alpha\ AND\ T_\beta)$, is

$\{(D_1, 0.5), (D_2, 0.1)\}$, since 0.5 is the minimum of 0.5 and 0.8 and since 0.1 is the minimum of 0.9 and 0.1.

Critics of such an approach argue that an appropriate scheme for ranking of the output for multiterm clauses should be sensitive to all of the terms in the query. As in the example above, the document D_2 belongs to the fuzzy set associated with $(T_\alpha \cap T_\beta)$ to a degree 0.1; this intersection membership value would not have been different, had the degree to which the document is in the set associated with T_α, been changed from 0.9 to 0.2. A similar phenomenon applies to unions. Nonetheless, the appeal of the above definitions for fuzzy-set operations lies in the fact that they retain most of the usual axioms of conventional set theory. Alternative definitions [YAGER 80] have been considered by many researchers in various applications. It is our intention not to expound on each of them, but to present several which may have practical applications in IR systems.

MMM Approach

We now consider the Mixed Max and Min (MMM) approach [FOXE 86]. For the two queries below

$$Q_{OR} = (T_1 \; OR \; T_2 \; OR \; \ldots \; OR \; T_k)$$

$$Q_{AND} = (T_1 \; AND \; T_2 \; AND \; \ldots \; AND \; T_k)$$

the similarities with respect to a document D, which has $(d_1, d_2, \ldots, d_k)$ indicating the corresponding term weights, are computed as follows:

$$SIM(Q_{or}, D) = Coeff_{or,1} \times max(d_1, d_2, \ldots, d_k) + Coeff_{or,2} \times min(d_1, d_2, \ldots, d_k)$$

$$SIM(Q_{and}, D) = Coeff_{and,1} \times min(d_1, d_2, \ldots, d_k) + Coeff_{and,2} \times max(d_1, d_2, \ldots, d_k)$$

Usually, it would be desirable that $Coeff_{or,1} > Coeff_{or,2}$, and $Coeff_{and,1} > Coeff_{and,2}$, since *OR* should be more similar to *max* than to *min*, and *AND* should be more similar to *min* than to *max*. For simplicity, the following settings are suggested:

$$Coeff_{or,2} = 1 - Coeff_{or,1}$$

$$Coeff_{and,2} = 1 - Coeff_{and,1}$$

In this study. the above suggestion is taken. Thus, it is noted that when the first coefficients of both *AND* and *OR* are set to 1, the MMM approach becomes the classical fuzzy-set approach.

What about the negation operation? Though the *NOT* formula $d_{\bar{o}} = 1 - d_o$ is generally reasonable for most purposes, it is important to observe that the use of *NOT* has two different implications. One form of *NOT* serves to positively identify objects that can fall into one of several categories but not a single specific category, much as a request for 'foreign' books is expressed as a request for books that are '*NOT* English'. The second form, as in the request 'Information Retrieval but *NOT* fuzzy', is more commonly used to re-scan and subsequently narrow down the set of items already retrieved.

Paice Approach

For the Paice approach [PAICE 84], we have

$$SIM(Q, D) = \frac{\sum_{i=1}^{n} r^{i-1} d_i}{\sum_{i=1}^{n} r^{i-1}} \qquad \begin{cases} Q = Q_{OR}, & \text{descending order of } d_i\text{'s} \\ Q = Q_{AND}, & \text{ascending order of } d_i\text{'s} \end{cases}$$

Depending upon whether the query Q is an OR query or an AND query, the above similarity computation is carried out with descending or ascending order of the term weights, d_i's. It is noted that each subsequent term weight receives a weighting factor which is a fixed ratio r (between 0 and 1) of the preceeding weighting factor. In this study, this scheme is so implemented that when $r = 0$, only the *max* in the sum is considered for an OR-query, and only the *min* in the sum is considered for an AND-query. Thus, with r set to 0, the Paice scheme behaves exactly as the classical fuzzy-set scheme.

To illustrate the Paice approach, we consider the case with $n = 2$. Suppose that a document D has d_1 and d_2 as the corresponding weights for the terms T_1 and T_2 used in the following two queries.

$$Q_{OR} = (T_1 \; OR \; T_2)$$

$$Q_{AND} = (T_1 \; AND \; T_2)$$

Using the Paice scheme with $r = 1/4$, we have

$$SIM(Q_{or}, D) = \frac{max\,(d_1, d_2) + 1/4\ min\,(d_1, d_2)}{1 + 1/4} = 4/5\ max\,(d_1, d_2) + 1/5\ min\,(d_1, d_2)$$

$$SIM(Q_{and}, D) = \frac{min\,(d_1, d_2) + 1/4\ max\,(d_1, d_2)}{1 + 1/4} = 4/5\ min\,(d_1, d_2) + 1/5\ max\,(d_1, d_2)$$

Incidentally, with $n = 2$ the Paice approach is just the same as the MMM approach.

As in the P-norm model, neither the MMM nor the Paice approach conform to the distributive property of conventional Boolean logic. Under all these approaches, it is generally not the case that the following two logically equivalent Boolean queries

FIND A AND (B OR C)

FIND (A AND B) OR (A AND C)

would lead to the same retrieval result. As demonstrated in [PAICE 84], discrepancies tend to appear in terms of the rank orders for the two forms of queries. We will explore later ways to determine which ranking more closely approximates user desires.

Topological Model

The Topological model [CATER 87] has been proposed as a generalization of the P-norm model. Cater and Kraft constructed the TIRS system based upon the following topological paradigm:

- The document space is a metric space, in which it is possible to find the distance between any pair of points.

- The document space is a product space, since each document is represented by a set of attribute terms. Arguing against the requirement for a fixed number of attribute terms in automatic indexing systems which rely on term frequencies in the text of the document, Cater and Kraft make a specific assumption in TIRS that there is available an infinite but countable collection of attribute terms.

- Each attribute term is weighted with an appropriate totally ordered set. Real-value weights within the interval [0,1] suffice for our purpose here.

- A query is considered to be a finite set of points in the document space, with each point in the set representing a possible "perfect" document.

In TIRS, the document space *DS* is defined as a countably infinite metric space as below:

$$DS = \overset{\infty}{\underset{i=1}{\mathrm{X}}}[0, 1]$$

A document D will be a point in DS; D can have only a finite number of non-zero terms. An arbitrary point in DS can be denoted as $D = (d_1, d_2, \ldots, d_n, 0, \ldots)$ where $\{d_i\}_{i=1}^{n}$ are elements of [0,1], with d_n being the last non-zero element.

As mentioned, a query is treated as a finite set of document descriptions. The evaluation of the query proceeds by first constructing the projection space with only the coordinates (attribute terms) given in the query and then finding the documents closest to the query with the metric of the defined document space.

Cater & Kraft have identified different forms of queries that TIRS can handle when supplied by users with various levels of experience. The simplest form is that of a query consisting of a single point in DS. Such a query can be interpreted either as a vector space model query, or as a Boolean query containing only AND operators. For a given query of this type denoted by $Q = (q_1, q_2, \ldots, q_k, 0, \ldots)$ and a clause weight w, all documents that are of distance w or less away from Q are retrieved:

$$D_Q^{TIRS} = \{x \mid dist(Q, x) \leq w\}$$

The second form is that of a query set containing $j > 1$ elements as in $Q = \{Q_1, Q_2, \ldots, Q_j\}$, where each $Q_i = (q_{i1}, q_{i2}, \ldots, q_{ik}, 0, \ldots)$ is of the previous form. If w_i is the weight of the query point Q_i, then the retrieved set of documents is given by

$$D_Q^{TIRS} = \{x \mid \exists i\, (i \in 1 \ldots j)\, dist(Q_i, x) \leq w_i\}$$

A document is retrieved if it is sufficiently close to some query point. The documents retrieved can thus be ranked in order of their overall distances from the closest query points. Equivalent to this type of query is the Boolean query using AND and OR which

is expressed in disjunctive normal form (DNF), with weights associated with each disjunct.

For a "traditionally experienced" user who is familiar with standard Boolean queries, the input query would often be one with a collection of weighted attribute terms connected by the binary Boolean operators *AND* and *OR*, as well as the unary Boolean operator *NOT*. As long as the input query is non-contradictory, there exists an equivalent disjunctive normal form (DNF) suitable for TIRS. The TIRS retrieval process then begins after each atom of the form *NOT*(*A*, *weight*) within the DNF query is replaced by the atom (*A*, 1 − *weight*). A DNF query can be easily converted into a set of points in the document space, each formed from a clause in that representation of the query. Associated with *ith* point obtained from the query is a relevance ball, whose radius essentially corresponds to the clause weight, w_i. Once the balls have been assigned, points within the balls are simply retrieved as the output documents.

The TIRS scheme allows the experienced user to adjust the relevance ball radii, or clause weights, so as to maximize the recall-precision product. Consider a simple query as below:

$$(<\text{information}, 0.8> \ AND \ <\text{retrieval}, 0.8>)$$
$$OR$$
$$(<\text{information}, 0.8> \ AND \ <\text{science}, 0.4>)$$

Assuming again that the attribute terms given are in the first few in *DS*, we have the following associated query points:

$$\begin{bmatrix} (<0.8, 0.8, 0.0, 0.0, ...>, 2.5) \\ (<0.8, 0.0, 0.4, 0.0, ...>, 1.0) \end{bmatrix}$$

The first point corresponding to the clause 'information retrieval' has a clause weight of 2.5, and the second corresponding to the clause 'information science' has a clause weight of 1.0. The document points will then be retrieved accordingly, with each retrieved point ranked by its distance from its closest query point in the *DS*. It is noted that a user requesting documents on 'information retrieval', and much less on 'information science' should set the clause weight for the first to be relatively larger than that for the second, since the first clause describes the user's need more appropriately than the second.

Note that in the representation above a term weight of zero is assigned to each term not present in the original DNF clause. In our implementation of TIRS, however, we treat a term being absent as a "don't care" (i.e. any value would do, including 0 and 1) and so add disjuncts with value 0 and with value 1 for all combinations of don't cares.

Chapter 3

Retrieval Experiments

IR Test Collections

Three IR test collections are used in this study so as to avoid the results being unnecessarily influenced by the characteristics of a particular data set. These IR test collections [SALT 83a] are CISI, CACM and INSPEC. The CISI collection contains 1460 articles on library science selected based on citation data from the Institute of Scientific Information (ISI). The CACM collection contains 3204 articles from the *Communications of ACM* published between 1958 and 1979. The INSPEC collection has 12684 articles from *Computer and Control Abstracts* which are basically concerned with electrical engineering and computer science.

Each article in the collections is considered as a document, and is given an identification number. A sample document and a sample query taken from CISI are

shown in Figure 6. The document is divided into several fields described by a set of concept types. The beginning of a document is marked by a line which contains the document ID. For example, document 18 in that figure has four fields -- namely, the TITLE, the AUTHOR, the DATE and the WORDS. The WORDS and TITLE fields are described by a single concept type which we will simply call WORD.

A typical query is given in prefix notation Boolean form, and the query terms are essentially of concept type WORD unless they are otherwise specified. Each query term can also be given a weight. In this study, the document terms are weighted with suitable real numbers, and query terms are only allowed to have binary weights (hereafter, are unweighted). The Boolean connectors *AND* and *OR* in the query have associated parameter values, the adjustment of which forms an important part of our experimental study.

Method of Experiments

The SMART retrieval program [FOXE 83b and BUCK 85] has been extended by this author so as to be able to run experiments using the P-norm, MMM, Paice and TIRS schemes. Each IR test collection above contains its own set of Boolean queries and their corresponding relevance judgements. Relevance judgements were established by experts who had examined individual documents in the collection and decided which document is relevant to which query. For CISI, full relevance information is available, but for the other two collections, it is only approximated.

The document D^{18} is as follows:

.I 18
.T
Selective ***Dissemination*** of ***Information***
.A
Mauerhoff, G. R.
.B
1974
.W
The present contribution does not duplicate previous studies but complements the earlier publications and closes the few gaps that exist in the literature prior to 1966 and after 1971. Additionally, it is a bold attempt to evaluate critically and objectively the history of the mechanized selective ***dissemination*** of ***information*** (SDI) as reflected in the literature, from the initial description by Luhn (1958, 1961b, c) to the post-1970 period when the SDI boom began losing ground to the more popular on-line interactive systems. The review therefore questions and interprets the concept of SDI, its implementation, and its evolution in the light of work performed by many companies, ***government agencies***, universities, societies, and libraries during the last fourteen years.

The query Q^{35} is as follows:

$$Q^{35} = (\ AND\ (<\text{government}, 1.0>,\ OR\ (<\text{information}, 1.0>, <\text{dissemination}, 1.0>, <\text{agencies}, 1.0>, <\text{projects}, 1.0>)))$$

Note that each of the terms in the query is given a weight 1.0. The corresponding terms are highlighted in the document above, and the weights of the terms are given below. Since the term 'projects' is not found in this document, it is assigned a zero document weight.

$$d_{\text{government}} = 0.28904$$
$$d_{\text{information}} = 0.09098$$
$$d_{\text{dissemination}} = 0.35416$$
$$d_{\text{agencies}} = 0.38384$$
$$d_{\text{projects}} = 0.00000$$

Figure 6. Sample Document and Query from CISI Collection

With relevance judgements, the SMART system is able to compute definitive recall and precision measures for each query and for each level of recall. The recall represents the proportion of relevant documents retrieved, whereas the precision represents the proportion of retrieved documents that are actually relevant. Here, we are essentially concerned with the average precision [SALT 83b] for the set retrieved by each query over the set of all queries and at three standard recall levels, 0.25, 0.50 and 0.75. Since there

are weaknesses in the use of average precision as the sole measure of retrieval effectiveness, the E-measure is also taken into consideration in this study.

The E-measure, which was first introduced by [SWETS 69], is thoroughly discussed in [RIJSB 79, pp. 174-175]. The E-measure is a weighted combination of precision and recall. The lower the E-measure, the greater is the retrieval effectiveness. The E-measure is computed based on a given β-level and a set of retrieved documents. The β-level is used to reflect the emphasis on recall or precision. Setting β to 1 implies attaching equal importance to both recall and precision, while setting β to 0.5 or 2 implies attaching half or twice as much importance to recall as to precision. The set of retrieved documents is defined by establishing a cutoff point in the document ranking. To effect a realistic comparison, we compute the E-measures using the top 30 documents.

The parameter values or coefficients associated with the Boolean connectors, *AND* and *OR*, are regarded as independent variables in this experimental study. The dependent variable can be either the average precision or the E-measure. The parameters are varied from one experimental run to another, and both average precisions and E-measures are obtained for each retrieval scheme for the purpose of comparing their effectiveness.

In order to provide a better understanding of how the changes in the parameter values on the Boolean operators affect the retrieval performance, Multiple Linear Regression techniques will be used in identifying possible prediction models for average precisions.

Similarity Computations

The following subsections present further details of similarity computations for each of the schemes that are being considered. An example of similarity computation for each scheme is carried out based upon the sample document and the sample query given in Figure 6.

P-norm Similarity

For the *OR* clause (with $p_{OR} = 1.5$), we have

$$SIM_{OR,1.5} = \left\{ \frac{.3542^{1.5} + .0910^{1.5} + .3838^{1.5} + 0^{1.5}}{1+1+1+1} \right\}^{1/1.5}$$

$$= \left\{ \frac{.2108 + .0274 + .2378 + 0}{4} \right\}^{1/1.5}$$

$$= \left\{ \frac{.4760}{4} \right\}^{1/1.5}$$

$$= .2419$$

For the *AND* clause (with $p_{AND} = 1.5$), we have

$$SIM_{AND,1.5} = 1 - \left\{ \frac{(1-.2419)^{1.5} + (1-.2890)^{1.5}}{1+1} \right\}^{1/1.5}$$

$$= 1 - \left\{ \frac{.6601 + .5995}{2} \right\}^{1/1.5}$$

$$= 1 - .6298^{1/1.5} = .2653$$

Hence, the P-norm similarity between the document, D^{18} and the query, Q^{35} for $p_{AND} = 1.5$ and $p_{OR} = 1.5$ is given by

$$SIM_{P\_NORM}(Q^{35}, D^{18}) = 0.2653$$

MMM Similarity

For the OR clause (with $Coeff_{OR} = 0.6$), we have

$$SIM_{OR,0.6} = .6 \times max\{.3542, .0910, .3838, 0\} + (1 - .6) \times min\{.3542, .0910, .3838, 0\}$$
$$= .6 \times .3838 + .4 \times 0 = .2303$$

For the AND clause (with $Coeff_{AND} = 0.5$), we have

$$SIM_{AND,0.5} = .5 \times min\{.2890, .2303\} + (1 - .5) \times max\{.2890, .2303\}$$
$$= .5 \times .2303 + .5 \times .2890 = .2596$$

Hence, the MMM similarity between the document, D^{18} and the query, Q^{35} for $Coeff_{AND} = 0.5$ and $Coeff_{OR} = 0.6$ is given by

$$SIM_{MMM}(Q^{35}, D^{18}) = 0.2596$$

Paice Similarity

For the OR clause (with $r_{OR} = 0.6$), we have

$$SIM_{OR,0.6} = \frac{(.6^0 \times .3838) + (.6^1 \times .3542) + (.6^2 \times .0910) + (.6^3 \times 0)}{.6^0 + .6^1 + .6^2 + .6^3}$$
$$= \frac{.3838 + .2125 + .0328 + 0}{2.176}$$
$$= \frac{.6291}{2.176} = .2891$$

For the AND clause (with $r_{AND} = 1.0$), we have

$$SIM_{AND,1.0} = \frac{(1^0 \times .2891) + (1^1 \times .2891)}{1^0 + 1^1}$$
$$= .2891$$

Hence, the PAICE similarity between the document, D^{18} and the query, Q^{35} for $r_{AND} = 1.0$ and $r_{OR} = 0.6$ is given by

$$SIM_{PAICE}(Q^{35}, D^{18}) = 0.2891$$

TIRS Similarity

For the given sample query in Figure 6, a disjunctive representation is

$$
\begin{aligned}
Q_{DNF}^{35} = \; & OR\,(\,AND(<\text{government, } 1.0>, <\text{information, } 1.0>)), \\
& (\,AND\,(<\text{government, } 1.0>, <\text{dissemination, } 1.0>)), \\
& (\,AND\,(<\text{government, } 1.0>, <\text{agencies, } 1.0>)), \\
& (\,AND\,(<\text{government, } 1.0>, <\text{projects, } 0.0>))
\end{aligned}
$$

Note that there are altogether five terms or literals in the given Boolean query. A min-term is a conjunction of the literals where each appears exactly once and is either complemented or uncomplemented. Thus, with five literals, the given Boolean query has up to 2^5 min-terms. In this experimental study, we first obtain the set of associated min-terms, and then select each one that satisfies the given query (i.e. find all min-terms for which the original given query evaluates to 'true'). Assuming that the attribute terms are stored in lexicographic order (agencies, dissemination, government, information, project) and that the terms in the query are the first few in the attribute list, the vectors corresponding to the set of associated min-terms obtained are as follows:

$$
\begin{aligned}
&<0.0, 0.0, 1.0, 1.0, 0.0, 0.0, ...>, \\
&<0.0, 1.0, 1.0, 0.0, 0.0, 0.0, ...>, \\
&<1.0, 0.0, 1.0, 0.0, 0.0, 0.0, ...>, \\
&<0.0, 0.0, 1.0, 0.0, 1.0, 0.0, ...>, \\
&<1.0, 1.0, 1.0, 0.0, 0.0, 0.0, ...>, \\
&<1.0, 0.0, 1.0, 1.0, 0.0, 0.0, ...>, \\
&<1.0, 0.0, 1.0, 0.0, 1.0, 0.0, ...>,
\end{aligned}
$$

$< 0.0, 1.0, 1.0, 1.0, 0.0, 0.0, ... >$,

$< 0.0, 1.0, 1.0, 0.0, 1.0, 0.0, ... >$,

$< 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, ... >$,

$< 1.0, 1.0, 1.0, 1.0, 0.0, 0.0, ... >$,

$< 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, ... >$,

$< 0.0, 1.0, 1.0, 1.0, 1.0, 0.0, ... >$,

$< 1.0, 0.0, 1.0, 1.0, 1.0, 0.0, ... >$,

$< 1.0, 1.0, 1.0, 1.0, 1.0, 0.0, ... >$

Each of the above associated min-terms forms a query point in the document space. One then assigns relevance balls to each of the query points and retrieves all the documents that are within these balls in ranked order of increasing distance. In this study, we take the relevance ball radius to be in all cases arbitrarily large, and we match each of the query points against the document using the INNER PRODUCT similarity function, returning the highest as the overall similarity between the query and the document.

The INNER PRODUCT similarity between two points $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ is defined as follows:

$$INNER\_PRODUCT(X, Y) = \sum_{i=1}^{n} x_i y_i$$

Using this method of computation, the overall highest similarity taken as the TIRS similarity between the document, D^{18} and the query, Q^{35} is

$$SIM_{TIRS}(Q^{35}, D^{18}) = 1.1180$$

A Note On Boolean Query DNF Conversion

To run the TIRS scheme, all the input queries must be in Disjunctive Normal Form (DNF). Note that only non-contradictory Boolean queries possess DNF forms. Converting a non-contradictory Boolean query into its most complete DNF form essentially comes down to building a set of minimal disjuncts or min-terms each of which satisfying it. But theoretically, such is in general an NP-complete problem. In practice, finding the set of associated min-terms each of which satisfying a non-contradictory Boolean query is a very tedious process, especially when the given Boolean query is somewhat lengthy. And, when conversions are to be performed on a large number of Boolean queries, a scheme for automating such a process is desirable. A procedure used in this study for finding the set of associated min-terms for each Boolean query is described below.

- Take all the terms in the given Boolean query and form the set of different combinations. For a query of n terms, there will be $(2^n - 1)$ combinations, not including the one with no query terms.

- Each combination is taken as a single document vector. Then, evaluate the original query against each of the vectors formed from the set of all combinations.

- The set of min-terms satisfying the original query includes all of the vectors for which the Boolean query evaluates to 'true'.

Consider a simple query $Q^{Bool} = A\ AND\ (B\ OR\ C)$. The possible combinations of the query terms are:

1. A
2. B
3. C
4. A, B
5. A, C
6. B, C
7 A, B, C

Taking each of the above combinations as a single document vector, we then proceed to evaluate the query Q^{Bool} against it. Here, the query evaluates to 'true' only for the set of vectors <A, B>, <A, C> and <A, B, C>. Thus, in this case, the query above has three associated min-terms, and logically it can be expressed as follows:

$$Q^{Bool} = AB + C = ABC + A\overline{B}C + AB\overline{C}.$$

Complexity of Computations

As we have seen, each of the above retrieval schemes takes a different approach to similarity computation. We now want to consider the computational overhead required in each of these schemes. In doing so, we first classify the computational steps, based on their relative costs, into the following four categories (beginning with the category of highest cost):

CAT 1: Exponentiation

CAT 2: Multiplication and Division

CAT 3: Addition and Subtraction

CAT 4: Comparison

Based on the computational steps we have performed on document 18 and query 35, we provide the following summary of overhead involved in each scheme. Note that initialization steps that need to be carried out only one time for each query prior to actually computing its similarity with respect to the each of the documents in the collection are not considered. For P-norm, we are using $p = 1.5$ and so have a good deal of expensive exponentiation computations. Clearly using $p = 1$ would lead to much lower overall cost. For the MMM scheme, there is a need for finding the *max* and *min* of a set of term weights each time an *AND* or *OR* clause is considered. With the recursive *MAXMIN* (*divide-and-conquer*) algorithm found in [AHO 74, p. 61], the number of comparisons required for a set of two elements is 1, and that required of a set of n elements, where n is a power of 2, is given by $3/2 \times n - 2$. Thus for a set of 4 elements, there will be a total $3/2 \times 4 - 2$ (i.e., 4) comparisons in finding the *max* and

min. For Paice, the term weights have to be sorted in ascending or descending order, depending on whether an *AND* clause or an *OR* is being considered. As indicated by [AHO 76], with an $n \log(n)$ sorting algorithm, we possibly require only 5 comparison steps for arranging a set of 4 real numbers in ascending or descending order. Also, in the Paice similarity computation, the coefficients for the terms in the sum of the numerator are obtained by incremental multiplications rather than by exponentiations. For TIRS, there are altogether 15 associated query points. Using INNER PRODUCT for similarity computation with each query point, we need altogether 43 multiplications and 28 additions. But, since all the query terms are unweighted, the multiplications can be omitted. Also, finding the maximum of the similarities computed for all query points takes 14 comparisons.

We summarize the computational costs as below, using a simple assignment of costs to the four categories of operations. The relative unit cost assumed for a computation of category 1 or 2 is 1.5, and that of a computation of category 3 or 4 is 1.0.

Computational Overheads Based on Document 18 and Query 35

		Retrieval Schemes			
Category	**Unit Cost**	**P-norm**	**MMM**	**Paice**	**TIRS**
1	1.5	8	0	0	0
2	1.5	2	4	6	0
3	1.0	7	2	4	28
4	1.0	0	5	5	14
Total Cost	-	22	13	18	42

From the total cost provided in the summary for each retrieval scheme, we clearly see that the TIRS scheme is more computationally intensive than any of the other

schemes. The MMM scheme is the most efficient scheme, followed by Paice, P-norm and finally TIRS.

It is, however, noted that when $p = 1$ is used, the exponentiation steps can be omitted. This will result in an overall cost of only 10 (i.e., with 2 divisions and 7 additions) for the P-norm scheme. Thus with such a setting for p, the P-norm scheme in fact turns out to be less costly than any of the other schemes.

Characteristics of Queries

The computational cost for the TIRS scheme is excessive when the DNF form of a query contains a large number of min-terms. Hence, queries with too many min-terms in the CACM collection are omitted for the TIRS experimental runs. When there is an abundance of min-terms in the query set, the TIRS runs have to be carried out in subdivided batches of queries, from which the final results are combined, and an evaluation of overall retrieval effectiveness is then made. Tables 1(a) and 1(b) show the distributions of query length and the corresponding number of min-terms in the CISI and CACM query sets, respectively.

All queries in CISI are used in all experimental runs. Of the 35 queries in CISI, 19 (greater than 50 percent) have more than 40 min-terms, and 13 (about 30 percent) have more than 100 min-terms. Of the 64 queries in CACM, only the 52 not marked with '*' are used for P-norm, MMM, and Paice runs. Only queries not marked with '*' or '@' (i.e., a total of 50 out of the 62 CACM queries) are used in TIRS runs. The queries marked wiht '@' are omitted because they contain terms not of the relevant concept types being considered in the TIRS runs. Of the 64 CACM queries, 9 (about 14 percent) have 40 or more min-terms. As can be seen from the tables, the distribution of the number of min-terms in the CACM query set is more skewed than that of the CISI. This may be explained by the fact that the CACM query set contains mostly 'homogeneous' queries, which are either strictly *AND* or strictly *OR*.

Qid	#Terms	#Min-terms	Qid	#Terms	#Min-terms
3	3	3	23	7	45
14	3	3	12	6	49
20	4	7	16	6	49
22	4	7	4	7	94
28	4	7	11	7	105
30	4	9	13	7	105
31	4	9	19	7	105
10	5	15	2	9	375
21	5	15	9	9	381
25	5	15	24	10	675
26	5	15	17	10	735
35	5	15	15	10	795
8	5	21	18	10	961
27	5	21	1	11	1023
29	5	25	32	11	1917
34	5	25	5	12	3255
6	6	45	7	12	3825
33	6	45			

Table 1(a). Distributions of #Terms and #Min-terms in CISI Query Set

Qid	#Terms	#Min-terms	Qid	#Terms	#Min-terms
3	3	1	36	4	3
10	2	1	52*	3	3
12	3	1	56*	3	3
13	3	1	42	5	5
14	2	1	16	4	7
15	2	1	39	4	7
17	2	1	55*	5	7
19	2	1	57@	4	7
20	2	1	1	4	9
24	3	1	5	5	9
26	3	1	11	4	9
27	4	1	29	5	13
28	3	1	48	4	13
30	2	1	49	5	13
34*	2	1	59	5	15
35*	2	1	54*	5	23
38	2	1	37	6	29
47*	3	1	4	6	31
50*	3	1	33	6	31
51*	2	1	43	6	31
61	2	1	44	6	31
62	2	1	7	8	35
63	2	1	32	7	39
64	1	1	18	6	43
8	2	1	40	7	63
2@	2	3	46*	7	63
6	5	3	60	9	153
9	3	3	21	8	193
22	3	3	58	8	234
23	3	3	45	11	1103
25	3	3	53*	14	16114
31	5	3	41*	15	24171

Table 1(b). Distributions of #Terms and #Min-terms in CACM Query Set

Chapter 4

Performance Results

Experimental Analysis on CISI Collection

P-norm Runs

Table 2 shows the average precision values obtained on CISI using the P-norm scheme for coefficients of the Boolean operators uniformly set at 1, 6, 12 and 50. Table 3, on the other hand, shows average precisions obtained for coefficients set between 1.0 and 4.0 with intervals of 0.25. (Note that the highest average precisions are enclosed in parentheses in these and later tables. The same will be done with results using other schemes. In tables showing E-measures, the best E-measures are also enclosed in parentheses.)

As shown by Table 2, the P-norm scheme does not perform well with large coefficients of AND and OR. The P-norm scheme tends to behave like the conventional strict Boolean scheme at large coefficients of *AND* and *OR*. The average precision of the standard Boolean scheme on the CISI is 0.1123, while the average precision for the P-norm scheme with both of the coefficients set to 50 is 0.1348.

The P-norm scheme becomes a version of the vector-processing scheme when both the coefficients of *AND* and *OR* are equal to 1. As shown in Table 3, the average precision obtained with $Coeff_{AND}$ and $Coeff_{OR}$ set at 1.0 is 0.1957.

The average precisions are somewhat less sensitive to the changes in the coefficient of *AND* than to changes in the coefficient of *OR*. Such phenomenon is also depicted in the surface plot shown in Figure 7(a). To further highlight this observation, the graphs in Figures 7(b) and 7(c) show how average precisions vary with the coefficient of *AND* at constant coefficients of *OR*, and how average precisions vary with the coefficient of *OR* at constant coefficients of *AND*, respectively.

The best average precision among all of these P-norms runs is 0.2008, and it occurs with both the $Coeff_{AND}$ and $Coeff_{OR}$ set at the level 1.50. The curves in Figure 7(b) do not show any large variation in average precision with respect to changes in the coefficient of *AND* at constant coefficients of *OR*. Except for the boundary case of $Coeff_{OR} = 1$, the curves show a slow decrease as the $Coeff_{AND}$ increases beyond the peak value location. However, the curves in Figure 7(b) show that beyond the peak value, the average precisions decrease rapidly with increases in the coefficient of *OR* at constant coefficients of *AND*.

It is noted that the P-norm scheme, at its best, shows an improvement of 79 percent over the standard Boolean scheme in terms of average precision.

A comparison of the best performance result with P-norm and those with other schemes under consideration will be made in a later section.

C_{AND}	C_{OR} 1	6	12	50
1	(.1957)	.1876	.1833	.1402
6	.1953	.1801	.1793	.1388
12	.1897	.1778	.1777	.1369
50	.1881	.1710	.1691	.1348

Table 2. Average Precision Values with P-norm Scheme on CISI for the set of Coefficients: 1, 6, 12 and 50.

Table 3. Average Precision Values With P-norm Scheme on CISI													
C_{AND}	C_{OR} 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.1957	.1972	.1988	.2004	.1975	.1959	.1937	.1916	.1911	.1904	.1899	.1894	.1886
1.25	.1963	.1970	.1997	.2001	.1982	.1947	.1924	.1923	.1919	.1908	.1895	.1885	.1878
1.50	.1968	.1977	(.2008)	.1991	.1965	.1944	.1924	.1922	.1917	.1909	.1901	.1891	.1884
1.75	.1968	.1990	.2002	.1979	.1976	.1944	.1930	.1922	.1914	.1907	.1902	.1895	.1891
2.00	.1971	.1983	.1999	.1991	.1975	.1942	.1932	.1919	.1912	.1905	.1906	.1899	.1895
2.25	.1966	.1983	.2001	.1989	.1983	.1950	.1922	.1917	.1915	.1910	.1904	.1897	.1896
2.50	.1970	.1977	.1990	.2003	.1979	.1941	.1925	.1919	.1914	.1910	.1907	.1895	.1890
2.75	.1986	.1983	.1993	.1992	.1972	.1942	.1925	.1921	.1910	.1909	.1899	.1889	.1888
3.00	.1974	.1989	.2000	.1997	.1966	.1931	.1919	.1917	.1914	.1905	.1892	.1883	.1883
3.25	.1967	.1998	.1997	.1990	.1962	.1931	.1923	.1912	.1909	.1898	.1891	.1886	.1878
3.50	.1976	.1995	.1991	.1986	.1958	.1937	.1918	.1908	.1908	.1895	.1881	.1878	.1871
3.75	.1980	.1985	.1990	.1988	.1963	.1925	.1911	.1908	.1903	.1888	.1878	.1868	.1866
4.00	.1993	.1981	.1995	.1986	.1949	.1922	.1906	.1898	.1891	.1882	.1872	.1870	.1855

Figure 7(a). The P-norm Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

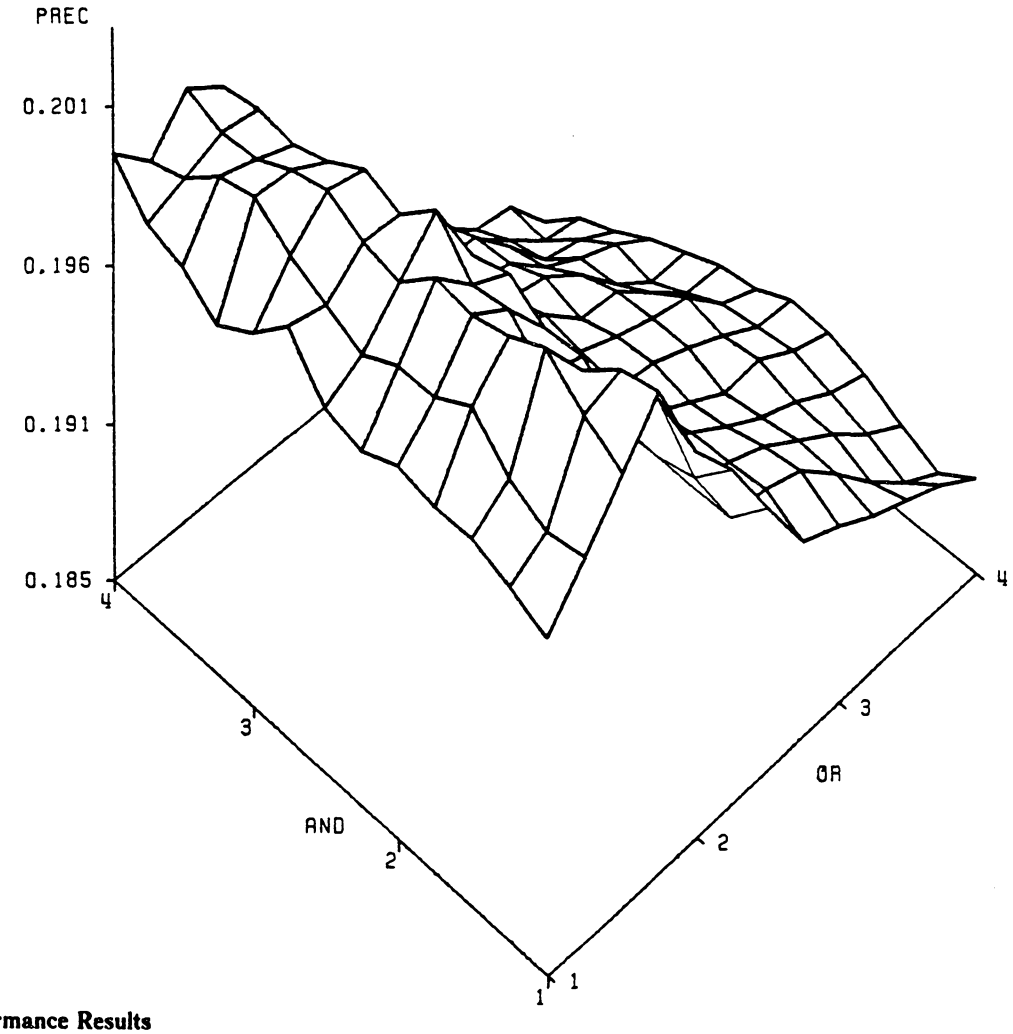


Figure 7(b). The P-norm Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

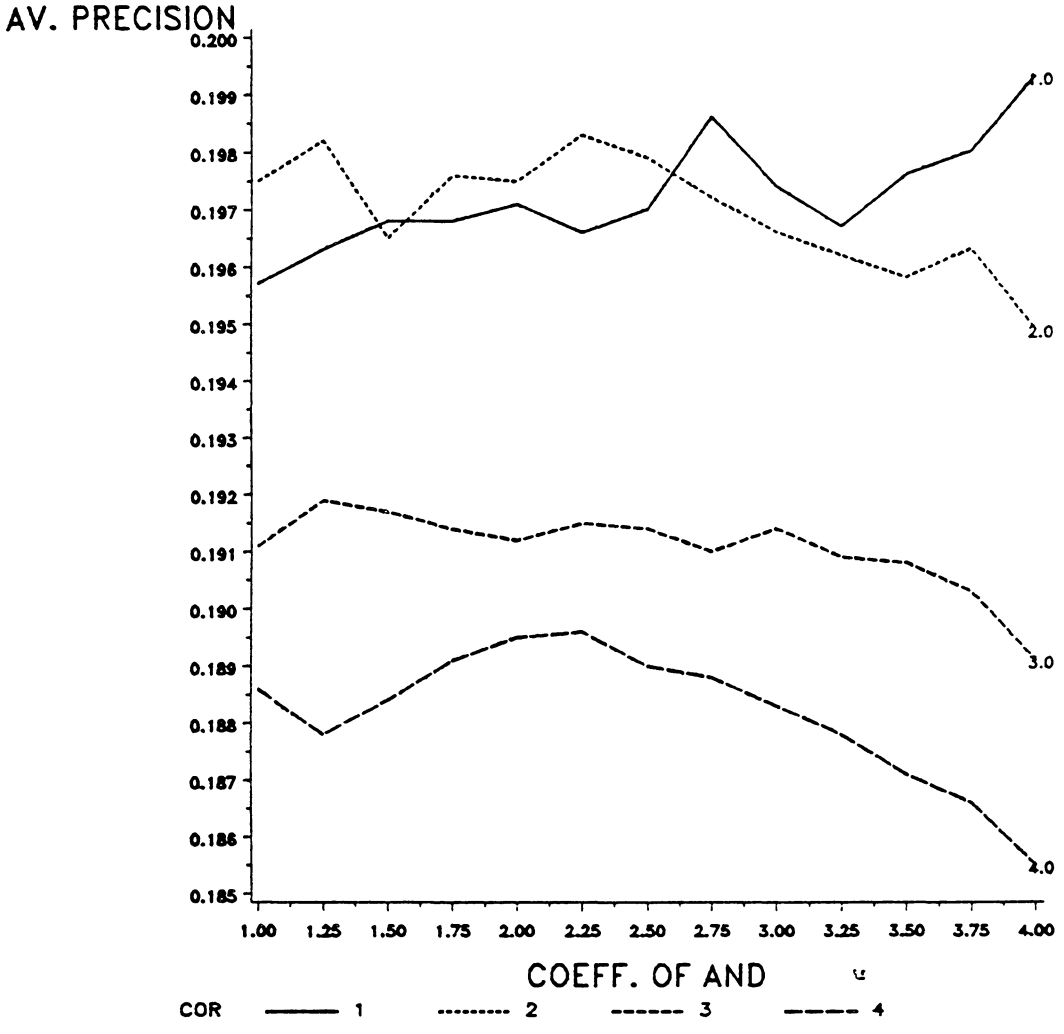
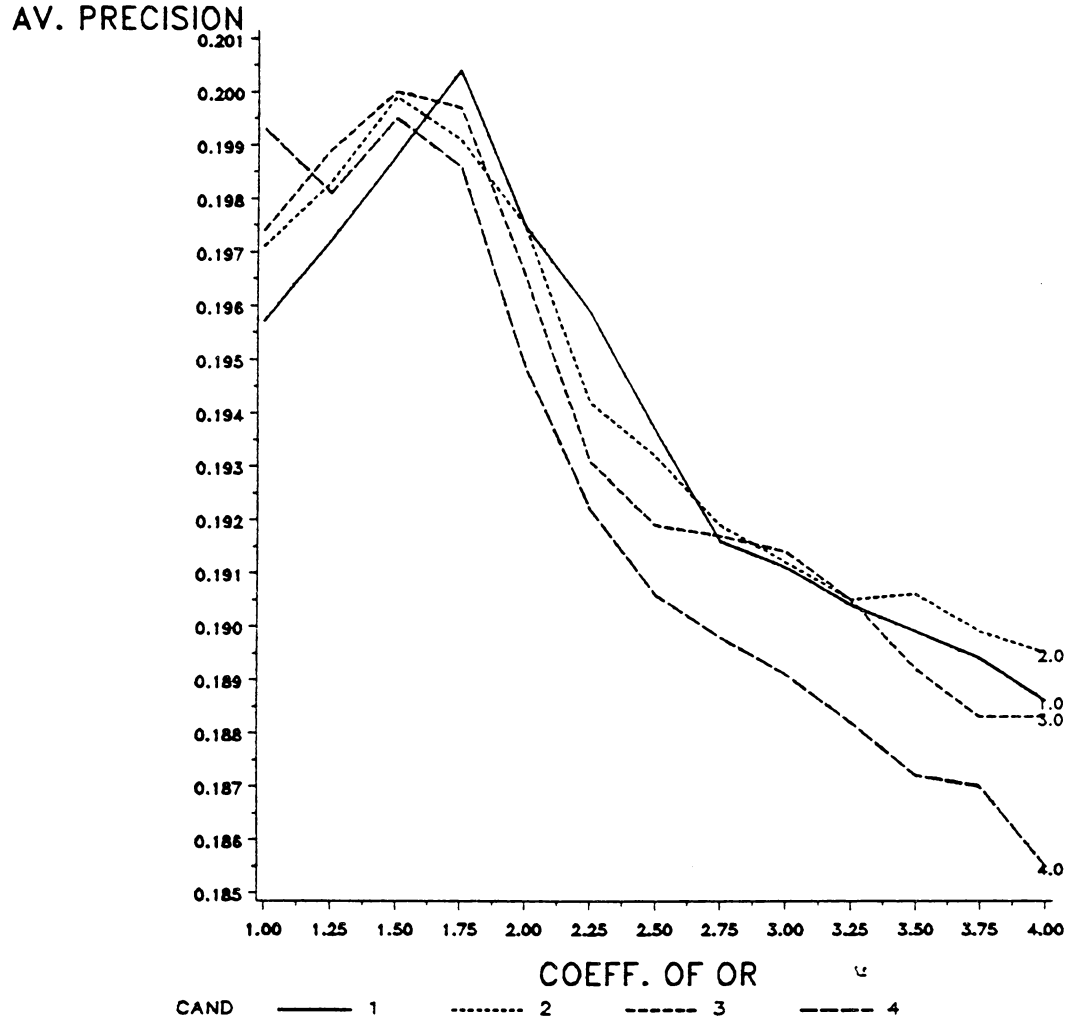


Figure 7(c). The P-norm Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$



Paice Runs

Table 4 shows the average precision values obtained on CISI using the Paice scheme for coefficients of the Boolean operators set between 0.0 and 1.0 with intervals of 0.1. For this scheme, the average precision seems to be affected to a much larger degree by changes in the coefficient of *OR* than by changes in the coefficient of *AND*, as can be seen from the surface plot and graphs in Figures 8(a), (b) and (c).

The peak average precision occurs at 0.1987 with $Coeff_{AND} = 1.0$, and $Coeff_{OR} = 0.6$. This peak average precision is a little lower than that of the P-norm runs. From the curves in Figure 8(c), it is noted that the average precision values increase with increases in the coefficient of *OR* at constant coefficients of *AND*, up to but not beyond the level of the coefficient of *OR* for which the peak average precision occurs. The curves in Figure 8(b) show that, at constant levels of $Coeff_{OR}$, the average precision values increase rapidly with increases in the $Coeff_{AND}$ for the range between 0 and 0.1, beyond which the growth of the average precisions seems more steady. This sudden rise of average precision values with increases in the $Coeff_{AND}$ at constant levels of $Coeff_{OR}$ may be explained by the fact that the Paice formulation for the *AND* operator emphasizes the more heavily weighted terms when the $Coeff_{AND}$ is at the high end of its range, as a result of its requiring the *AND* similarity computation be carried out in ascending order of term weights.

The Paice scheme effectively gives some form of weighted average of all query terms, as opposed to the classical fuzzy-set scheme which simply takes the *max* or the *min*. Thus, when one sets a coefficient of *AND* or *OR* to 1.0, one intends to give equal importance to each individual query term in the retrieval process. As shown by the Paice

runs, the coefficient of *AND* should be set at around 1.0 to achieve high average precision.

When a query is constructed by the user using *AND*, all terms included must serve equally well or play similar roles in describing his search intent. Otherwise, the user would have chosen to use *OR* instead. So, the fuzzy subsets associated with terms within an *AND*-query are often not needed to be discriminated against one another during retrieval, and therefore, setting the coefficient of *AND* to 1.0 is sensible. On the other hand, a user frequently includes in an *OR*-query terms which can serve only somewhat to describe his intents. Under such a circumstance, some terms in the user query may play more important roles than the others in describing the user's need; thus it would be wise that all fuzzy subsets associated with terms in an *OR* query be 'fairly' weighted in the retrieval process so as to achieve a high performance search.

For this study, the Paice scheme has been so implemented that when the $Coeff_{OR}$ is 0, only the *max* in the sum of the Paice formula is considered, and when the $Coeff_{AND}$ is 0, only the *min* is considered. Thus, at those coefficients, the Paice *OR* and *AND* are respectively the classical fuzzy *OR* and *AND*. The average precision obtained with those coefficients is 0.1291, as can be seen in Table 4.

Also, when both the $Coeff_{OR}$ and the $Coeff_{AND}$ are equal to 1, the Paice scheme behaves like a vector-processing scheme. The average precision obtained with those coefficients is 0.1957, which as we have expected turns out to be the same as the corresponding value for the P-norm. This average precision is lower than the overall peak value by a small margin. However, the average precision of the classical fuzzy-set scheme (0.1291), as compared with the overall peak value (0.1987) is exceeded by 0.0696. The Paice scheme, at its peak performance, has an improvement of 54 percent in terms

of average precision over the classical fuzzy-set scheme, and an improvement of 77 percent over the standard Boolean scheme.

Table 4. Average Precision Values with Paice Scheme on CISI											
C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.1291	.1260	.1255	.1264	.1269	.1275	.1296	.1290	.1320	.1345	.1387
0.1	.1725	.1719	.1725	.1735	.1742	.1774	.1807	.1802	.1822	.1846	.1842
0.2	.1755	.1775	.1783	.1787	.1802	.1828	.1829	.1867	.1869	.1914	.1930
0.3	.1753	.1783	.1810	.1822	.1830	.1865	.1881	.1893	.1897	.1896	.1899
0.4	.1764	.1798	.1825	.1819	.1852	.1883	.1905	.1905	.1903	.1917	.1937
0.5	.1772	.1798	.1837	.1852	.1868	.1896	.1918	.1932	.1925	.1948	.1937
0.6	.1792	.1825	.1865	.1869	.1874	.1913	.1936	.1928	.1923	.1956	.1948
0.7	.1794	.1842	.1869	.1879	.1879	.1924	.1933	.1933	.1949	.1957	.1956
0.8	.1804	.1845	.1875	.1891	.1901	.1951	.1954	.1966	.1958	.1962	.1969
0.9	.1796	.1851	.1882	.1897	.1916	.1971	.1980	.1976	.1955	.1965	.1962
1.0	.1792	.1846	.1880	.1919	.1940	.1976	(.1987)	.1972	.1957	.1954	.1957

Figure 8(a). The Paice Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

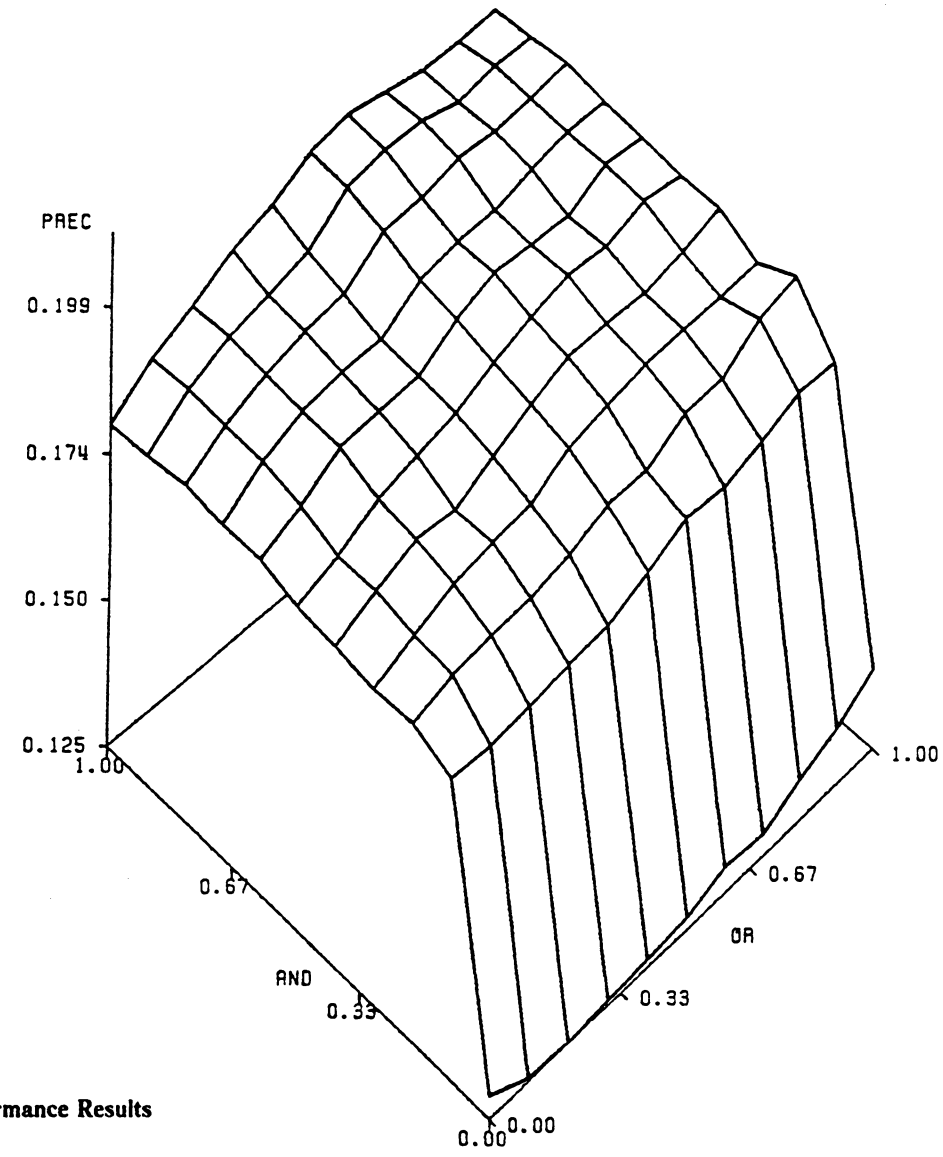


Figure 8(b). The Paice Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

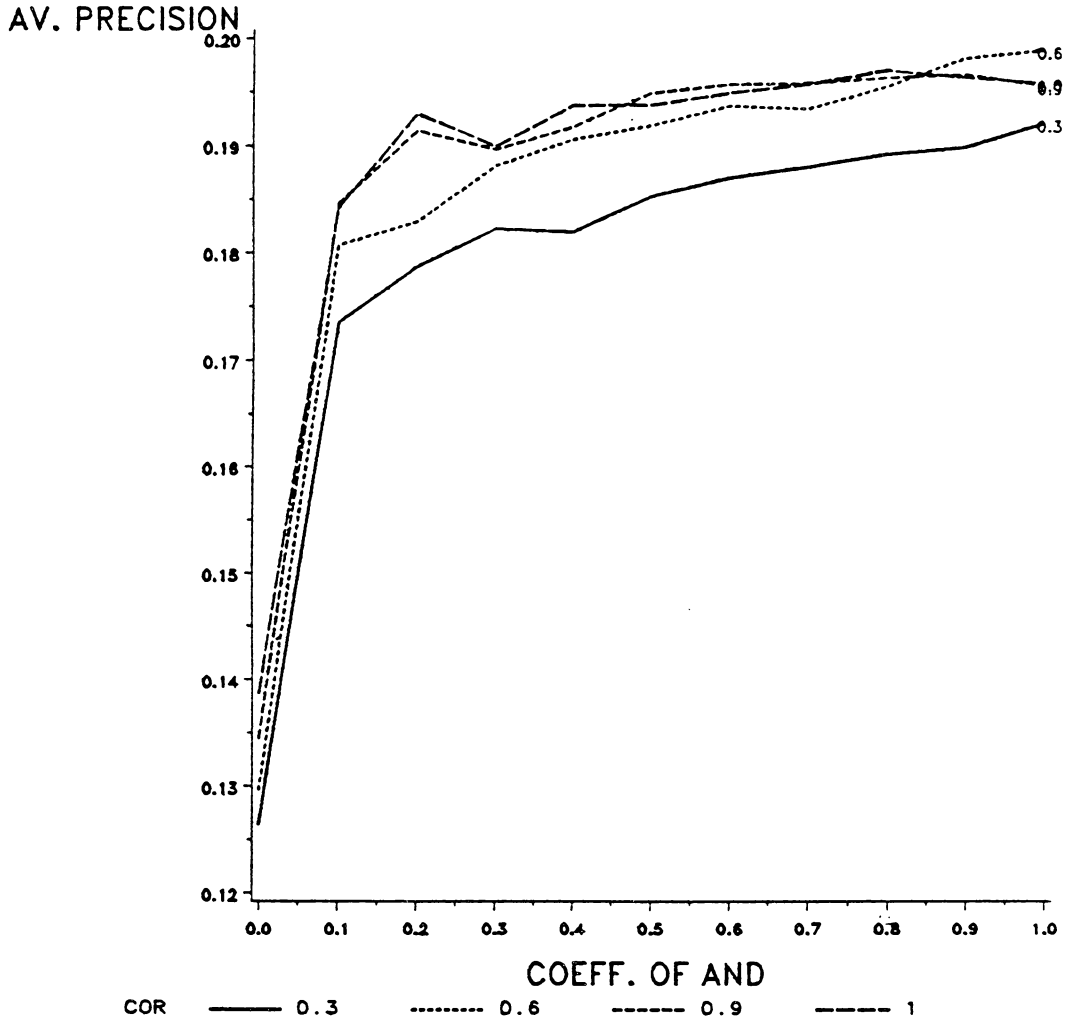
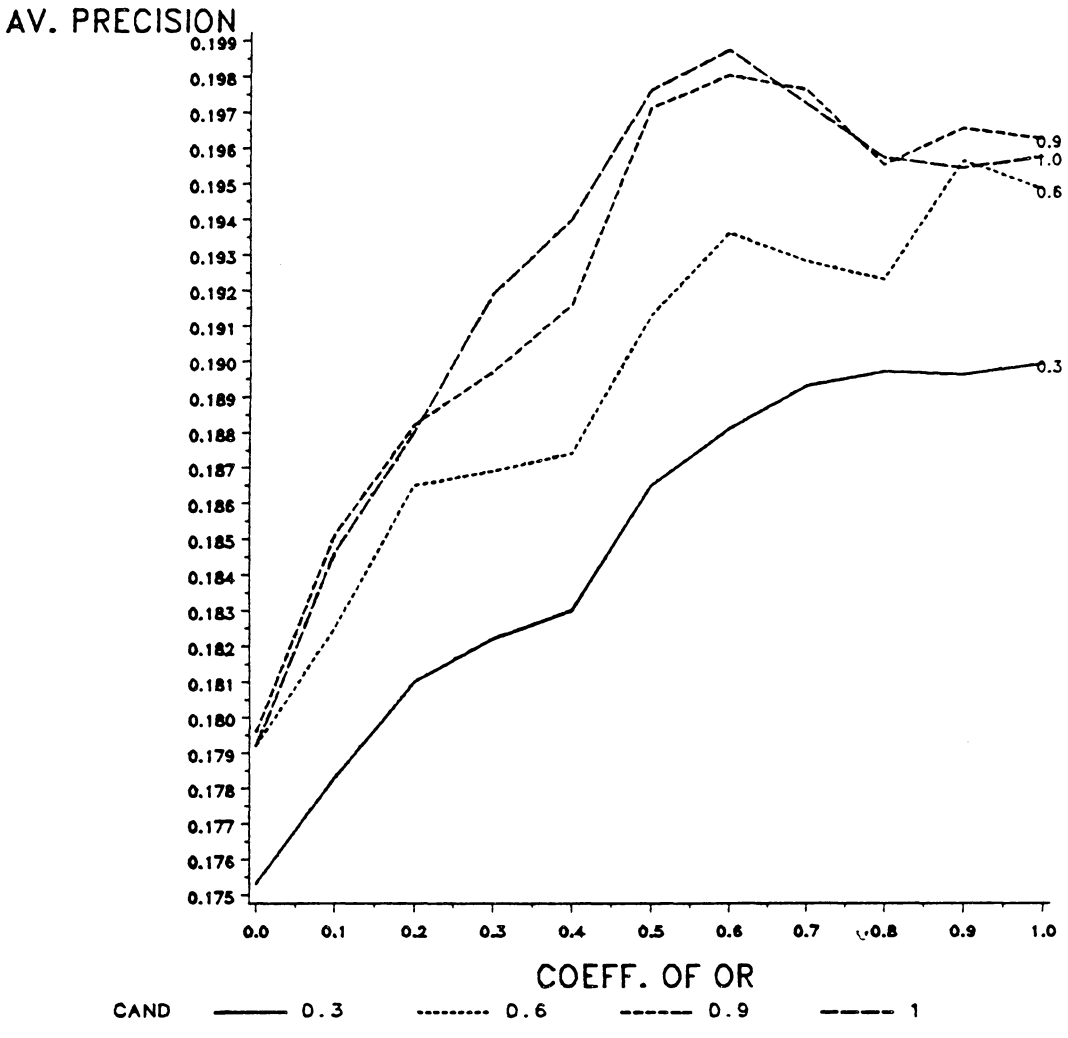


Figure 8(c). The Paice Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$



MMM Runs

Table 5 shows the average precision values obtained on CISI using the MMM scheme for coefficients of the Boolean operators set between 0.0 and 1.0 with intervals of 0.1. For this scheme, a similar set of graphs to those in previous sections is presented in Figures 9(a), (b) and (c). The peak average precision occurs at 0.1889 with $Coeff_{AND} = 0.5$ and $Coeff_{OR} = 0.6$. This value is lower than both that of the P-norm and the Paice runs.

It is observed that changes in the coefficient of *OR* do not seem to affect the average precision as much as the changes in the coefficient of *AND*. The curves in Figure 9(c) are relatively flat with respect to the whole range of values for the coefficient of *OR*, and they run parallel to one another for the various coefficients of *AND*. However, the curves in Figure 9(b), at constant coefficients of *OR*, rise with increasing values of the coefficient of *AND*. The rates of change of average precision with respect to the coefficient of *AND* are not the same at the various constant coefficients of *OR*. This suggests that there is an interaction effect between the two predictor variables.

Instead of defining *OR* using *max* alone and *AND* using *min* alone, the MMM scheme suggests a linear combination of the two for each case. As expected, higher average precision results were obtained with the coefficients of *AND* and *OR* set at levels 0.5 and above than otherwise. This supports the desirable situation we have suggested about the *OR* being set closer to *max* than *min*, and the *AND* being set closer to *min* than *max*, as previously outlined in our discussion on the MMM retrieval model.

As compared with the sets of curves and surface plot in the previous two retrieval schemes, the set of curves and surface plot in this MMM scheme are relatively smooth. However, it is to be cautioned that when both the coefficients of *AND* and *OR* approach 1.0, the MMM scheme is essentially that of the classical fuzzy-set case. Though the average precision values depicted by the curves in Figure 9(c) witness a uniform increase up to the 1.0 level set for the coefficient of *OR*, the average precision values shown by the curves in Figure 9(b) fall steeply when the coefficient of *AND* approaches the value 1.0. This clearly indicates that the classical fuzzy-set scheme, in general, does not perform as well as MMM. The average precision with both of the coefficients set at 1.0 is 0.1291, which is exactly the same as that from Paice. The MMM scheme shows an improvement of 46 percent in terms of average precision over the classical fuzzy-set scheme, and also an improvement of 68 percent over the standard Boolean scheme.

With the $Coeff_{OR}$ and $Coeff_{AND}$ set to 0, the average precision obtained is 0.1138. Under such a setting, we have essentially switched the classical fuzzy *OR* and *AND* operations. Two other peculiar settings for the coefficients that are worth considering are (a) $Coeff_{OR} = 1$ and $Coeff_{AND} = 0$, and (b) $Coeff_{OR} = 0$ and $Coeff_{AND} = 1$. The first case changes the *AND* to *OR*, and has an average precision of 0.1449; the second changes the *OR* to *AND* , and has an average precision of 0.0370. It seems that changing *OR* to *AND* has resulted in not retrieving anything useful at all. On the other hand, changing *AND* to *OR* may have rendered the search queries too broad, with the average precision jeopardized but far less seriously than in the previous case.

Table 5. Average Precision Values With MMM Scheme on CISI											
C_{AND}	C_{OR} 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.1138	.1741	.1736	.1752	.1745	.1729	.1725	.1691	.1622	.1518	.1449
0.1	.1139	.1818	.1821	.1835	.1809	.1790	.1792	.1778	.1723	.1606	.1527
0.2	.1137	.1827	.1822	.1850	.1831	.1818	.1839	.1820	.1750	.1659	.1575
0.3	.1135	.1834	.1837	.1872	.1867	.1856	.1852	.1850	.1789	.1727	.1631
0.4	.1135	.1833	.1849	.1879	.1874	.1868	.1858	.1854	.1822	.1774	.1711
0.5	.1135	.1846	.1862	.1871	.1874	.1886	(.1889)	.1881	.1863	.1838	.1800
0.6	.1139	.1857	.1852	.1861	.1874	.1872	.1863	.1868	.1861	.1841	.1813
0.7	.1139	.1835	.1826	.1842	.1837	.1837	.1848	.1828	.1819	.1804	.1780
0.8	.1139	.1806	.1817	.1799	.1781	.1796	.1792	.1796	.1793	.1791	.1774
0.9	.1139	.1784	.1731	.1729	.1725	.1757	.1746	.1729	.1728	.1738	.1743
1.0	.0370	.1264	.1263	.1262	.1264	.1293	.1267	.1266	.1256	.1265	.1291

Figure 9(a). The MMM Scheme on CISI: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

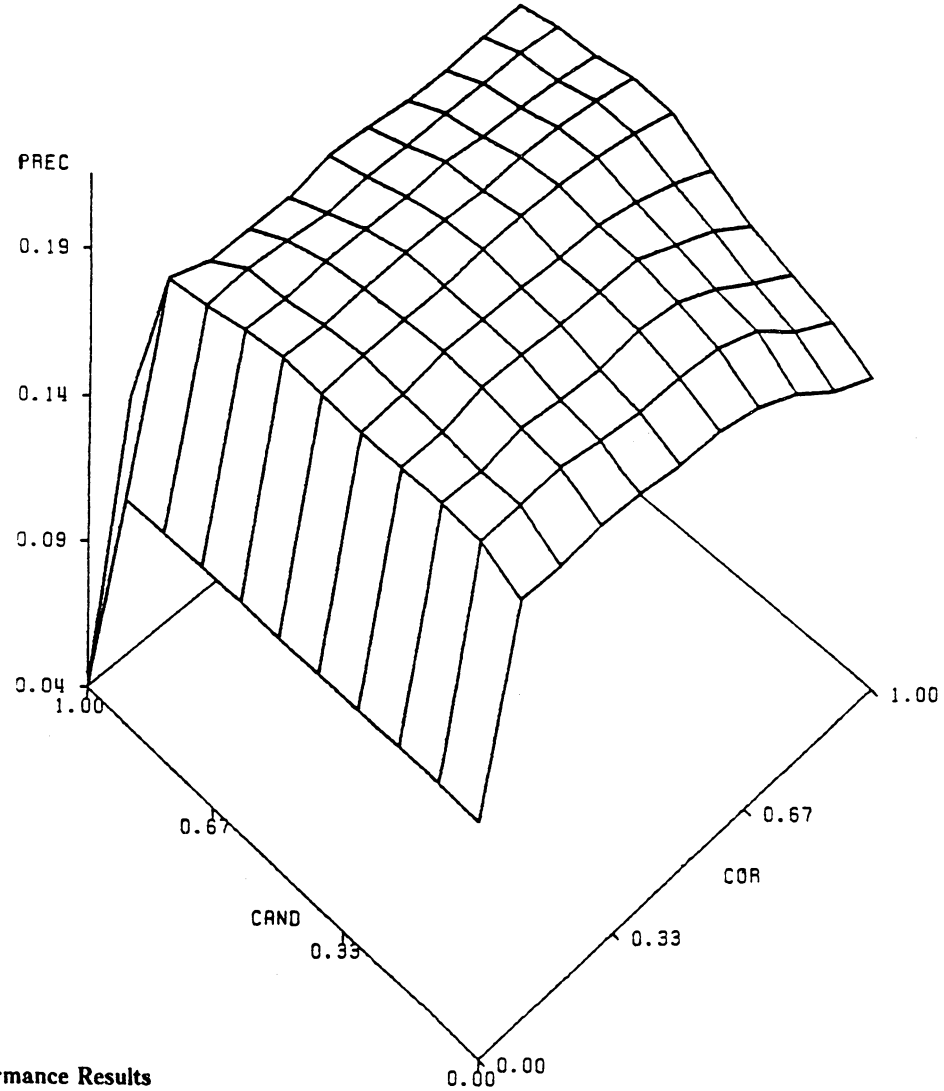


Figure 9(b). The MMM Scheme on CISI:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

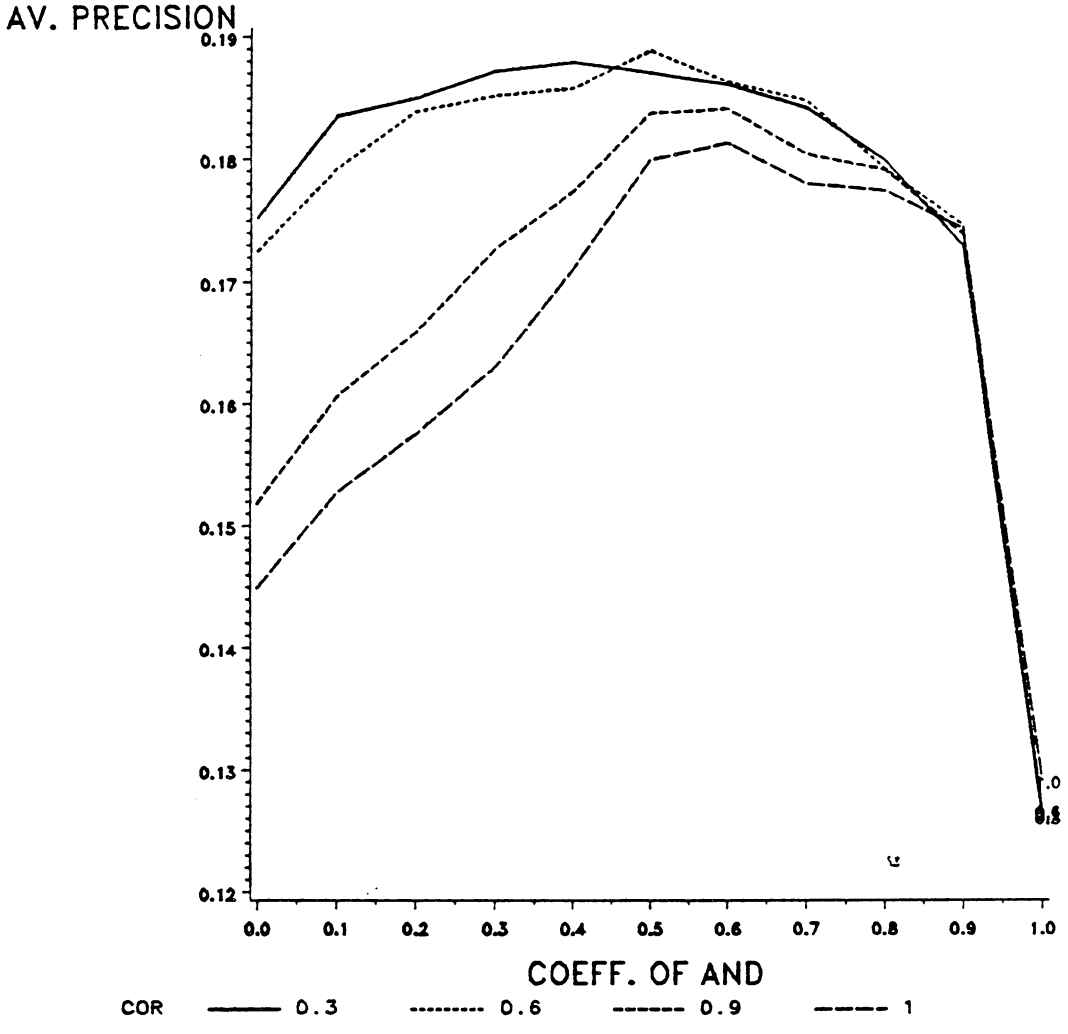
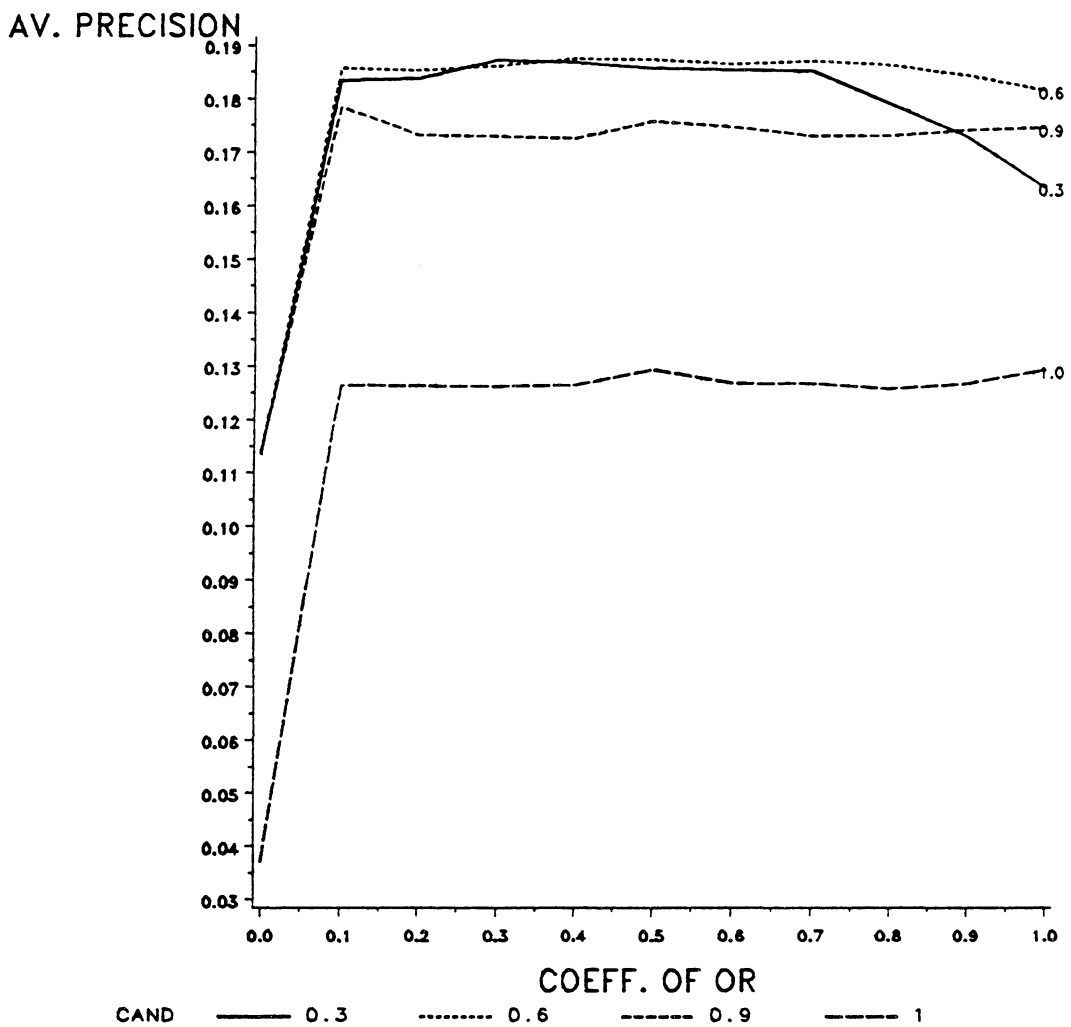


Figure 9(c). The MMM Scheme on CISI:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$



TIRS Runs

In this study, the idea of 'relevance ball radius' is not exactly made full use of in the way implied by the original scheme of Cater & Kraft, since we are interested in ranking rather than retrieving at arbitrary thresholds. The relevance ball radius does not actually form an independent variable in our TIRS scheme, since in all cases, it is taken to be arbitrarily large. The average precision obtained from the TIRS experimental run is 0.1645. This average precision result seems to be lower than any of the peak average precisions from the experimental runs of aforementioned schemes.

The TIRS average precision is a 46 percent improvement over that of the standard Boolean scheme, and a 27 percent improvement over that of the classical fuzzy-set scheme.

TIRS runs were particularly time consuming, and had to be done in partial batches of queries. For the same set of queries, the TIRS scheme involves more computational overhead than the other schemes. This is so because the TIRS scheme does not just consider a query as given, but rather the set of associated min-terms, from which many query points are obtained for matching.

Looking back at Table 1(a), we see that more than 50 percent of CISI queries have 40 or above min-terms, and almost 25 percent have 675 or above min-terms. While each of the other schemes has to match each of the 35 CISI queries, the TIRS scheme has to match each of the individual min-terms for each of the queries.

Prediction Models

In order to understand the effects of varying strictness of *AND* and *OR* that result from manipulating the parameters available in each of the retrieval schemes, multiple linear regression techniques [MYERS 86] were used to obtain prediction models of retrieval performance. The SAS stepwise regression procedures (namely, FORWARD SELECTION and MAXR) [SAS 85, pp. 763-774] enabled us to find a set of candidate prediction models. But, there remained the difficult task of selecting the model that best predicts retrieval performances. A model that is too simple may suffer from biased regression coefficients and biased predictions. On the other hand, a model that is overly complicated can result in large variances, both in the regression coefficients and in the predictions. Thus, one always attempts to balance the tradeoffs between the two.

The s^2 and R^2 values are generally used to ascertain the quality of a fitted model. The s^2 represents the mean square error of the model, while the R^2 represents the proportion of variation in the response data that is explained by the fitted model. Another criterion that we can use in selecting models is the Mallows $C(p)$ statistic. It is essentially a measure of bias plus variance. In a normal procedure, p which is the number of variables plus the intercept, should be close to the function value of $C(p)$ in order to judge that the model contains no biased estimates.

Table 6a shows a summary of the results for the SAS forward selection procedure performed on CISI. The table presents the sequential steps in which the predictor variables are entered, and the corresponding partial F and p statistics. The F-statistic outlined for each step may be viewed as a ratio that expresses the variance explained by

the variable entered divided by variance due to the model established up to and including that variable.

With P-norm, C_{OR} is entered first, followed by $C_{AND}C_{OR}$ and C_{AND}. With Paice, C_{AND} is entered first, followed by its square and its cube; C_{OR} and C_{OR}^2 are not entered until the last two steps. With MMM, C_{AND}^3 is entered first, followed by C_{AND}^2 and then $C_{AND}C_{OR}$. Also, with this scheme, the variable $C_{AND}C_{OR}$, whose F and p statistics are respectively 16.4131 and 0.0001, is entered early in the selection process; there seems to be significant interaction effects between C_{AND} and C_{OR}. The partial R^2 for the entry of $C_{AND}C_{OR}$ is 0.0814. The $C_{AND}C_{OR}$ interaction variable is entered second for the case of P-norm, with $F = 20.668$, $p = 0.0001$, and partial $R^2 = 0.0164$. Thus, adding the variable $C_{AND}C_{OR}$ in the third step in the forward selection of prediction variables for average precision with MMM contributes more towards the overall model R^2 than in the second step of selection with P-norm.

Table 6b shows the best 3-, 4-, and 5-variable models obtained for each retrieval scheme using the SAS MAXR procedure. In terms of R^2, all the models for P-norm and Paice are well-fitted. But, for MMM, only the 5-variable model ($R^2 = 0.7778$) is reasonably better fitted than the 3- and 4-variable models. In terms of R^2 and $C(p)$ combined, all the 5-variable models are well-fitted, and the 4-variable model for Paice is also acceptable.

To further obtain even better fittings with multiple regression, we proceed to omit results at the boundary values of coefficients which seem to constitute the set of undesirable outliers. For P-norm, we take only the range of coefficients between 1.5 and 4.0; for Paice, we take the range between 0.2 and 1.0 and for MMM, the range between

0.1 and 0.9. The stepwise regression results and predictions models for best 3, 4 and 5 variables are given in Tables 7a and 7b.

The prediction models in Table 7b are superior to the corresponding ones in Table 6b, especially for the case of MMM. All the best 3-variable models are well-fitted. For P-norm, the best 3-variable model has $R^2 = 0.9655$ and $C_p = 22.6088$; for Paice, $R^2 = 0.9032$ and $C_p = 31.6090$, and for MMM, $R^2 = 0.8717$ and $C_p = 19.8613$.

Table 6a. Stepwise Regression Results on CISI Collection
Summary of SAS Forward Selection Procedure
Dependent Variable : Average Precision

P-NORM SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{OR}	0.8523	0.8523	198.661	963.5694	0.0001
2	$C_{AND}C_{OR}$	0.0164	0.8686	160.396	20.6678	0.0001
3	C_{AND}	0.0023	0.8709	156.849	2.8795	0.0916
4	C^3_{AND}	0.0065	0.8774	142.869	8.6818	0.0037
5	C^2_{OR}	0.0007	0.8781	143.233	0.8885	0.3473
6	C^3_{OR}	0.0565	0.9346	6.012	140.0753	0.0001

PAICE SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{AND}	0.4346	0.4346	392.936	91.4874	0.0001
2	C^2_{AND}	0.2285	0.6632	188.818	80.0541	0.0001
3	C^3_{AND}	0.1377	0.8009	66.574	80.9498	0.0001
4	C_{OR}	0.0709	0.8718	4.639	64.1354	0.0001
5	C^2_{OR}	0.0027	0.8745	4.160	2.5194	0.1152

MMM SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C^3_{AND}	0.1802	0.1802	336.003	26.1489	0.0001
2	C^2_{AND}	0.1579	0.3380	250.758	28.1470	0.0001
3	$C_{AND}C_{OR}$	0.0814	0.4195	207.760	16.4131	0.0001
4	C^3_{OR}	0.0615	0.4810	175.755	13.7562	0.0003
5	C_{OR}	0.1349	0.6159	103.229	40.3828	0.0001
6	C^2_{OR}	0.1748	0.7907	8.633	95.2328	0.0001
7	C_{AND}	0.0048	0.7955	8.000	2.6327	0.1075

Table 6b. Selected Prediction Models of Average Precisions On CISI Best 3-, 4- and 5-variable Models Obtained by MAXR

P-NORM SCHEME

$$\hat{Prec} = 0.2022 + 0.0006C_{AND} - 0.0030C_{OR} - 0.0004C_{AND}C_{OR}$$

$s^2 = 0.000002$
$R^2 = 0.870895$
$C_p = 156.8495$

$$\hat{Prec} = 0.2006 + 0.0017C_{AND} - 0.0030C_{OR} - 0.0004C_{AND}C_{OR} - 0.0001C_{AND}^3$$

$s^2 = 0.000002$
$R^2 = 0.877386$
$C_p = 142.8692$

$$\hat{Prec} = 0.1831 + 0.0006C_{AND} + 0.0247C_{OR} - 0.0004C_{AND}C_{OR} - 0.0119C_{OR}^2 + 0.0015C_{OR}^3$$

$s^2 = 0.000001$
$R^2 = 0.928109$
$C_p = 19.99215$

PAICE SCHEME

$$\hat{Prec} = 0.1388 + 0.2961C_{AND} - 0.5201C_{AND}^2 + 0.2818C_{AND}^3$$

$s^2 = 0.000067$
$R^2 = 0.800910$
$C_p = 66.57433$

$$\hat{Prec} = 0.1312 + 0.2961C_{AND} + 0.0151C_{OR} - 0.5201C_{AND}^2 + 0.2818C_{AND}^3$$

$s^2 = 0.000043$
$R^2 = 0.871794$
$C_p = 4.638687$

$$\hat{Prec} = 0.1296 + 0.2961C_{AND} + 0.0258C_{OR} - 0.5201C_{AND}^2 - 0.0107C_{OR}^2 + 0.2818C_{AND}^3$$

$s^2 = 0.000043$
$R^2 = 0.874542$
$C_p = 4.159648$

MMM SCHEME

$$\hat{Prec} = 0.1760 + 0.0986C_{AND}C_{OR} - 0.0783C_{AND}^3 - 0.0392C_{OR}^3$$

$s^2 = 0.000390$
$R^2 = 0.448802$
$C_p = 191.5618$

$$\hat{Prec} = 0.1302 + 0.1724C_{OR} + 0.1890C_{OR}^2 - 0.1503C_{OR}^2 - 0.2239C_{AND}^3$$

$s^2 = 0.000238$
$R^2 = 0.666208$
$C_p = 73.43508$

$$\hat{Prec} = 0.1169 + 0.3834C_{OR} + 0.1890C_{AND}^2 - 0.7036C_{OR}^2 - 0.2239C_{AND}^3 + 0.3689C_{OR}^3$$

$s^2 = 0.000160$
$R^2 = 0.777767$
$C_p = 13.47937$

Table 7a. Stepwise Regression Results on CISI Collection
(With Boundary Values Omitted)
Summary of SAS Forward Selection Procedure
Dependent Variable : Average Precision

P-NORM SCHEME
(with independent variables set within range 1.5-4.0)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{OR}	0.8764	0.8764	369.416	843.4623	0.0001
2	C^2_{OR}	0.0558	0.9322	151.906	97.0460	0.0001
3	C^3_{AND}	0.0334	0.9655	22.609	113.2801	0.0001
4	C_{AND}	0.0017	0.9672	18.034	5.9108	0.0166
5	$C_{AND}C_{OR}$	0.0032	0.9704	6.780	2.7404	0.1006
6	C^3_{OR}	0.0007	0.9711	8.000	0.7804	0.3789
7	C^2_{AND}	0.0002	0.9713	8.000	0.7804	0.3789

PAICE SCHEME
(with independent variables set within range 0.2-1.0)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	$C_{AND}C_{OR}$	0.7277	0.7277	217.217	211.0717	0.0001
2	C_{AND}	0.0155	0.7432	202.444	4.7157	0.0329
3	C_{OR}	0.1438	0.8869	49.135	97.9143	0.0001
4	C^3_{OR}	0.0340	0.9209	14.420	32.6664	0.0001
5	C^2_{AND}	0.0114	0.9323	4.127	12.6078	0.0007

MMM SCHEME
(with independent variables set within range 0.1-0.9)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C^3_{AND}	0.5741	0.5741	252.187	118.6277	0.0001
2	C^2_{AND}	0.2701	0.8442	39.717	150.8182	0.0001
3	C_{AND}	0.0275	0.8717	19.861	18.4528	0.0001
4	C^3_{OR}	0.0144	0.8861	10.416	10.7598	0.0015
5	$C_{AND}C_{OR}$	0.0100	0.8962	4.445	8.1214	0.0055

Table 7b. Selected Prediction Models of Average Precisions On CISI (With Boundary Values Omitted) Best 3-, 4- and 5-variable Models Obtained by MAXR

P-NORM SCHEME
(with independent variables set within range 1.5-4.0)

$$\hat{Prec} = 0.2185 - 0.0141C_{OR} + 0.0017C_{OR}^2 - 0.00004C_{AND}^3$$

$s^2 = 0.000001$
$R^2 = 0.965530$
$C_p = 22.60875$

$$\hat{Prec} = 0.2157 + 0.0025C_{AND} - 0.0141C_{OR} - 0.0006C_{AND}^2 + 0.0017C_{OR}^2$$

$s^2 = 0.000001$
$R^2 = 0.967346$
$C_p = 17.46445$

$$\hat{Prec} = 0.2130 + 0.0035C_{AND} - 0.0131C_{OR} - 0.0004C_{AND}C_{OR} - 0.0006C_{AND}^2 + 0.0017C_{OR}^2$$

$s^2 = 0.000001$
$R^2 = 0.970528$
$C_p = 6.946225$

PAICE SCHEME
(with independent variables set within range 0.2-1.0)

$$\hat{Prec} = 0.1720 + 0.0130C_{AND} + 0.0230C_{OR} - 0.0087C_{OR}^3$$

$s^2 = 0.000003$
$R^2 = 0.903167$
$C_p = 31.60897$

$$\hat{Prec} = 0.1684 + 0.0189C_{AND} + 0.0290C_{OR} - 0.0100C_{AND}C_{OR} - 0.0087C_{OR}^3$$

$s^2 = 0.000002$
$R^2 = 0.920930$
$C_p = 14.41978$

$$\hat{Prec} = 0.1657 + 0.0299C_{AND} + 0.0290C_{OR} - 0.0100C_{AND}C_{OR} - 0.0091C_{AND}^2 - 0.0087C_{OR}^3$$

$s^2 = 0.000002$
$R^2 = 0.932309$
$C_p = 4.126878$

MMM SCHEME
(with independent variables set within range 0.1-0.9)

$$\hat{Prec} = 0.1895 - 0.1317C_{AND} + 0.4394C_{AND}^2 - 0.3636C_{AND}^3$$

$s^2 = 0.000040$
$R^2 = 0.871724$
$C_p = 19.86132$

$$\hat{Prec} = 0.1914 - 0.1317C_{AND} + 0.4394C_{AND}^2 - 0.3636C_{AND}^3 - 0.0086C_{OR}^3$$

$s^2 = 0.000036$
$R^2 = 0.886137$
$C_p = 10.41597$

$$\hat{Prec} = 0.1937 - 0.1413C_{OR} + 0.0191C_{AND}C_{OR} + 0.4394C_{AND}^2 - 0.3636C_{AND}^3 - 0.0190C_{OR}^3$$

$s^2 = 0.000033$
$R^2 = 0.896175$
$C_p = 4.444911$

Discussion

Scheme	Best Precision	Rank	Best E-measure	Rank
P-NORM	.2008	1	.7940	1
PAICE	.1987	2	.7945	2
MMM	.1889	3	.8060	3
TIRS	.1645	4	.8331	4

Table 8. Relative Ranks of Schemes by Average Precision and E-measure on CISI.

Table 8 shows the summary of best performance measures of all the schemes and their relative ranks. On this CISI collection, the P-norm scheme is superior to all the others in terms of average precision, and it is followed by Paice, MMM and lastly TIRS. While the best average precisions of P-norm, Paice and MMM are close to one another, that of the TIRS scheme seems a little lower.

In each of our experimental runs, the SMART retrieval program also produces a 'top-ranked' file containing information about the top ten documents retrieved for each query. Table 9 shows the information we obtained for query 35 with the four different retrieval schemes. The first column shows the document ID. Associated with each document are its relative rank, a flag indicating if it is relevant or not, and its similarity with the query in question. Consider the first row from the table for P-norm. Document 18 is ranked highest in the retrieved set, is judged to be relevant with respect to query 35, and has a similarity of 0.2653. This similarity result can serve to validate against the 'hand-computed' value obtained in the section on 'Similarity Computations'

which appeared earlier. The same kind of validations can also be carried out for the other schemes.

As seen in Table 9, document 18 is ranked consistently the highest in the retrieved set by all the schemes except MMM. With P-norm and Paice, 7 out of the ten top-ranked retrieved documents are relevant, and with MMM, 6 are relevant. However, with TIRS, only 5 out of the ten top-ranked retrieved documents are relevant. It is also interesting to note that while P-norm, Paice and MMM each ranks document 385 the second of the ten top-ranked, TIRS ranks it the third. TIRS ranks document 375 the second, while P-norm and Paice each rank it third and MMM ranks it sixth. The MMM scheme ranks document 286 first, while the P-norm and Paice schemes rank it fourth and fifth respectively. Surprisingly for TIRS, document 286 is not ranked at all within the ten top-ranked retrieved documents, even though it is a relevant document. Documents 385, 375 and 286 are shown in Figures 10, 11 and 12 respectively. Though document 375 is not a relevant document, it seems to be quite relevant.

(a) P-NORM SCHEME $C_{and} = 1.50$, $C_{or} = 1.50$

document ID	rank	relevant	similarity
18	1	1	0.265303
385	2	1	0.253013
375	3	0	0.232333
286	4	1	0.231278
1145	5	1	0.227007
130	6	1	0.206087
402	7	1	0.201781
403	8	1	0.197633
1245	9	0	0.196567
421	10	0	0.192213

(b) PAICE SCHEME $C_{and} = 1.0$, $C_{or} = 0.6$

document ID	rank	relevant	similarity
18	1	1	0.289074
385	2	1	0.274840
375	3	0	0.248752
1145	4	1	0.247774
286	5	1	0.245275
130	6	1	0.224691
402	7	1	0.216784
1245	8	0	0.213371
403	9	1	0.210157
421	10	0	0.207812

(c) MMM SCHEME $C_{and} = 0.5$, $C_{or} = 0.6$

document ID	rank	relevant	similarity
286	1	1	0.270544
385	2	1	0.259676
18	3	1	0.259676
1145	4	1	0.259676
130	5	1	0.241611
375	6	0	0.238232
1245	7	0	0.225072
1449	8	0	0.222579
1031	9	0	0.217841
402	10	1	0.216784

(d) TIRS SCHEME

document ID	rank	relevant	similarity
18	1	1	1.118024
375	2	0	1.109683
385	3	1	0.978384
1362	4	0	0.945677
1207	5	1	0.841416
1145	6	1	0.782064
1415	7	0	0.727307
910	8	0	0.722732
130	9	1	0.709542
947	10	0	0.687792

Table 9. Ten Top-ranked Documents Retrieved with Query 35 on CISI

The document, D^{385} is as follows (with query terms highlighted for reader's convenience):

.I 385
.T
Evaluative Research Principles and Practice in Public Service and Social Action Programs
.A
Suchman, E.A.
.W
In these days of large ***government*** programs intended to reduce poverty, develop communities, prevent delinquency and crime, control disease, and reconstruct cities, the predominant rhetoric is that of planning, pilot ***projects***, experimental and demonstration programs - and evaluation. Those who seek to select for support the more promising plans and ***projects*** submitted to funding ***agencies*** have become habituated to the ritualistic inclusion in the proposal of a final section on Evaluation. In most cases this section consists of sometimes grandiose but usually vague statements of intent and procedure for assessing the impact of the proposed action. In some cases there is an elegant, highly academic, and impractical scheme worked out in meticulous detail by an obviously talented research consultant. In a few treasured instances there is a well-considered, realistic, and workmanlike plan for getting some fairly reliable answers to the questions of what worked and why.

Figure 10. Document 385 from CISI collection

The document, D^{375} is as follows:

.I 375
.T
Encyclopedia of ***Information*** Systems and Services
.A
Kruzas, A.T.
.W
The processing and transfer of ***information*** is an important activity of many thousands of libraries, research institutes, educational institutions, professional and trade associations, non-profit organizations, publishing houses, ***government agencies***, and others. All of these groups are already listed in a variety of existing directories. This publication, on the other hand, has selected from the above groups, those organizations and services which are principally concerned with storage, retrieval, and ***dissemination*** of ***information***, and in addition, are innovative, experimental, or non-conventional. A major emphasis is on computerization, micrographics, networks, advanced reference services, ***information*** centers, and data banks.
The Encyclopedia of ***Information*** Systems and Services includes descriptions of the following types of services and facilities:

Information Centers
Computerized Systems and Services
Networks and Cooperative Programs
Data Banks
Documentation Centers
Information Storage and Retrieval Systems
Micrographic Systems and Services
Research Centers and Projects
Clearinghouses and Referral Centerss
Consulting and Planning Organizations and Services
Information Offices
Industrial Research ***Information*** Centers
Professional Associations
Specialized Library Reference Services

Figure 11. Document 375 from CISI collection

The document, D^{286} is as follows:

.I 286
.T
User's Reaction to Microfiche A Preliminary Study
.A
Lewis, Ralph W.
.W
Recent emphasis placed on the use of microfiche by large ***government agencies*** has increased the pressure in libraries supporting ***government*** research to make greater use of microfiche.. Negative and apathetic user attitudes, expressed by researchers, indicate that expanded efforts to overcome resistance if the great potential of microfiche is to be realized.. Efforts in microphotography, expended on technical achievement in the past, should be directed toward understanding the user and his needs to discover why he avoids microforms and how to overcome his resistance to them..

Figure 12. Document 286 from CISI collection

The E-measures

Since some argue that the users must be allowed to attach different relative importance to precision and recall, we will next consider E-measures obtained in addition to average precision values from our experimental runs. The E-measure is a performance measure that is computed with respect to a parameter (β) set by the user. The β-level represents the importance the user attaches to recall as opposed to precision. As mentioned before, the lower the E-measure, the better is the retrieval performance. The E-measure is based on a set oriented view of retrieval, where for each query the recall and precision of a retrieved set are determined. Since our comparison is of ranking, we have chosen to obtain values for sets representing the best 30 documents.

Tables 11, 12, and 13 present the complete sets of E-measures from the P-norm, PAICE and MMM runs on CISI. We have summarized the best E-measures at β-levels 0.5, 1.0 and 2.0 along with those of TIRS in Table 10. As we can see from the table, the best E-measure at β-level 1.0 ranks the four different retrieval schemes in the same order as the best average precision, namely P-norm, PAICE, MMM and TIRS.

Scheme	Best E-measure $\beta = 0.5$	Rank	Best E-measure $\beta = 1.0$	Rank	Best E-measure $\beta = 2.0$	Rank
P-NORM	.7616	1	.7940	1	.8060	2
PAICE	.7623	2	.7945	2	.8056	1
MMM	.7725	3	.8060	3	.8210	3
TIRS	.8065	4	.8331	4	.8408	4

Table 10. Relative Ranks of Schemes by E-measure at β-levels 0.5, 1.0 and 2.0 on CISI

Again, at β-level 0.5, the E-measure ranks the retrieval schemes in the same order. However, at β-level 2.0, the best E-measures seem to indicate marginally better retrieval performance with Paice than P-norm. Thus, when one wants to emphasize recall as twice as important as precision, one might settle on the Paice scheme rather than P-norm. But, such an inference is simply not valid until we have further seen similar results from experimental runs on other collections and in any case, the difference between the P-norm and Paice schemes is clearly not of any real significance.

Table 11(a). E-measures with P-norm Scheme on CISI at $\beta = 0.5$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.7678	.7654	.7653	.7661	.7704	.7723	.7716	.7724	.7715	.7743	.7768	.7774	.7789
1.25	.7674	.7660	.7668	.7658	.7730	.7740	.7741	.7707	.7697	.7756	.7763	.7783	.7782
1.50	.7676	.7638	.7659	.7659	.7732	.7738	.7741	.7724	.7679	.7729	.7755	.7757	.7780
1.75	.7676	.7622	.7674	.7697	.7748	.7729	.7732	.7715	.7713	.7711	.7738	.7767	.7775
2.00	.7656	.7622	.7682	.7684	.7730	.7732	.7728	.7715	.7704	.7734	.7711	.7741	.7742
2.25	.7656	.7622	.7700	.7673	.7727	.7731	.7702	.7686	.7675	.7687	.7732	.7750	.7757
2.50	.7622	.7638	.7678	.7705	.7701	.7683	.7675	.7689	.7655	.7689	.7700	.7728	.7742
2.75	.7649	.7655	.7669	.7690	.7690	.7680	.7673	.7666	.7673	.7720	.7719	.7741	.7762
3.00	.7649	.7666	.7663	.7669	.7681	.7682	.7663	.7686	.7651	.7713	.7729	.7741	.7763
3.25	.7668	.7665	.7636	.7662	.7675	.7679	.7669	.7669	.7683	.7725	.7726	.7732	.7768
3.50	.7671	.7650	.7622	.7647	.7666	.7675	.7658	.7677	.7663	.7689	.7716	.7737	.7757
3.75	.7676	.7638	.7627	.7644	.7668	.7675	.7673	.7676	.7662	.7699	.7721	.7733	.7757
4.00	.7676	.7623	(.7616)	.7644	.7665	.7661	.7665	.7679	.7661	.7698	.7697	.7709	.7742

Table 11(b). E-measures with Paice Scheme on CISI at $\beta = 0.5$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8188	.8161	.8164	.8122	.8120	.8100	.8051	.8049	.8014	.8033	.8046
0.1	.8059	.7966	.8012	.7974	.7892	.7875	.7851	.7852	.7818	.7799	.7810
0.2	.7966	.7893	.7910	.7925	.7859	.7828	.7788	.7747	.7765	.7727	.7708
0.3	.7933	.7833	.7817	.7808	.7786	.7760	.7749	.7724	.7714	.7708	.7667
0.4	.7862	.7761	.7743	.7751	.7728	.7721	.7705	.7690	.7667	.7694	.7667
0.5	.7879	.7774	.7760	.7712	.7679	.7642	.7661	.7663	.7636	.7628	.7644
0.6	.7862	.7747	.7732	.7679	.7709	.7687	.7649	.7643	.7651	.7661	(.7623)
0.7	.7839	.7781	.7738	.7714	.7717	.7702	.7671	.7664	.7673	.7647	.7647
0.8	.7848	.7760	.7741	.7737	.7750	.7762	.7686	.7678	.7667	.7656	.7634
0.9	.7857	.7787	.7709	.7709	.7729	.7742	.7719	.7678	.7688	.7672	.7661
1.0	.7903	.7803	.7753	.7723	.7730	.7719	.7689	.7691	.7696	.7696	.7678

Table 11(c). E-measures with MMM Scheme on CISI at $\beta = 0.5$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8508	.8161	.8177	.8108	.8055	.7996	.7976	.7992	.8098	.8210	.8385
0.1	.8516	.8103	.8082	.8022	.7940	.7883	.7849	.7916	.7965	.8097	.8215
0.2	.8502	.8063	.8000	.7886	.7803	.7827	.7807	.7896	.7949	.8065	.8159
0.3	.8458	.8036	.7945	.7854	.7829	.7823	.7824	.7811	.7902	.7979	.8076
0.4	.8450	.7995	.7942	.7866	.7815	(.7725)	.7817	.7834	.7822	.7912	.7995
0.5	.8415	.7975	.7940	.7855	.7783	.7784	.7770	.7825	.7796	.7855	.7895
0.6	.8421	.7961	.7928	.7818	.7773	.7792	.7793	.7775	.7803	.7804	.7842
0.7	.8421	.8013	.7925	.7850	.7813	.7783	.7777	.7768	.7806	.7825	.7881
0.8	.8421	.7979	.7912	.7811	.7821	.7831	.7824	.7880	.7919	.7875	.7936
0.9	.8421	.7986	.7915	.7924	.7957	.7941	.7967	.7972	.8043	.8035	.8060
1.0	.9794	.8149	.8155	.8149	.8155	.8137	.8135	.8158	.8179	.8137	.8188

Table 12(a). E-measures with P-norm Scheme on CISI at $\beta = 1.0$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.7994	.7971	.7969	.7979	.8022	.8039	.8032	.8037	.8029	.8050	.8076	.8074	.8085
1.25	.7991	.7976	.7980	.7971	.8045	.8054	.8054	.8022	.8013	.8059	.8068	.8081	.8080
1.50	.7994	.7958	.7973	.7972	.8048	.8051	.8054	.8038	.7995	.8034	.8060	.8058	.8077
1.75	.7994	.7941	.7984	.8008	.8060	.8042	.8046	.8030	.8028	.8016	.8045	.8070	.8078
2.00	.7979	.7941	.7991	.8000	.8043	.8046	.8040	.8030	.8018	.8042	.8026	.8054	.8054
2.25	.7979	.7941	.8006	.7987	.8037	.8044	.8020	.8006	.7995	.8003	.8050	.8059	.8065
2.50	.7947	.7955	.7989	.8022	.8017	.8007	.8001	.8011	.7979	.8007	.8017	.8040	.8051
2.75	.7966	.7970	.7984	.8009	.8011	.8001	.7998	.7992	.7995	.8037	.8036	.8056	.8071
3.00	.7966	.7981	.7980	.7995	.8001	.8005	.7984	.8002	.7966	.8032	.8048	.8056	.8074
3.25	.7987	.7985	.7955	.7988	.7997	.7998	.7987	.7979	.7996	.8040	.8043	.8046	.8078
3.50	.7991	.7973	.7944	.7975	.7988	.7997	.7973	.7986	.7976	.8001	.8032	.8049	.8064
3.75	.7992	.7960	.7952	.7970	.7992	.7997	.7985	.7983	.7974	.8013	.8035	.8043	.8064
4.00	.7992	.7946	(.7940)	.7970	.7987	.7981	.7978	.7991	.7973	.8012	.8011	.8019	.8048

Table 12(b). E-measures with Paice Scheme on CISI at $\beta = 1.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8422	.8402	.8406	.8374	.8381	.8364	.8323	.8319	.8291	.8308	.8318
0.1	.8310	.8230	.8273	.8242	.8174	.8165	.8145	.8151	.8118	.8111	.8124
0.2	.8231	.8173	.8187	.8205	.8149	.8124	.8097	.8057	.8077	.8051	.8032
0.3	.8211	.8117	.8104	.8097	.8086	.8068	.8069	.8038	.8026	.8023	.7984
0.4	.8147	.8054	.8045	.8056	.8037	.8038	.8022	.8008	.7989	.8003	.7985
0.5	.8164	.8070	.8071	.8033	.7999	.7963	.7980	.7976	.7955	(.7945)	.7965
0.6	.8157	.8052	.8047	.7994	.8016	.7997	.7965	.7956	.7963	.7983	.7950
0.7	.8137	.8095	.8052	.8019	.8022	.8011	.7978	.7973	.7981	.7967	.7971
0.8	.8147	.8063	.8040	.8038	.8049	.8062	.7991	.7985	.7980	.7975	.7954
0.9	.8150	.8092	.8006	.8007	.8028	.8044	.8024	.7994	.8002	.7992	.7982
1.0	.8194	.8108	.8048	.8018	.8030	.8025	.7992	.8003	.8013	.8011	.7994

Table 12(c). E-measures with MMM Scheme on CISI at $\beta = 1.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8709	.8431	.8452	.8386	.8334	.8286	.8261	.8278	.8372	.8482	.8639
0.1	.8715	.8371	.8354	.8301	.8228	.8175	.8141	.8210	.8254	.8369	.8488
0.2	.8705	.8343	.8287	.8186	.8117	.8142	.8113	.8199	.8236	.8344	.8438
0.3	.8676	.8330	.8250	.8168	.8145	.8136	.8139	.8117	.8202	.8270	.8368
0.4	.8669	.8303	.8261	.8179	.8129	(.8060)	.8128	.8150	.8132	.8219	.8303
0.5	.8646	.8278	.8247	.8158	.8098	.8095	.8077	.8123	.8098	.8153	.8188
0.6	.8650	.8264	.8237	.8131	.8089	.8095	.8094	.8088	.8112	.8114	.8142
0.7	.8650	.8307	.8225	.8151	.8117	.8092	.8089	.8081	.8108	.8118	.8167
0.8	.8650	.8273	.8207	.8112	.8118	.8123	.8117	.8165	.8194	.8154	.8207
0.9	.8650	.8263	.8204	.8214	.8245	.8223	.8249	.8245	.8298	.8293	.8312
1.0	.9843	.8408	.8412	.8408	.8412	.8395	.8392	.8406	.8415	.8381	.8422

Table 13(a). E-measures with P-norm Scheme on CISI at $\beta = 2.0$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.8117	.8092	.8086	.8104	.8154	.8169	.8162	.8162	.8155	.8172	.8205	.8198	.8208
1.25	.8114	.8096	.8095	.8087	.8174	.8181	.8181	.8148	.8140	.8179	.8194	.8205	.8204
1.50	.8117	.8081	.8089	.8088	.8177	.8178	.8181	.8163	.8123	.8155	.8187	.8182	.8199
1.75	.8117	(.8060)	.8099	.8122	.8186	.8169	.8174	.8158	.8156	.8138	.8173	.8196	.8212
2.00	.8105	(.8060)	.8104	.8116	.8171	.8174	.8166	.8158	.8145	.8167	.8159	.8192	.8193
2.25	.8105	(.8060)	.8118	.8102	.8159	.8171	.8150	.8136	.8125	.8131	.8192	.8193	.8198
2.50	.8069	.8072	.8102	.8144	.8141	.8138	.8135	.8142	.8108	.8139	.8148	.8171	.8180
2.75	.8085	.8085	.8100	.8132	.8137	.8131	.8130	.8125	.8123	.8169	.8168	.8188	.8200
3.00	.8085	.8099	.8097	.8124	.8126	.8136	.8110	.8124	.8080	.8165	.8182	.8188	.8208
3.25	.8110	.8105	.8071	.8117	.8123	.8123	.8112	.8094	.8116	.8171	.8175	.8177	.8211
3.50	.8114	.8093	.8062	.8104	.8115	.8123	.8090	.8100	.8094	.8124	.8164	.8179	.8191
3.75	.8113	.8080	.8073	.8098	.8119	.8123	.8100	.8095	.8093	.8138	.8166	.8171	.8191
4.00	.8113	.8066	(.8060)	.8098	.8113	.8106	.8093	.8113	.8092	.8137	.8136	.8142	.8169

Table 13(b). E-measures with Paice Scheme on CISI at $\beta = 2.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8509	.8492	.8496	.8468	.8479	.8459	.8423	.8420	.8394	.8410	.8418
0.1	.8405	.8333	.8373	.8346	.8282	.8275	.8260	.8268	.8230	.8235	.8252
0.2	.8332	.8284	.8296	.8315	.8261	.8239	.8219	.8175	.8198	.8177	.8155
0.3	.8326	.8229	.8217	.8207	.8204	.8190	.8199	.8161	.8142	.8140	.8095
0.4	.8261	.8166	.8167	.8178	.8161	.8169	.8147	.8135	.8111	.8109	.8096
0.5	.8277	.8189	.8206	.8168	.8130	.8089	.8106	.8094	.8074	(.8056)	.8084
0.6	.8282	.8177	.8180	.8120	.8133	.8116	.8084	.8072	.8079	.8109	.8072
0.7	.8263	.8232	.8186	.8136	.8139	.8130	.8092	.8088	.8095	.8091	.8097
0.8	.8277	.8192	.8158	.8155	.8165	.8178	.8104	.8103	.8100	.8098	.8075
0.9	.8279	.8230	.8121	.8122	.8145	.8163	.8143	.8118	.8123	.8116	.8106
1.0	.8320	.8245	.8166	.8132	.8148	.8147	.8104	.8126	.8141	.8133	.8117

Table 13(c). E-measures with MMM Scheme on CISI at $\beta = 2.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8821	.8564	.8590	.8524	.8470	.8426	.8400	.8418	.8506	.8616	.8761
0.1	.8826	.8497	.8484	.8430	.8362	.8310	.8279	.8352	.8391	.8490	.8614
0.2	.8818	.8475	.8423	.8326	.8265	.8291	.8259	.8346	.8368	.8469	.8568
0.3	.8796	.8473	.8400	.8321	.8299	.8291	.8295	.8266	.8346	.8407	.8511
0.4	.8791	.8458	.8422	.8334	.8287	.8233	.8281	.8308	.8287	.8378	.8461
0.5	.8774	.8412	.8383	.8291	.8239	.8234	(.8210)	.8252	.8225	.8283	.8316
0.6	.8777	.8399	.8375	.8271	.8230	.8220	.8220	.8221	.8249	.8252	.8269
0.7	.8777	.8437	.8356	.8281	.8248	.8221	.8220	.8214	.8236	.8236	.8280
0.8	.8777	.8398	.8330	.8226	.8234	.8238	.8231	.8275	.8300	.8263	.8311
0.9	.8777	.8373	.8316	.8326	.8363	.8332	.8359	.8350	.8394	.8391	.8407
1.0	.9871	.8504	.8507	.8504	.8507	.8491	.8488	.8498	.8500	.8472	.8509

Experimental Analysis on CACM and INSPEC Collections

The same set of experimental runs were performed on the CACM collection as on CISI. However, experimental runs on the INSPEC collection were carried out with the coefficients of AND and OR set at larger intervals and only within the range of peak retrieval performances observable from runs on the previous two collections. Since TIRS did not do well on both the CISI and CACM collections as compared to any of the other schemes and since the INSPEC collection has more documents and longer queries, we did not think it was worthwhile for us to run TIRS on the INSPEC collection. On INSPEC, we have also made no experimental runs in this study with P-norm using any particularly large coefficients of *AND* and *OR*.

CACM

Collected Results

Table 14 shows the P-norm runs on the CACM collection for coefficients set at 1, 6, 12 and 50, while Table 15 shows runs with coefficients set between 1.0 and 4.0 with intervals of 0.25. Again, it is clear that the P-norm scheme does not generally perform well when large coefficients are used. At $Coeff_{AND}$ and $Coeff_{OR}$ equal 50, the average precision obtained with the P-norm scheme is 0.2580, while the average precision obtained for the standard Boolean scheme is 0.1577. However, the overall best average

precision for the P-norm runs is 0.3249; it occurs at $C_{AND} = 1.00$ and $C_{OR} = 1.25$. The P-norm scheme, at its best, shows an improvement of 106 percent in terms of average precision over the standard Boolean scheme.

Table 16 presents the Paice runs on the CACM collection for $Coeff_{AND}$ and $Coeff_{OR}$ set between 0 and 1 with intervals of 0.1. The overall best average precision for the Paice runs, occurring at $C_{AND} = 1.0$ and $C_{OR} = 0.7$, is 0.3215, which is a little lower than that for P-norm. In terms of average precision, the Paice scheme has an improvement of 104 percent over the standard Boolean scheme. Also, both the P-norm and Paice schemes show an improvement of about 85 percent in retrieval effectiness (as indicated by their best average precisions) over the classical fuzzy-set scheme which has an average precision of 0.1745.

Table 17 presents the MMM runs on the CACM collection for $Coeff_{AND}$ and $Coeff_{OR}$ set between 0 and 1 with intervals of 0.1. The overall best average precision for MMM runs, occurring at $C_{AND} = 0.9$, and $C_{OR} = 0.4$, is 0.3300. This peak average precision is higher than that for both the P-norm runs and the Paice runs. The MMM scheme has improvement of 109 percent over the standard Boolean scheme and 89 percent over the classical fuzzy-set scheme in terms of average precision.

For TIRS, the average precision obtained is 0.2804. As with CISI, the average precision for TIRS on this collection is lower than the best average precisions of P-norm, Paice and MMM. In terms of average precision, the TIRS scheme has an improvement of 78 percent over the standard Boolean scheme and also an improvement of 61 percent over the classical fuzzy-set scheme.

It is noted that similar observations for the boundary cases of the various retrieval schemes on the CISI collection can be made for CACM and INSPEC. As on the CISI collection, analysis of the same kind with surface plots and graphs showing changes in average precisions with respect to changes in $Coeff_{AND}$ at constant levels of $Coeff_{OR}$, and vice versa, can be carried out with the average precision results. The surface plots and related graphs for various schemes on CACM are given in Figures 13, 14 and 15.

As compared to the corresponding surface plot and related curves on CISI, those found in this section for the P-norm scheme on CACM show a little more variation in average precision with respect to both the coefficients of *AND* and *OR*. On the CACM collection, the surface plot for the Paice scheme is virtually identical to that on CISI, while the related curves are close to their counterparts on CISI. Similarly, the surface plot for the MMM scheme is just as smooth as that on CISI, and the related curves show the same behavior as their counterparts on CISI.

C_{AND}	C_{OR} 1	6	12	50
1	(.3122)	.3102	.3093	.3025
6	.2977	.2965	.2975	.2862
12	.2975	.2943	.2932	.2835
50	.2850	.2733	.2706	.2580

Table 14. Average Precision Values with P-norm Scheme on CACM for set of Coefficients: 1, 6, 12 and 50.

Table 15. Average Precision Values with P-norm Scheme on CACM													
C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.3122	(.3249)	.3201	.3187	.3172	.3170	.3176	.3184	.3178	.3177	.3168	.3141	.3136
1.25	.3206	.3217	.3186	.3175	.3162	.3161	.3170	.3172	.3163	.3138	.3138	.3132	.3127
1.50	.3181	.3186	.3171	.3148	.3148	.3148	.3143	.3141	.3127	.3126	.3124	.3117	.3095
1.75	.3171	.3170	.3172	.3147	.3149	.3124	.3112	.3111	.3116	.3108	.3110	.3098	.3089
2.00	.3188	.3178	.3169	.3154	.3138	.3125	.3133	.3123	.3121	.3115	.3116	.3113	.3108
2.25	.3191	.3161	.3160	.3143	.3150	.3128	.3121	.3129	.3127	.3123	.3120	.3111	.3108
2.50	.3167	.3167	.3152	.3148	.3140	.3135	.3133	.3130	.3129	.3129	.3129	.3118	.3106
2.75	.3142	.3144	.3134	.3126	.3123	.3108	.3104	.3112	.3103	.3099	.3097	.3091	.3086
3.00	.3115	.3130	.3107	.3091	.3088	.3091	.3092	.3094	.3093	.3090	.3083	.3073	.3067
3.25	.3097	.3096	.3092	.3076	.3079	.3077	.3078	.3082	.3079	.3078	.3078	.3069	.3066
3.50	.3102	.3102	.3095	.3081	.3080	.3085	.3083	.3088	.3087	.3085	.3085	.3079	.3075
3.75	.3088	.3088	.3077	.3064	.3068	.3071	.3072	.3077	.3080	.3074	.3074	.3064	.3060
4.00	.3079	.3072	.3065	.3057	.3058	.3061	.3065	.3069	.3064	.3064	.3063	.3057	.3052

Figure 13(a). The P-norm Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

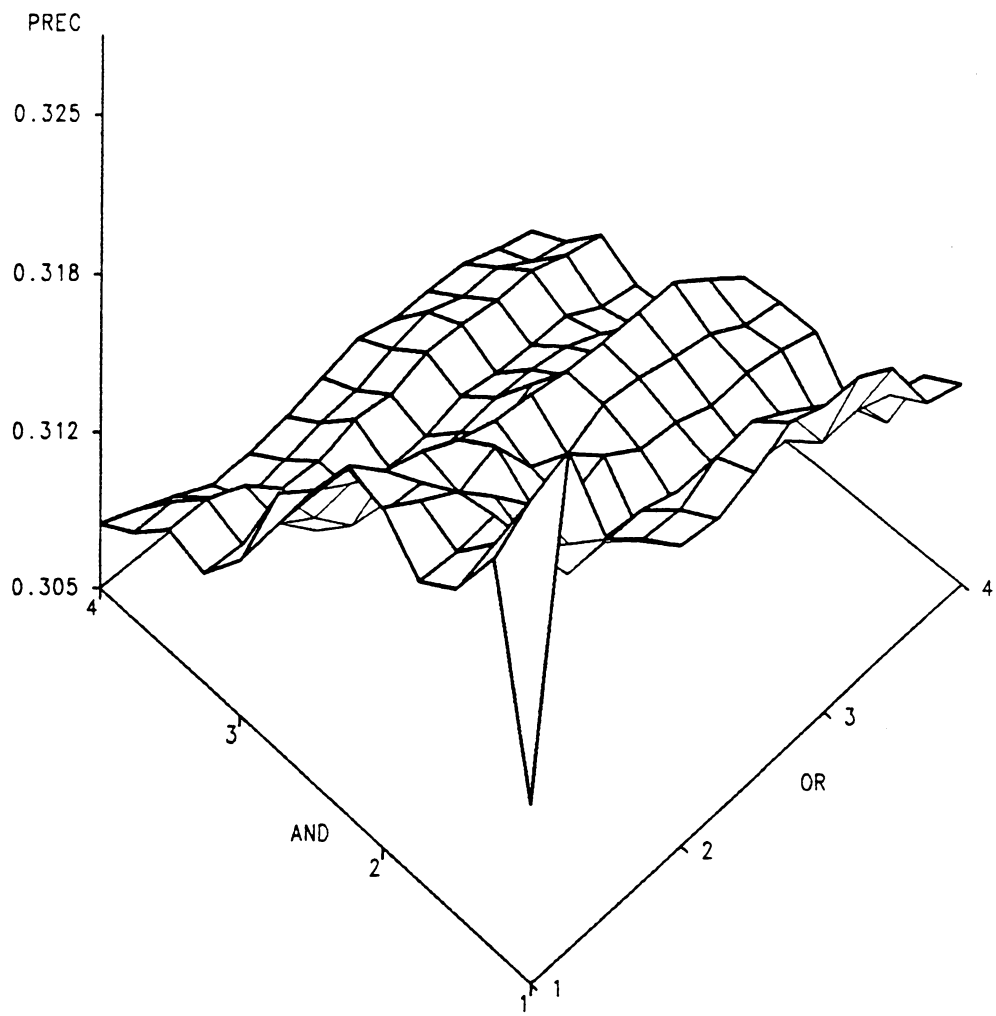


Figure 13(b). The P-norm Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

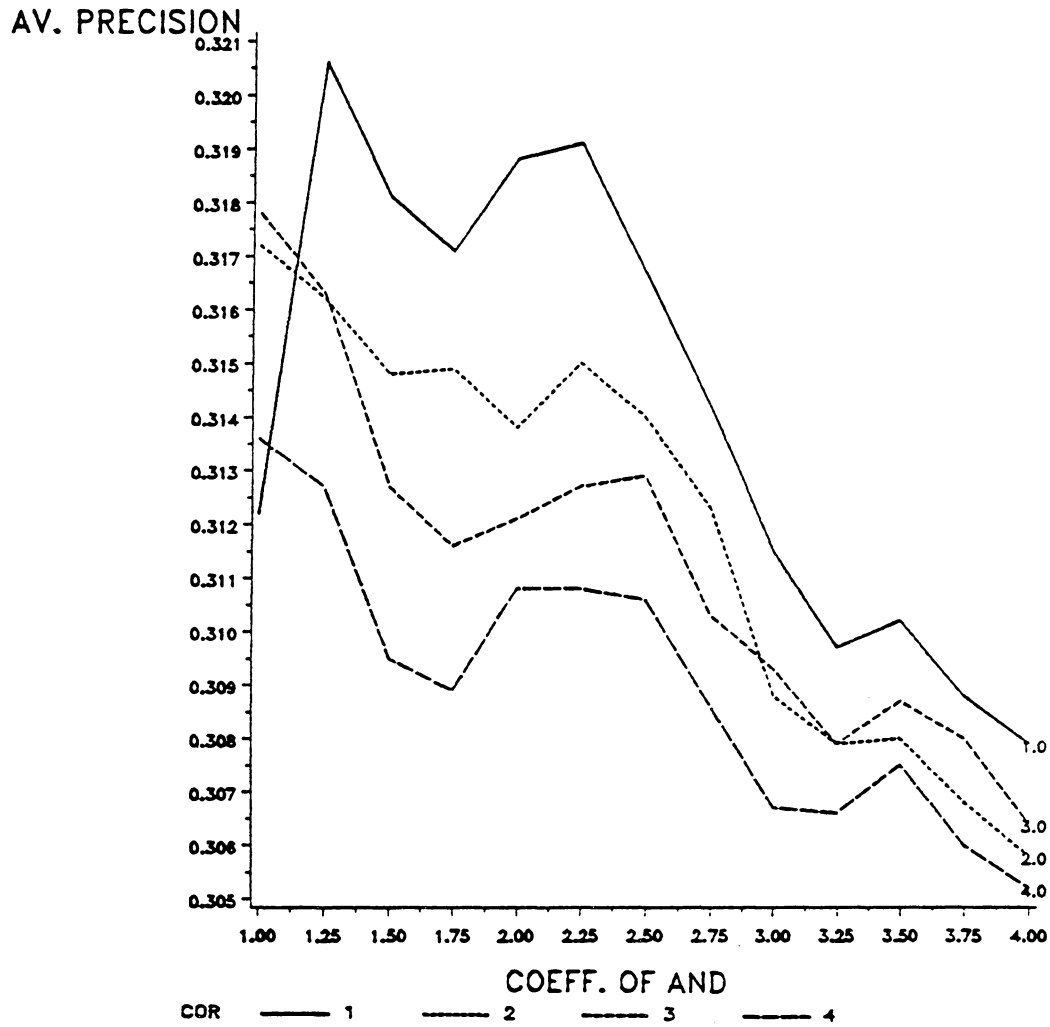


Figure 13(c). The P-norm Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$

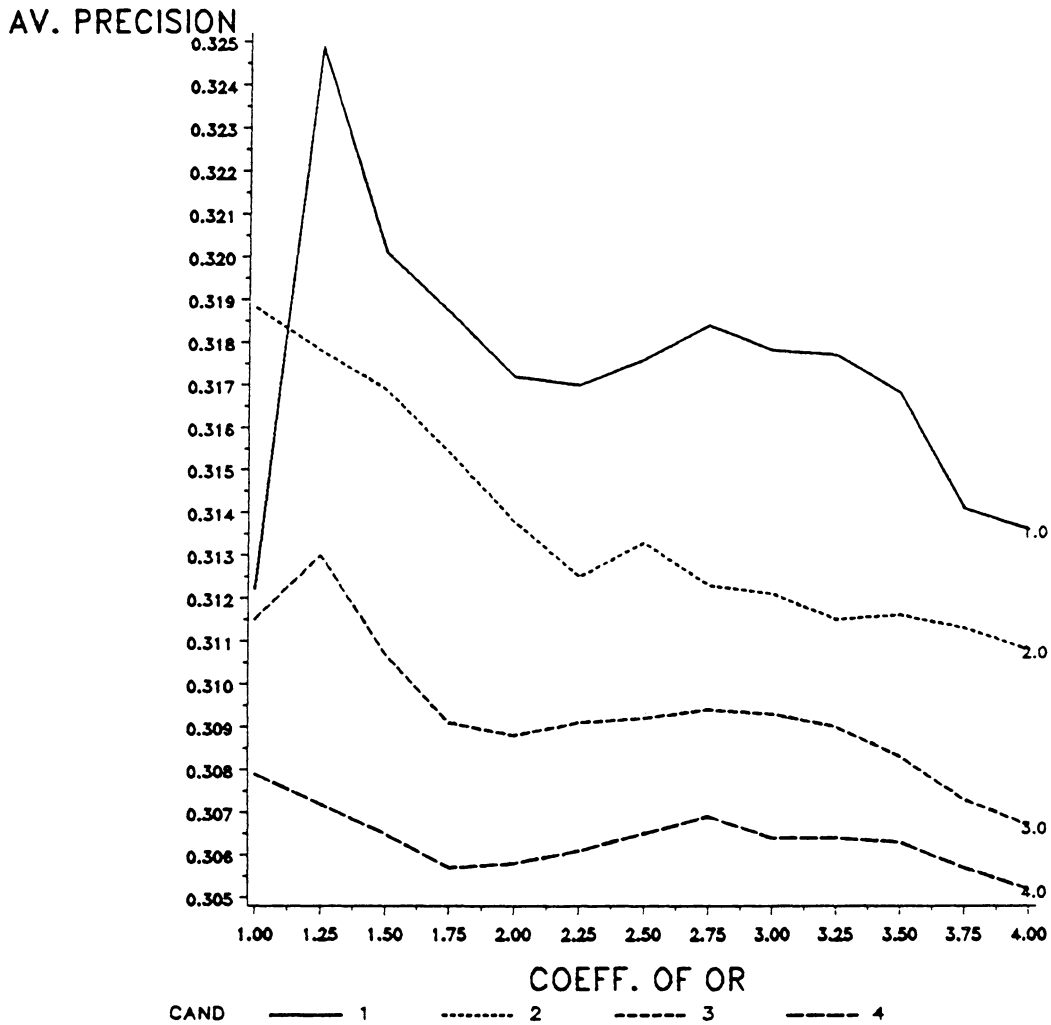


Table 16. Average Precision Values with Paice Scheme on CACM											
C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.1745	.1731	.1720	.1720	.1723	.1723	.1723	.1723	.1733	.1744	.1743
0.1	.2977	.2980	.3001	.3021	.3025	.3020	.3039	.3054	.3050	.3046	.3044
0.2	.2952	.2946	.2995	.3006	.3012	.3014	.3048	.3047	.3041	.3039	.3035
0.3	.2992	.2988	.3032	.3036	.3065	.3063	.3086	.3089	.3075	.3065	.3086
0.4	.3015	.3016	.3058	.3086	.3091	.3086	.3108	.3105	.3108	.3124	.3119
0.5	.3006	.3015	.3066	.3079	.3100	.3100	.3117	.3121	.3137	.3130	.3138
0.6	.3021	.3039	.3106	.3111	.3114	.3118	.3141	.3150	.3154	.3165	.3171
0.7	.3013	.3054	.3095	.3103	.3105	.3112	.3137	.3151	.3146	.3156	.3177
0.8	.3024	.3078	.3096	.3105	.3117	.3127	.3157	.3155	.3169	.3182	.3178
0.9	.3058	.3086	.3106	.3111	.3137	.3153	.3157	.3180	.3190	.3197	.3097
1.0	.3091	.3096	.3121	.3137	.3161	.3170	.3192	(.3215)	.3134	.3127	.3122

Figure 14(a). The Paice Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

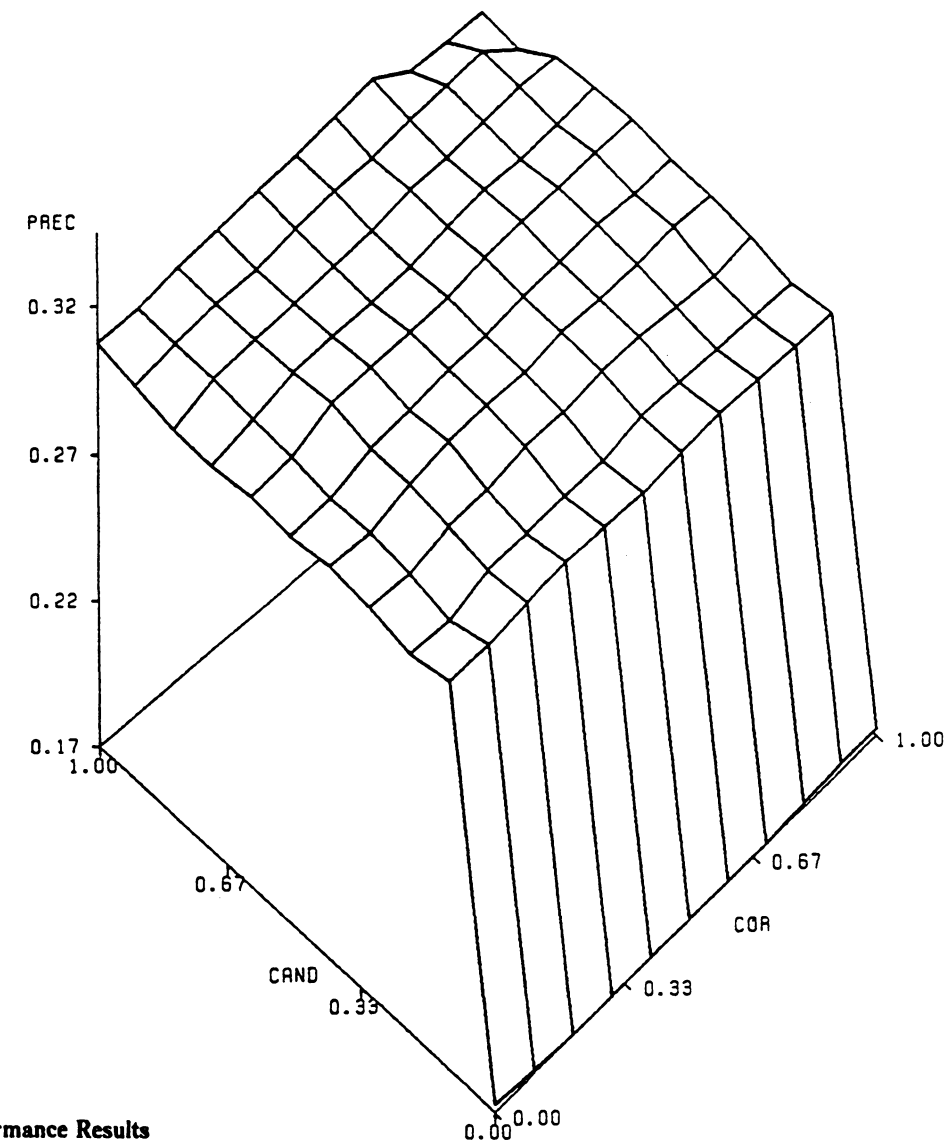


Figure 14(b). The Paice Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

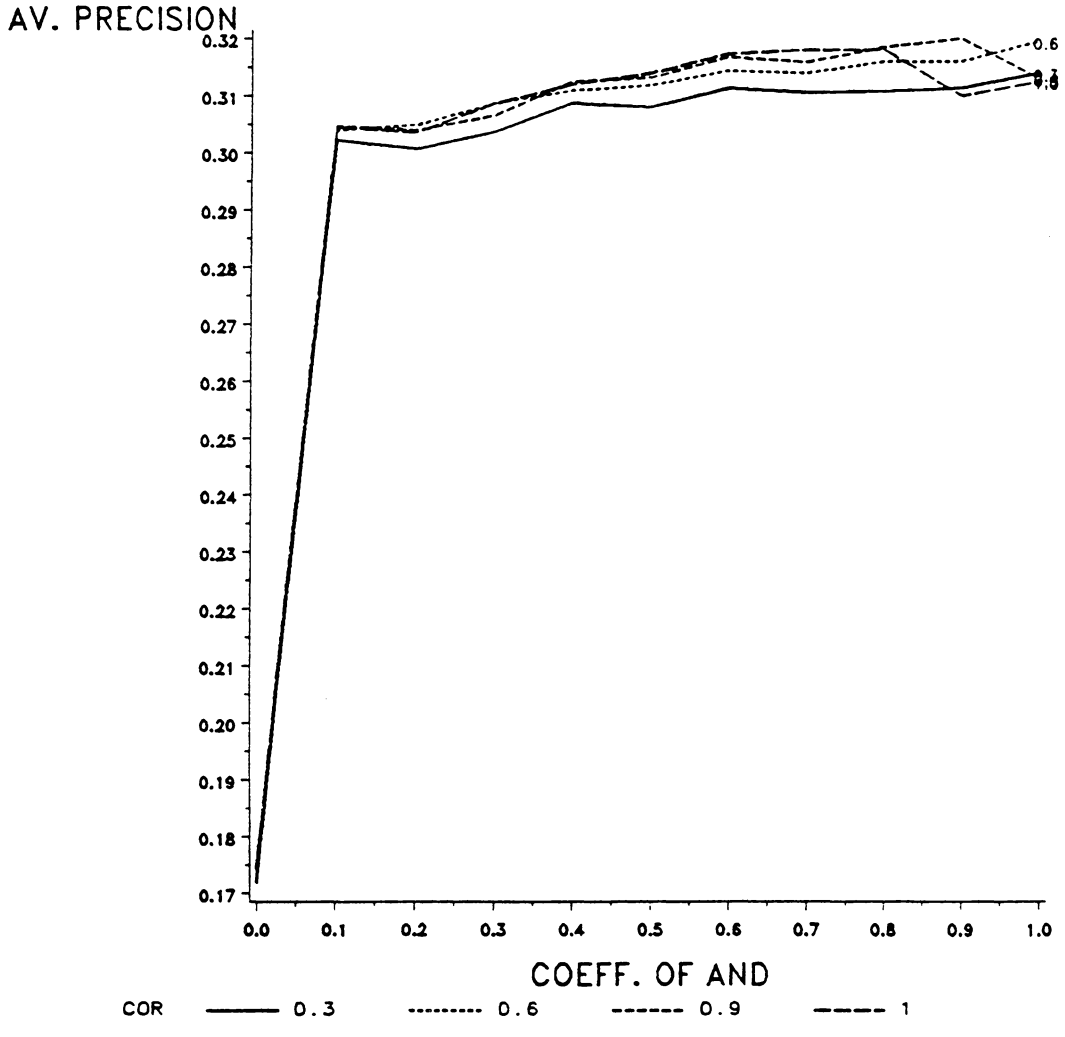


Figure 14(c). The Paice Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$

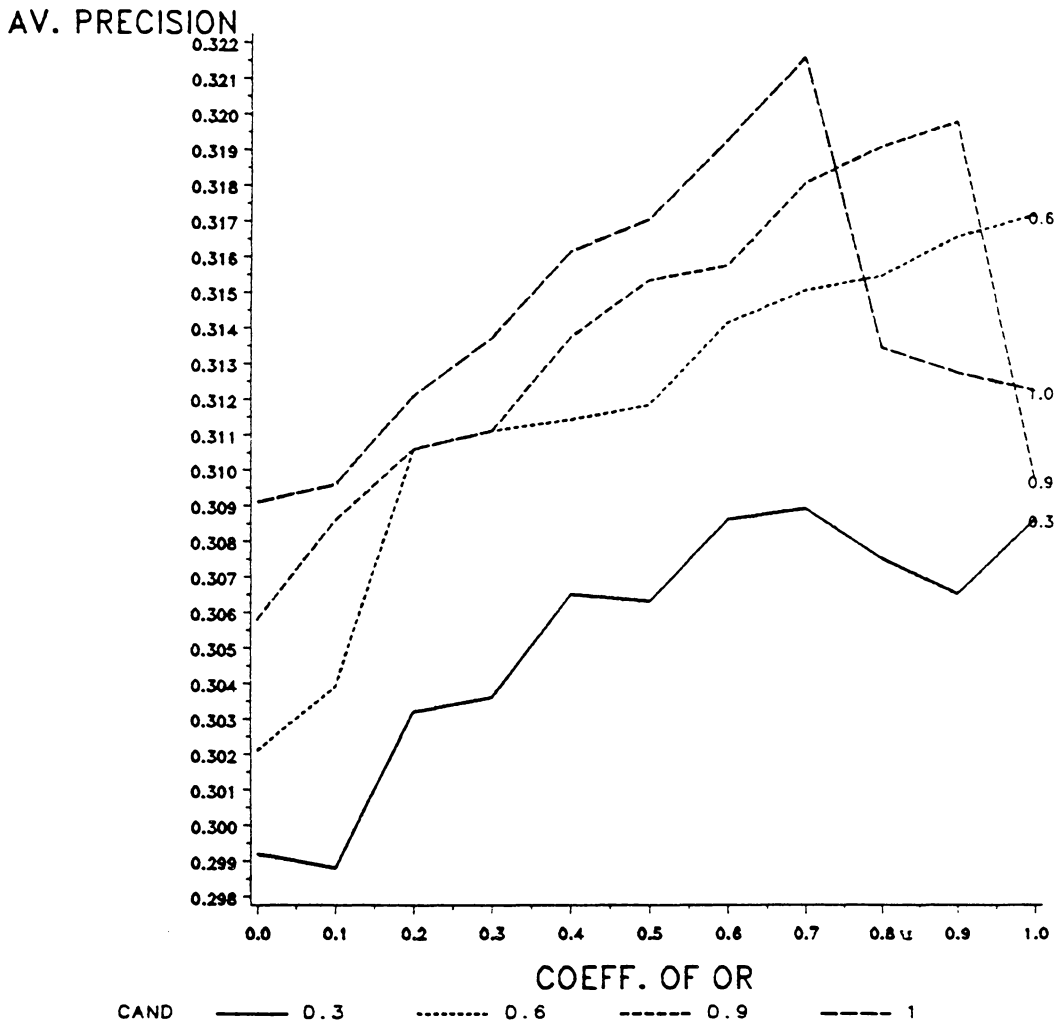


Table 17. Average Precision Values with MMM Scheme on CACM											
C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.2015	.2590	.2597	.2668	.2683	.2727	.2737	.2625	.2651	.2420	.2360
0.1	.2092	.2747	.2732	.2717	.2744	.2767	.2759	.2718	.2680	.2602	.2573
0.2	.2118	.2790	.2775	.2815	.2862	.2873	.2864	.2795	.2789	.2684	.2798
0.3	.2184	.2884	.2916	.2938	.2951	.2975	.2954	.2940	.2924	.2984	.2922
0.4	.2258	.3007	.3046	.3077	.3068	.3098	.3115	.3180	.3169	.3130	.3104
0.5	.2339	.3110	.3171	.3152	.3192	.3187	.3285	.3263	.3243	.3227	.3234
0.6	.2346	.3127	.3161	.3191	.3177	.3267	.3284	.3269	.3256	.3232	.3196
0.7	.2342	.3186	.3175	.3173	.3267	.3267	.3288	.3261	.3238	.3206	.3194
0.8	.2339	.3160	.3165	.3269	.3265	.3293	.3299	.3249	.3228	.3198	.3188
0.9	.2350	.3177	.3273	.3295	(.3300)	.3281	.3288	.3233	.3218	.3200	.3193
1.0	.0627	.1752	.1737	.1754	.1740	.1724	.1724	.1720	.1717	.1732	.1745

Figure 15(a). The MMM Scheme on CACM: Surface Plot
Average Precision vs. $Coeff_{AND}$ and $Coeff_{OR}$

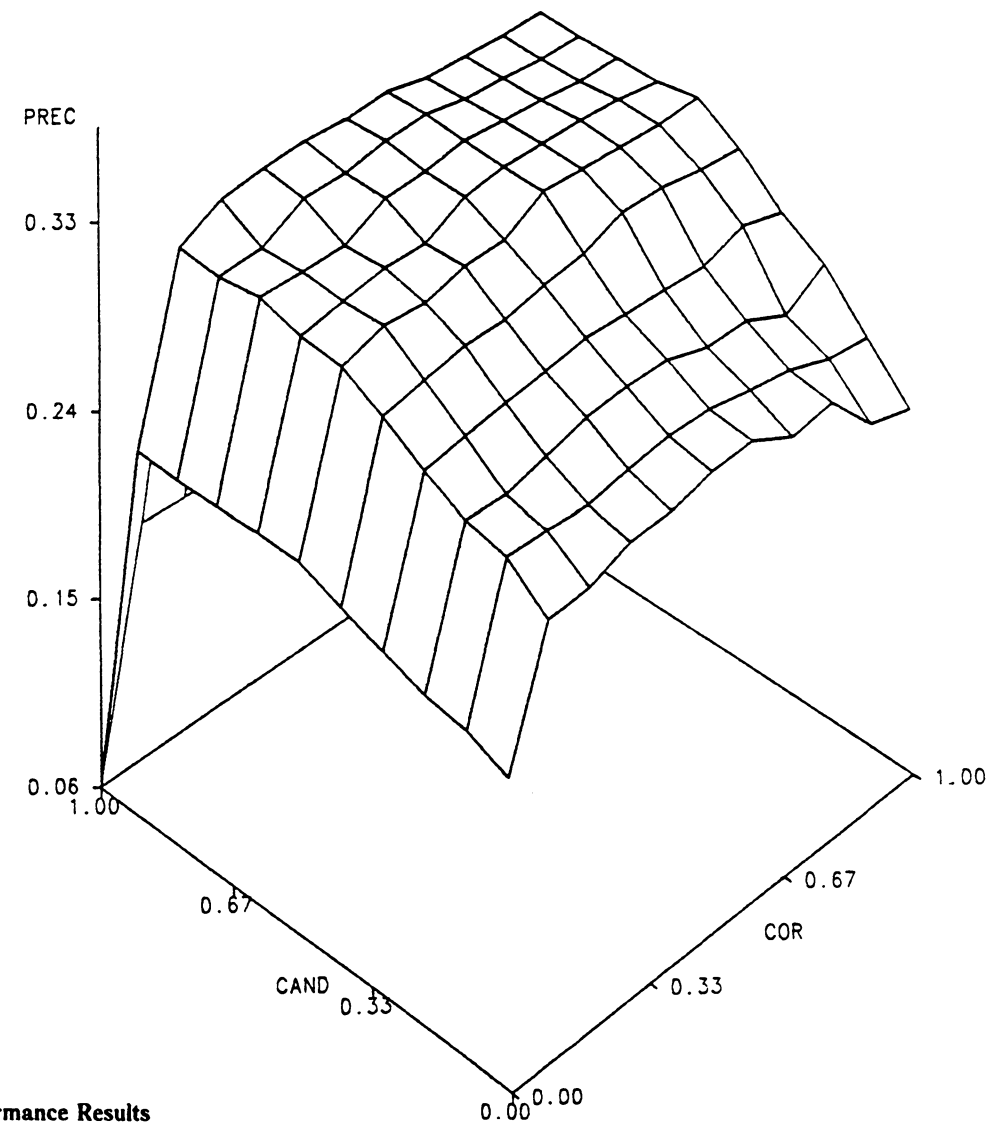


Figure 15(b). The MMM Scheme on CACM:
Average Precision vs. $Coeff_{AND}$ for Various Levels of $Coeff_{OR}$

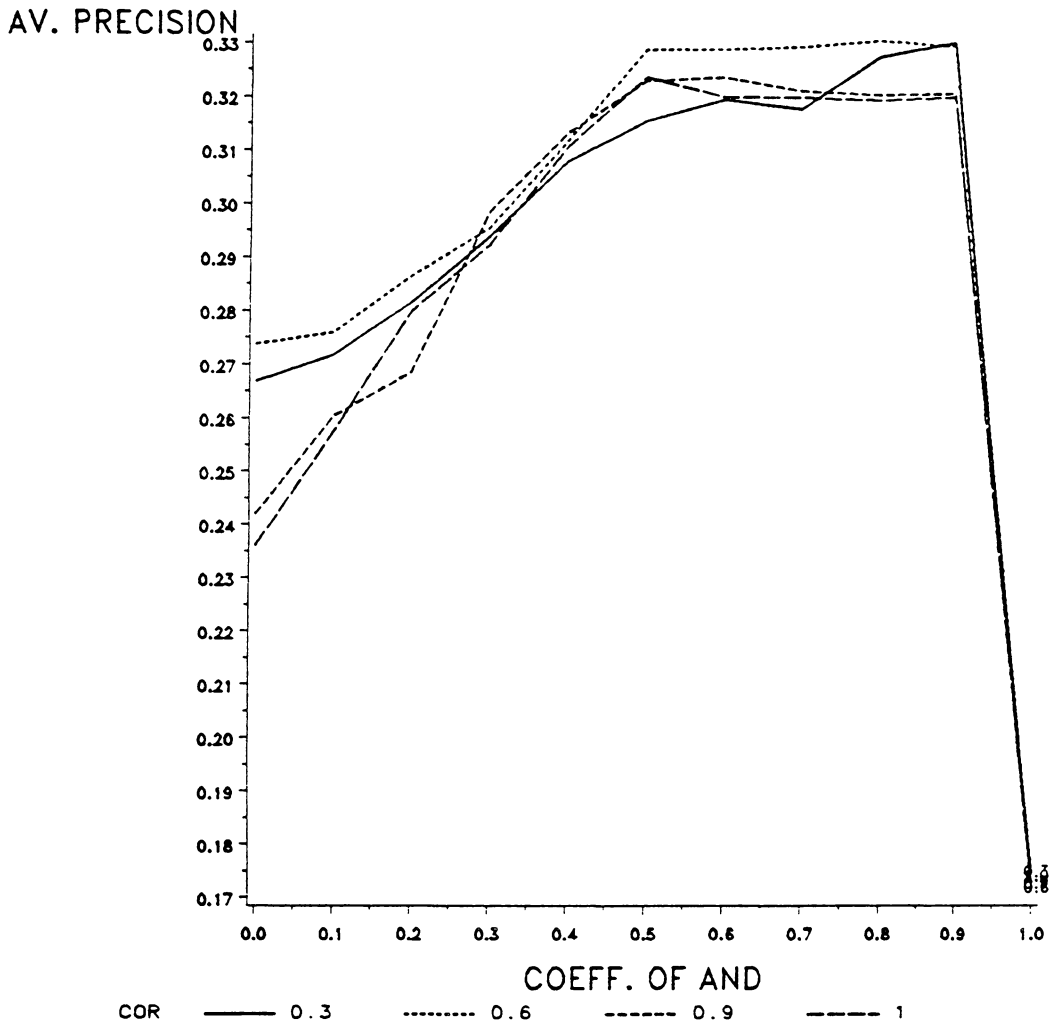
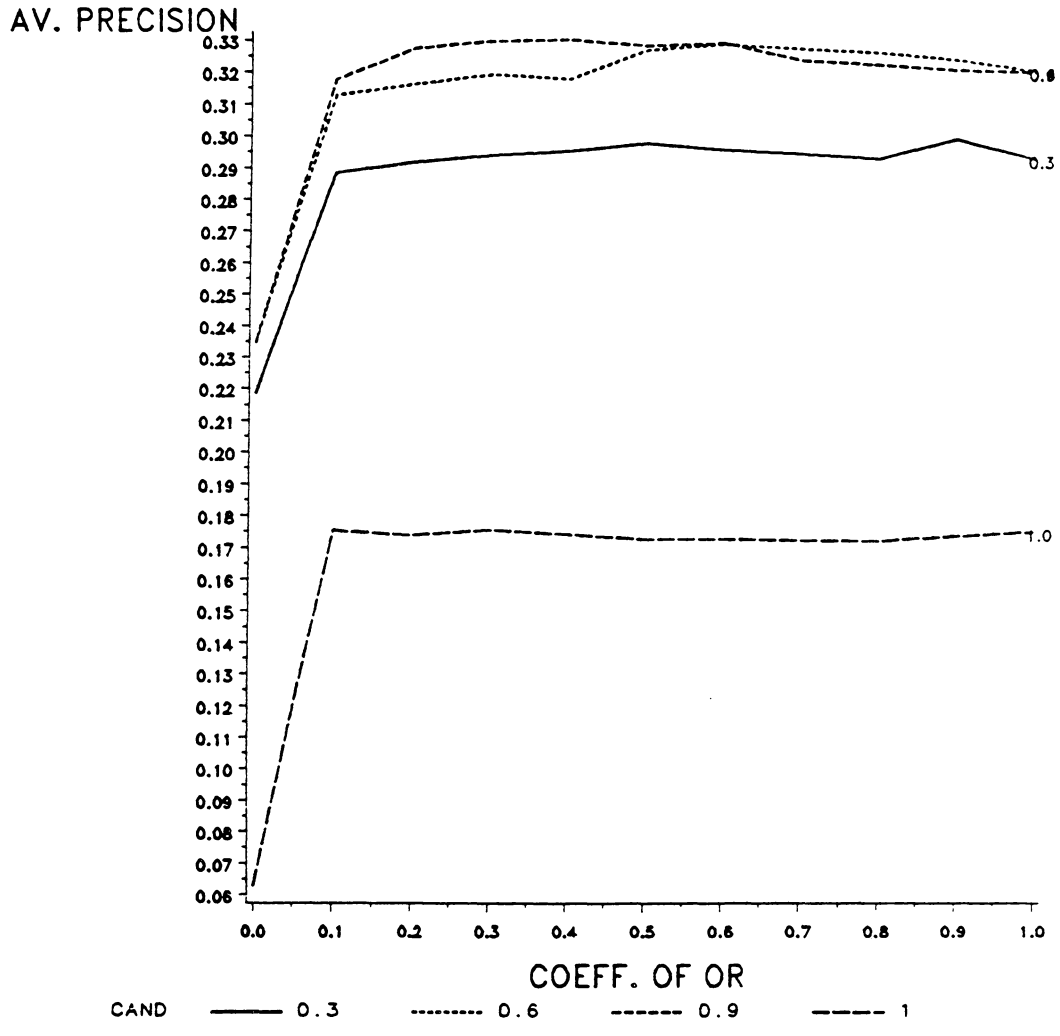


Figure 15(c). The MMM Scheme on CACM:
Average Precision vs. $Coeff_{OR}$ for Various Levels of $Coeff_{AND}$



Prediction Models

Table 18a shows a summary of the results for the SAS forward selection procedure performed on CACM. Again, the table presents the sequential steps in which the predictor variables are entered. With P-norm, C_{AND} is entered first, followed by C_{OR} and $C_{AND}C_{OR}$. With Paice, the $C_{AND}C_{OR}$ is entered first followed by C_{AND} and C_{OR}. With MMM, the C^3_{AND} variable is entered first, followed by C^2_{AND} and C_{OR}. Here, the sequence in which the predictor variables are entered with the P-norm or Paice scheme is quite different from the corresponding sequence on CISI. However, for the MMM scheme, the first two steps of the sequence in which the C^3_{AND} and C^2_{AND} are entered, correspondingly match those on CISI. For the MMM scheme, it is noted that the variable $C_{AND}C_{OR}$ is entered only in the last step on CACM, while it is entered in the third step on CISI.

Table 18b shows the best 3-, 4-, and 5-variable models obtained for each retrieval scheme using the SAS MAXR procedure. On the CACM collection, it seems that these prediction models (particularly the Paice scheme) are not as well-fitted when measured in terms of R^2 as is the case on CISI. However, in terms of R^2 and $C(p)$ combined, the set of prediction models with the P-norm scheme on CACM are considerably better than the corresponding set on CISI. It is interesting to observe that, for the set on CACM, the average precision in each model shows consistent relationships with respect to the variables C_{AND}, C_{OR}, and $C_{AND}C_{OR}$. This is not quite the case for the set on CISI; as can been seen from Table 6b, the regression coefficient for the variable C_{OR} switches sign from negative to positive, as one moves from the best 4-variable model to the best 5-variable model. While the best 5-variable model for the MMM scheme on CACM is

quite acceptable in terms of R^2 and $C(p)$ combined, none of the models shown here for the Paice scheme seems well-fitted at all.

Tables 19a and 19b show the stepwise regression results and the prediction models for the best 3, 4 and 5 variables obtained with the omission of boundary values. As on CISI, all the best 3-variable models are well-fitted for all schemes. For P-norm, the best 3-variable model has $R^2 = 0.8745$ and $C_p = 44.1628$; for Paice, $R^2 = 0.8318$ and $C_p = 26.0692$, and for MMM, $R^2 = 0.9304$ and $C_p = 30.8186$.

Table 18a. Stepwise Regression Results on CACM Collection
Summary of SAS Forward Selection Procedure
Dependent Variable : Precision

P-NORM SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{AND}	0.6888	0.6888	235.044	369.5810	0.0001
2	C_{OR}	0.1540	0.8427	39.130	162.5375	0.0001
3	$C_{AND}C_{OR}$	0.0240	0.8668	10.243	29.7611	0.0001
4	C_{OR}^2	0.0017	0.8684	10.115	2.0636	0.1528
5	C_{OR}^3	0.0046	0.8730	6.231	5.8756	0.0164
6	C_{AND}^3	0.0012	0.8742	6.703	1.5309	0.2178
7	C_{AND}^2	0.0005	0.8747	8.000	0.7030	0.4030

PAICE SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	$C_{AND}C_{OR}$	0.1349	0.1349	4.77122	2.1839	0.1616
2	C_{AND}	0.0448	0.1798	5.90186	0.7107	0.4144
3	C_{OR}	0.3078	0.4876	1.93475	7.2076	0.0199
4	C_{AND}^2	0.0627	0.5503	2.71899	1.5338	0.2413

MMM SCHEME

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{AND}^3	0.1076	0.1076	366.938	14.3539	0.0002
2	C_{AND}^2	0.4785	0.5862	109.431	136.4425	0.0001
3	C_{OR}	0.0547	0.6408	81.782	17.8090	0.0001
4	C_{OR}^2	0.0913	0.7321	34.263	39.5432	0.0001
5	C_{OR}^3	0.0394	0.7715	14.905	14.8238	0.0001
6	C_{AND}	0.0165	0.7881	7.938	8.8935	0.0035
7	$C_{AND}C_{OR}$	0.0036	0.7916	8.000	1.9379	0.1666

Table 18b. Selected Prediction Models of Average Precisions On CACM
Best 3-, 4- and 5-variable Models Obtained by MAXR

P-NORM SCHEME

$$\hat{Prec} = 0.3291 - 0.0053C_{AND} - 0.0034C_{OR} + 0.0007C_{AND}C_{OR}$$

$s^2 = 0.000002$
$R^2 = 0.866775$
$C_p = 10.24302$

$$\hat{Prec} = 0.3302 - 0.0053C_{AND} - 0.0045C_{OR} + 0.0007C_{AND}C_{OR} + 0.0002C_{OR}^2$$

$s^2 = 0.000002$
$R^2 = 0.868430$
$C_p = 10.11502$

$$\hat{Prec} = 0.3353 - 0.0053C_{AND} - 0.0119C_{OR} + 0.0007C_{AND}C_{OR} + 0.0034C_{OR}^2 - 0.0004C_{OR}^3$$

$s^2 = 0.000002$
$R^2 = 0.873008$
$C_p = 6.231063$

PAICE SCHEME

$$\hat{Prec} = 0.2217 + 0.1200C_{AND} + 0.1139C_{OR} - 0.1446C_{AND}C_{OR}$$

$s^2 = 0.000006$
$R^2 = 0.487565$
$C_p = 1.943750$

$$\hat{Prec} = 0.1684 + 0.2475C_{AND} + 0.1140C_{OR} - 0.1446C_{AND}C_{OR} - 0.0750C_{AND}^2$$

$s^2 = 0.000006$
$R^2 = 0.550275$
$C_p = 2.718988$

$$\hat{Prec} = 0.1952 + 0.2482C_{AND} - 0.1455C_{AND}C_{OR} - 0.0750C_{AND}^2 + 0.1555C_{OR}^2 - 0.0684C_{OR}^3$$

$s^2 = 0.000006$
$R^2 = 0.558496$
$C_p = 4.559592$

MMM SCHEME

$$\hat{Prec} = 0.2324 + 0.0369C_{OR} + 0.6276C_{AND}^2 - 0.6800C_{AND}^3$$

$s^2 = 0.000926$
$R^2 = 0.640830$
$C_p = 81.78215$

$$\hat{Prec} = 0.2068 + 0.2077C_{OR} + 0.6276C_{AND}^2 - 0.1709C_{OR}^2 - 0.6800C_{AND}^3$$

$s^2 = 0.000697$
$R^2 = 0.732140$
$C_p = 34.26332$

$$\hat{Prec} = 0.1918 + 0.4470C_{OR} + 0.6276C_{AND}^2 - 0.7981C_{OR}^2 - 0.6800C_{AND}^3 + 0.4182C_{OR}^3$$

$s^2 = 0.000599$
$R^2 = 0.771525$
$C_p = 14.90459$

Table 19a. Stepwise Regression Results on CACM Collection
(With Boundary Values Omitted)
Summary of SAS Forward Selection Procedure
Dependent Variable : Average Precision

P-NORM SCHEME
(with independent variables set within range 1.5-4.0)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C^2_{AND}	0.7051	0.7051	252.174	284.5151	0.0001
2	C_{OR}	0.1219	0.8270	101.516	83.1977	0.0001
3	$C_{AND}C_{OR}$	0.0474	0.8745	44.163	44.1856	0.0001
4	C^3_{AND}	0.0066	0.8811	37.862	6.4682	0.0123
5	C_{OR}	0.0148	0.8959	21.366	16.3156	0.0001

PAICE SCHEME
(with independent variables set within range 0.2-1.0)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{AND}	0.6003	0.6003	158.433	118.6355	0.0001
2	C_{OR}	0.1440	0.7443	75.631	43.9123	0.0001
3	C^2_{AND}	0.0875	0.8318	26.069	40.0757	0.0001
4	C^3_{OR}	0.0326	0.8643	8.897	18.2369	0.0001
5	C^3_{AND}	0.0093	0.8736	5.447	5.4903	0.0218
6	C^2_{OR}	0.0013	0.8749	6.698	0.7523	0.3886
7	$C_{AND}C_{OR}$	0.0012	0.8761	8.000	0.6981	0.4061

MMM SCHEME
(with independent variables set within range 0.1-0.9)

STEP	VARIABLE ENTERED	PARTIAL RSQUARE	MODEL RSQUARE	C(P)	F	PROB > F
1	C_{AND}	0.7933	0.7933	231.240	303.2170	0.0001
2	C^2_{OR}	0.1305	0.9238	38.675	133.5042	0.0001
3	$C_{AND}C_{OR}$	0.0066	0.9304	30.819	7.3101	0.0084
4	C^3_{OR}	0.0089	0.9393	19.504	11.1807	0.0013
5	C_{OR}	0.0113	0.9506	4.710	17.0878	0.0001
6	C^3_{AND}	0.0004	0.9510	6.119	0.5981	0.4418

Table 19b. Selected Prediction Models of Average Precisions On CACM (With Boundary Values Omitted) Best 3-, 4- and 5-variable Models Obtained by MAXR

P-NORM SCHEME
(with independent variables set within range 1.5-4.0)

$$\hat{Prec} = 0.3224 - 0.0039C_{OR} + 0.0009C_{AND}C_{OR} - 0.0010C_{AND}^2$$

$s^2 = 0.000001$
$R^2 = 0.874453$
$C_p = 44.16275$

$$\hat{Prec} = 0.3241 - 0.0043C_{OR} + 0.0011C_{AND}C_{OR} - 0.0016C_{AND}^2 + 0.0001C_{AND}^3$$

$s^2 = 0.000001$
$R^2 = 0.881084$
$C_p = 37.86209$

$$\hat{Prec} = 0.3061 + 0.0213C_{AND} - 0.0042C_{OR} + 0.0011C_{AND}C_{OR} - 0.0096C_{AND}^2 + 0.0011C_{AND}^3$$

$s^2 = 0.000001$
$R^2 = 0.895859$
$C_p = 21.36639$

PAICE SCHEME
(with independent variables set within range 0.2-1.0)

$$\hat{Prec} = 0.2913 + 0.0434C_{AND} + 0.0070C_{OR} - 0.0242C_{AND}^2$$

$s^2 = 0.000004$
$R^2 = 0.831797$
$C_p = 26.06915$

$$\hat{Prec} = 0.2908 + 0.0434C_{AND} - 0.0242C_{AND}^2 + 0.0307C_{OR}^2 - 0.0249C_{OR}^3$$

$s^2 = 0.000003$
$R^2 = 0.865539$
$C_p = 8.195776$

$$\hat{Prec} = 0.2855 + 0.0786C_{AND} - 0.0900C_{AND}^2 + 0.0307C_{OR}^2 + 0.0366C_{AND}^3 - 0.0249C_{OR}^3$$

$s^2 = 0.000003$
$R^2 = 0.874792$
$C_p = 4.745967$

MMM SCHEME
(with independent variables set within range 0.1-0.9)

$$\hat{Prec} = 0.2509 + 0.1861C_{AND} + 0.0111C_{AND}C_{OR} - 0.1230C_{AND}^2$$

$s^2 = 0.000029$
$R^2 = 0.930385$
$C_p = 30.81862$

$$\hat{Prec} = 0.2415 + 0.1917C_{AND} + 0.0332C_{OR} - 0.1230C_{AND}^2 - 0.0319C_{OR}^3$$

$s^2 = 0.000022$
$R^2 = 0.949027$
$C_p = 5.016495$

$$\hat{Prec} = 0.2444 + 0.1858C_{AND} + 0.0273C_{AND} + 0.0117C_{AND}C_{OR} - 0.1230C_{AND}^2 - 0.0319C_{OR}^3$$

$s^2 = 0.000021$
$R^2 = 0.950574$
$C_p = 4.710257$

Discussion

Scheme	Best Precision	Rank	Best E-measure	Rank
P-NORM	.3249	2	.7346	1
PAICE	.3215	3	.7358	2
MMM	.3300	1	.7429	3
TIRS	.2804	4	.7506	4

Table 20. Relative Ranks of Schemes by Average Precision and E-measure on CACM.

Table 20 shows the summary of best average precision values and E-measures with $\beta = 1$ on the CACM collection for all the retrieval schemes being considered and their relative ranks. In terms of average precisions, MMM is ranked first, followed by P-norm, Paice and lastly TIRS. The TIRS average precision is much lower than the best average precision for any of the other schemes. This set of rankings is not quite the same as that on CISI. Recall that on CISI, the order is P-norm, Paice, MMM and TIRS. Nonetheless, in terms of E-measures at β-level 1, the rankings of retrieval schemes on CACM are consistent with those on CISI.

Consider the ten top-ranked documents retrieved for each scheme using query 24 as shown in Table 21. For all of the schemes, 3 out of the ten top-ranked documents retrieved are relevant. Documents 1696, 268, and 749 are consistently ranked as respectively the first, second and third by each of the schemes in the ten top-ranked documents. Query 24 is shown in Figure 16, and documents 1696 and 749 are shown in Figures 18 and 19 respectively. Document 1696 is certainly a relevant document. Document 749 is a very short document, and it has a match of the query term

"stochastic". With respect to query 24, this document seems relevant, but it was not judged as one of the relevant documents in the collection.

From Table 21, it is interesting that P-norm, Paice and TIRS each ranks the same set of ten top-ranked document in exactly the same way. However, this is not so for MMM. On a closer look, we see that MMM did not retrieve the same set of ten top-ranked documents and that it ranked document 1892 more accurately than the other schemes.

(a) P-NORM SCHEME $C_{and} = 1.00, C_{or} = 1.25$

document ID	rank	relevant	similarity
1696	1	1	0.288768
268	2	1	0.227415
749	3	0	0.227415
1410	4	0	0.195959
1892	5	1	0.194885
1194	6	0	0.191667
2535	7	0	0.190238
2742	8	0	0.186343
1435	9	0	0.186343
1135	10	0	0.153422

(b) PAICE SCHEME $C_{and} = 1.0, C_{or} = 0.7$

document ID	rank	relevant	similarity
1696	1	1	0.288768
268	2	1	0.227415
749	3	0	0.227415
1410	4	0	0.195959
1892	5	1	0.194885
1194	6	0	0.191667
2535	7	0	0.190238
2742	8	0	0.186343
1435	9	0	0.186343
1135	10	0	0.153422

(c) MMM SCHEME $C_{and} = 0.9, C_{or} = 0.4$

document ID	rank	relevant	similarity
1696	1	1	0.068225
268	2	1	0.068224
749	3	0	0.068224
1892	4	1	0.042640
1540	5	0	0.042640
1194	6	0	0.040935
2742	7	0	0.039798
1435	8	0	0.039798
1235	9	0	0.039798
1410	10	0	0.039798

(d) TIRS SCHEME

document ID	rank	relevant	similarity
1696	1	1	0.866304
268	2	1	0.682244
749	3	0	0.682244
1410	4	0	0.587878
1892	5	1	0.584655
1194	6	0	0.575000
2535	7	0	0.570713
2742	8	0	0.559028
1435	9	0	0.559028
1135	10	0	0.460265

Table 21. Ten Top-ranked Documents Retrieved with Query 24 on CACM

The query, Q^{24} is as follows:

$Q^{24} = (\ AND\ (<\text{applied, 1.0}>,\ <\text{stochastic, 1.0}>,\ <\text{processes, 1.0}>))$

Figure 16. Query 24 from CACM Collection

The document, D^{1696} is as follows:

.I 1696
.T
An Algorithm for Identifying the Ergodic Subchains
and Transient States of a ***Stochastic*** Matrix
.W
An algorithm for identifying the ergodic subchains
and transient states of a ***stochastic*** matrix
is presented. Applications in Markov renewal programming
and in the construction of variable length
codes are reviewed, and an updating procedure for dealing
with certain sequences of ***stochastic*** matrices
is discussed. Computation times are investigated experimentally
and compared with those of another recently
propose method.
.B
CACM September, 1968
.A
Fox, B. L.
Landi, D. M.
.K
stochastic matrix, ergodic, chain identification
.C
5.39 5.5
.N
CA680905 JB February 22, 1978 9:04 AM

Figure 17. Document 1696 from CACM Collection

The document, D^{749} is as follows:

.I 749
.T
Note on ***Stochastic*** Matrices
.B
CACM September, 1963
.A
Dumey, A. I.
.N
CA630909 JB March 13, 1978 7:35 PM

Figure 18. Document 749 from CACM Collection

Table 22 presents the summary of the best E-measures at β-levels 0.5, 1.0 and 2.0, obtained from Tables 23, 24 and 25. Whether the β-level is set at 0.5, 1.0 or 2.0, the E-measure consistently ranks the retrieval schemes being considered in the order P-norm, Paice, MMM and TIRS.

The best E-measures for P-norm and Paice on the CACM collection do not seem to be significantly different from each other. The best E-measure of the MMM runs indicates that the MMM scheme does not do as well as P-norm or Paice, but performs relatively better than the TIRS scheme.

Scheme	Best E-measure $\beta = 0.5$	Rank	Best E-measure $\beta = 1.0$	Rank	Best E-measure $\beta = 2.0$	Rank
P-NORM	.7663	1	.7346	1	.6654	1
PAICE	.7676	2	.7358	2	.6668	2
MMM	.7746	3	.7429	3	.6740	3
TIRS	.7790	4	.7506	4	.6893	4

Table 22. Relative Ranks of Schemes by E-measure at β-levels 0.5, 1.0 and 2.0 on CACM

Table 23(a). E-measures with P-norm Scheme on CACM at $\beta = 0.5$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.7725	.7694	.7673	.7695	.7695	.7707	.7719	.7725	.7726	.7720	.7721	.7721	.7721
1.25	.7723	.7692	.7671	.7694	.7699	.7712	.7710	.7723	.7729	.7724	.7712	.7719	.7719
1.50	.7717	.7685	(.7663)	.7694	.7699	.7698	.7717	.7730	.7730	.7731	.7733	.7732	.7732
1.75	.7723	.7684	.7675	.7692	.7698	.7704	.7724	.7736	.7736	.7737	.7731	.7731	.7738
2.00	.7723	.7690	.7690	.7707	.7705	.7711	.7723	.7729	.7743	.7738	.7738	.7738	.7745
2.25	.7730	.7683	.7677	.7713	.7698	.7703	.7723	.7730	.7737	.7737	.7744	.7738	.7738
2.50	.7736	.7683	.7684	.7726	.7705	.7717	.7730	.7744	.7752	.7752	.7746	.7753	.7753
2.75	.7741	.7695	.7697	.7724	.7738	.7745	.7752	.7758	.7758	.7758	.7759	.7766	.7766
3.00	.7741	.7709	.7697	.7739	.7746	.7746	.7759	.7765	.7758	.7758	.7765	.7773	.7786
3.25	.7741	.7702	.7717	.7746	.7746	.7760	.7773	.7773	.7773	.7778	.7779	.7779	.7779
3.50	.7749	.7717	.7716	.7761	.7768	.7768	.7779	.7779	.7785	.7792	.7786	.7793	.7793
3.75	.7763	.7725	.7724	.7768	.7768	.7768	.7786	.7786	.7792	.7799	.7793	.7793	.7800
4.00	.7763	.7746	.7730	.7768	.7775	.7775	.7786	.7792	.7799	.7799	.7793	.7800	.7800

Table 23(b). E-measures with Paice Scheme on CACM at $\beta = 0.5$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8929	.8929	.8929	.8929	.8929	.8929	.8929	.8929	.8929	.8918	.8923
0.1	.7957	.7922	.7895	.7883	.7857	.7876	.7858	.7852	.7853	.7860	.7855
0.2	.7943	.7916	.7884	.7871	.7845	.7838	.7814	.7790	.7795	.7809	.7815
0.3	.7930	.7895	.7871	.7859	.7801	.7795	.7783	.7783	.7783	.7790	.7773
0.4	.7915	.7876	.7857	.7798	.7774	.7773	.7756	.7756	.7755	.7783	.7768
0.5	.7884	.7850	.7812	.7787	.7749	.7741	.7729	.7730	.7743	.7737	.7750
0.6	.7865	.7818	.7749	.7745	.7722	.7727	.7724	.7718	.7704	.7725	.7731
0.7	.7852	.7810	.7751	.7732	.7708	.7736	.7712	.7699	.7705	.7725	.7731
0.8	.7841	.7792	.7751	.7724	.7722	.7722	.7704	.7699	.7711	.7717	.7730
0.9	.7821	.7773	.7731	.7719	.7704	.7701	(.7676)	.7692	.7698	.7717	.7723
1.0	.7823	.7768	.7733	.7728	.7698	.7696	.7678	.7687	.7707	.7719	.7725

Table 23(c). E-measures with MMM Scheme on CACM at $\beta = 0.5$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8415	.8115	.8088	.8027	.8017	.8017	.8023	.8025	.8035	.8074	.8124
0.1	.8386	.7954	.7954	.7927	.7940	.7933	.7931	.7976	.7989	.7990	.8055
0.2	.8359	.7913	.7906	.7898	.7883	.7891	.7895	.7901	.7901	.7924	.7952
0.3	.8327	.7874	.7859	.7836	.7830	.7828	.7815	.7816	.7815	.7820	.7879
0.4	.8333	.7844	.7815	.7813	.7806	.7804	.7803	.7779	.7788	.7803	.7849
0.5	.8319	.7793	.7784	.7795	.7807	.7759	.7752	.7749	.7765	.7788	.7834
0.6	.8304	.7771	.7780	.7785	.7752	.7759	.7757	(.7746)	.7764	.7790	.7830
0.7	.8311	.7801	.7792	.7767	.7767	.7759	.7761	.7779	.7805	.7831	.7875
0.8	.8311	.7798	.7767	.7773	.7773	.7770	.7776	.7792	.7828	.7856	.7896
0.9	.8324	.7780	.7801	.7778	.7796	.7808	.7823	.7828	.7842	.7869	.7903
1.0	.9562	.8872	.8872	.8878	.8907	.8923	.8929	.8929	.8929	.8929	.8929

Table 24(a). E-measures with P-norm Scheme on CACM at $\beta = 1.0$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.7416	.7386	.7362	.7381	.7379	.7390	.7402	.7406	.7408	.7403	.7405	.7405	.7405
1.25	.7410	.7380	.7356	.7375	.7380	.7394	.7387	.7400	.7407	.7402	.7392	.7399	.7399
1.50	.7404	.7373	(.7346)	.7375	.7380	.7375	.7396	.7409	.7409	.7412	.7417	.7415	.7415
1.75	.7410	.7370	.7354	.7370	.7377	.7381	.7402	.7416	.7416	.7418	.7412	.7412	.7419
2.00	.7409	.7374	.7374	.7389	.7384	.7388	.7400	.7407	.7423	.7419	.7420	.7420	.7427
2.25	.7418	.7365	.7359	.7395	.7374	.7379	.7400	.7410	.7417	.7417	.7425	.7420	.7420
2.50	.7424	.7364	.7368	.7408	.7384	.7395	.7410	.7426	.7438	.7438	.7433	.7440	.7440
2.75	.7426	.7377	.7381	.7402	.7418	.7427	.7437	.7442	.7442	.7442	.7446	.7452	.7452
3.00	.7426	.7393	.7381	.7421	.7430	.7430	.7445	.7450	.7442	.7442	.7453	.7462	.7476
3.25	.7426	.7384	.7405	.7430	.7430	.7449	.7464	.7462	.7462	.7466	.7468	.7468	.7468
3.50	.7435	.7405	.7402	.7451	.7458	.7458	.7467	.7467	.7475	.7482	.7477	.7487	.7487
3.75	.7454	.7414	.7414	.7458	.7458	.7458	.7477	.7477	.7482	.7492	.7487	.7487	.7494
4.00	.7454	.7440	.7419	.7458	.7468	.7468	.7477	.7482	.7492	.7492	.7487	.7494	.7494

Table 24(b). E-measures with Paice Scheme on CACM at $\beta = 1.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8736	.8736	.8736	.8736	.8736	.8736	.8736	.8736	.8736	.8726	.8731
0.1	.7659	.7620	.7590	.7577	.7551	.7572	.7554	.7549	.7550	.7559	.7554
0.2	.7512	.7645	.7615	.7581	.7568	.7541	.7532	.7509	.7486	.7491	.7507
0.3	.7631	.7591	.7568	.7558	.7496	.7490	.7480	.7480	.7480	.7486	.7460
0.4	.7613	.7572	.7552	.7488	.7463	.7462	.7446	.7446	.7444	.7476	.7455
0.5	.7581	.7542	.7505	.7480	.7441	.7428	.7418	.7420	.7432	.7427	.7442
0.6	.7563	.7509	.7429	.7430	.7408	.7413	.7416	.7409	.7391	.7417	.7423
0.7	.7548	.7497	.7435	.7414	.7391	.7428	.7403	.7389	.7394	.7417	.7423
0.8	.7539	.7481	.7435	.7405	.7411	.7410	.7393	.7389	.7399	.7406	.7418
0.9	.7518	.7462	.7414	.7401	.7392	.7385	(.7358)	.7380	.7385	.7405	.7410
1.0	.7523	.7461	.7420	.7416	.7386	.7381	.7366	.7379	.7400	.7411	.7416

Table 24(c). E-measures with MMM Scheme on CACM at $\beta = 1.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8214	.7880	.7851	.7783	.7767	.7767	.7774	.7769	.7778	.7820	.7869
0.1	.8178	.7694	.7694	.7662	.7678	.7668	.7663	.7712	.7716	.7719	.7786
0.2	.8147	.7647	.7640	.7627	.7609	.7620	.7620	.7627	.7618	.7646	.7667
0.3	.8114	.7604	.7586	.7556	.7550	.7545	.7531	.7527	.7523	.7526	.7591
0.4	.8118	.7564	.7529	.7525	.7516	.7512	.7511	.7477	.7480	.7500	.7549
0.5	.8100	.7498	.7486	.7494	.7506	.7454	.7445	.7436	.7450	.7480	.7530
0.6	.8081	.7472	.7476	.7478	.7444	.7455	.7448	(.7429)	.7445	.7476	.7518
0.7	.8090	.7500	.7488	.7464	.7464	.7452	.7449	.7466	.7494	.7522	.7566
0.8	.8090	.7494	.7464	.7471	.7472	.7462	.7467	.7480	.7513	.7549	.7591
0.9	.8103	.7477	.7503	.7473	.7489	.7502	.7511	.7514	.7529	.7561	.7597
1.0	.9512	.8685	.8685	.8693	.8717	.8731	.8736	.8736	.8736	.8736	.8736

Table 25(a). E-measures with P-norm Scheme on CACM at $\beta = 2.0$

C_{AND} \ C_{OR}	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
1.00	.6742	.6712	.6685	.6698	.6695	.6706	.6718	.6719	.6722	.6718	.6720	.6720	.6720
1.25	.6730	.6700	.6672	.6686	.6690	.6708	.6693	.6706	.6713	.6709	.6700	.6707	.6707
1.50	.6723	.6693	.6657	.6686	.6690	.6682	.6706	.6719	.6719	.6722	.6738	.6732	.6732
1.75	.6729	.6686	(.6654)	.6669	.6682	.6686	.6710	.6724	.6724	.6727	.6722	.6722	.6729
2.00	.6724	.6685	.6684	.6699	.6691	.6693	.6704	.6711	.6733	.6729	.6731	.6731	.6738
2.25	.6737	.6672	.6664	.6704	.6675	.6680	.6704	.6719	.6726	.6726	.6735	.6731	.6731
2.50	.6740	.6669	.6677	.6717	.6688	.6700	.6719	.6739	.6764	.6764	.6759	.6766	.6766
2.75	.6738	.6681	.6691	.6704	.6724	.6745	.6760	.6764	.6764	.6764	.6772	.6779	.6779
3.00	.6738	.6702	.6691	.6737	.6750	.6750	.6769	.6773	.6764	.6764	.6789	.6801	.6815
3.25	.6738	.6689	.6727	.6750	.6750	.6782	.6802	.6799	.6799	.6803	.6805	.6805	.6805
3.50	.6749	.6726	.6720	.6787	.6794	.6794	.6802	.6802	.6815	.6822	.6817	.6832	.6832
3.75	.6782	.6740	.6745	.6794	.6794	.6794	.6818	.6818	.6822	.6837	.6832	.6832	.6839
4.00	.6782	.6779	.6750	.6794	.6809	.6809	.6818	.6822	.6837	.6837	.6832	.6839	.6839

Table 25(b). E-measures with Paice Scheme on CACM at $\beta = 2.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.8338	.8338	.8338	.8338	.8338	.8338	.8338	.8338	.8338	.8330	.8334
0.1	.7020	.6974	.6938	.6925	.6896	.6922	.6905	.6901	.6902	.6911	.6907
0.2	.7008	.6975	.6934	.6922	.6892	.6879	.6858	.6833	.6838	.6855	.6859
0.3	.6993	.6947	.6924	.6914	.6842	.6837	.6827	.6827	.6828	.6832	.6783
0.4	.6971	.6924	.6902	.6820	.6794	.6792	.6781	.6781	.6776	.6815	.6777
0.5	.6929	.6883	.6843	.6824	.6778	.6752	.6742	.6745	.6757	.6752	.6770
0.6	.6917	.6851	.6745	.6751	.6730	.6733	.6745	.6738	.6712	.6745	.6752
0.7	.6897	.6826	.6756	.6731	.6709	.6758	.6731	.6715	.6719	.6748	.6752
0.8	.6889	.6810	.6756	.6719	.6737	.6736	.6718	.6714	.6720	.6727	.6739
0.9	.6867	.6791	.6734	.6719	.6717	.6700	(.6668)	.6700	.6705	.6725	.6730
1.0	.6878	.6797	.6747	.6745	.6709	.6699	.6687	.6705	.6729	.6738	.6742

Table 25(c). E-measures with MMM Scheme on CACM at $\beta = 2.0$

C_{AND} \ C_{OR}	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	.7760	.7334	.7303	.7229	.7203	.7203	.7210	.7199	.7213	.7259	.7306
0.1	.7702	.7107	.7107	.7069	.7086	.7073	.7065	.7121	.7119	.7120	.7192
0.2	.7664	.7050	.7041	.7023	.6999	.7015	.7010	.7020	.7003	.7039	.7046
0.3	.7629	.6997	.6974	.6933	.6926	.6918	.6904	.6894	.6887	.6889	.6963
0.4	.7625	.6936	.6891	.6883	.6875	.6867	.6869	.6820	.6816	.6844	.6898
0.5	.7592	.6835	.6818	.6825	.6839	.6780	.6767	.6749	.6764	.6808	.6864
0.6	.7566	.6800	.6799	.6801	.6763	.6784	.6769	(.6740)	.6755	.6795	.6841
0.7	.7579	.6829	.6813	.6797	.6797	.6777	.6769	.6785	.6819	.6853	.6898
0.8	.7579	.6820	.6797	.6803	.6806	.6787	.6791	.6801	.6835	.6883	.6929
0.9	.7591	.6809	.6844	.6801	.6816	.6833	.6834	.6832	.6849	.6892	.6931
1.0	.9387	.8292	.8292	.8302	.8322	.8334	.8338	.8338	.8338	.8338	.8338

INSPEC

Collected Results

Tables 26(a), (b) and (c) present the average precision values obtained on the INSPEC collection using the P-norm, Paice and MMM schemes. Tables 27, 28 and 29 present the corresponding E-measures at β levels 0.5, 1.0 and 2.0, respectively. The best scores are summarized in Tables 30 and 31, along with their relative ranks to aid comparison of relative retrieval effectiveness of the schemes as before on CISI and CACM. Note that at $Coeff_{AND}$ and $Coeff_{OR}$ equal to 1, the average precision value with the MMM scheme is 0.1497, which is also the average precision value with the classical fuzzy-set scheme. The P-norm, Paice and MMM schemes show an improvement of about 100 percent on INSPEC over the classical fuzzy-set scheme in terms of average precisions. The average precision value obtained on INSPEC with the standard Boolean scheme is 0.0998. The P-norm, Paice and MMM schemes show an improvement of close to 200 percent on INSPEC over the standard Boolean scheme in terms of average precisions.

Table 26. Average Precision Values on INSPEC

(a) P-NORM SCHEME

C_{AND}	C_{OR} 1.00	1.25	1.50	1.75	2.00
1.00	.2972	.3044	(.3093)	.3077	.3070
1.25	.2982	.3061	(.3093)	.3072	.3065
1.50	.3001	.3059	.3081	.3064	.3057
1.75	.3012	.3056	.3068	.3060	.3043
2.00	.3010	.3062	.3072	.3062	.3051

(b) PAICE SCHEME

C_{AND}	C_{OR} 0.70	0.80	0.90	0.95
0.70	.3022	.3009	.3019	.3010
0.80	.3038	.3042	.3032	.3020
0.90	(.3061)	.3051	.3037	.3023
0.95	.3055	.3040	.3024	.3009

(C) MMM SCHEME

C_{AND}	C_{OR} 0.40	0.60	0.80	1.00
0.40	.2819	(.2948)	.2910	.2581
0.60	.2821	.2936	.2885	.2731
0.80	.2790	.2801	.2779	.2702
1.00	.1641	.1624	.1583	.1497

Table 27. E-measures at $\beta = 0.5$ on INSPEC

(a) P-NORM SCHEME

C_{AND}	C_{OR} 1.00	1.25	1.50	1.75	2.00
1.00	.7084	.7056	.7247	.7021	.7012
1.25	.7099	.7055	.7026	.7019	.7010
1.50	.7090	.7048	.7020	(.7003)	.7025
1.75	.7100	.7057	.7015	.7022	.7018
2.00	.7096	.7059	.7025	.7028	.7027

(b) PAICE SCHEME

C_{AND}	C_{OR} 0.70	0.80	0.90	0.95
0.70	.7033	.7065	.7065	.7079
0.80	(.7003)	.7036	.7062	.7073
0.90	.7245	.7045	.7080	.7089
0.95	.7017	.7033	.7069	.7079

(c) MMM SCHEME

C_{AND}	C_{OR} 0.40	0.60	0.80	1.00
0.40	.7245	.7380	.7210	.7531
0.60	.7220	(.7110)	.7146	.7290
0.80	.7332	.7296	.7316	.7412
1.00	.8009	.8040	.8038	.8111

Table 28. E-measures at $\beta = 1.0$ on INSPEC

(a) P-NORM SCHEME

C_{AND}	C_{OR} 1.00	1.25	1.50	1.75	2.00
1.00	.7196	.7171	.7287	.7137	.7128
1.25	.7210	.7170	.7140	.7138	.7123
1.50	.7203	.7164	.7137	(.7116)	.7142
1.75	.7212	.7173	.7133	.7137	.7136
2.00	.7210	.7179	.7145	.7147	.7145

(b) PAICE SCHEME

C_{AND}	C_{OR} 0.70	0.80	0.90	0.95
0.70	.7146	.7180	.7181	.7192
0.80	(.7122)	.7156	.7180	.7188
0.90	.7288	.7162	.7193	.7203
0.95	.7136	.7150	.7184	.7191

(c) MMM SCHEME

C_{AND}	C_{OR} 0.40	0.60	0.80	1.00
0.40	.7333	.7410	.7309	.7610
0.60	.7316	(.7216)	.7253	.7403
0.80	.7424	.7393	.7412	.7515
1.00	.8136	.8164	.8165	.8235

Table 29. E-measures at $\beta = 2.0$ on INSPEC

(a) P-NORM SCHEME

C_{AND}	C_{OR} 1.00	1.25	1.50	1.75	2.00
1.00	.7081	.7059	.7094	.7026	.7019
1.25	.7095	.7055	.7023	.7030	(.7012)
1.50	.7091	.7051	.7024	.6999	.7036
1.75	.7097	.7061	.7020	.7025	.7026
2.00	.7096	.7072	.7036	.7039	.7037

(b) PAICE SCHEME

C_{AND}	C_{OR} 0.70	0.80	0.90	0.95
0.70	.7031	.7066	.7068	.7075
0.80	(.7009)	.7045	.7068	.7075
0.90	.7094	.7048	.7077	.7093
0.95	.7023	.7037	.7073	.7077

(c) MMM SCHEME

C_{AND}	C_{OR} 0.40	0.60	0.80	1.00
0.40	.7211	.7211	.7189	.7487
0.60	.7192	(.7097)	.7145	.7311
0.80	.7308	.7283	.7301	.7420
1.00	.8119	.8147	.8155	.8219

Scheme	Best Precision	Rank	Best E-measure $\beta = 1.0$	Rank
P-NORM	.3093	1	.7116	1
PAICE	.3061	2	.7122	2
MMM	.2948	3	.7216	3

Table 30. Relative Ranks of Schemes by Average Precision and E-measure on INSPEC.

Scheme	Best E-measure $\beta = 0.5$	Rank	Best E-measure $\beta = 2.0$	Rank
P-NORM	.7003	1	.7012	2
PAICE	.7003	1	.7009	1
MMM	.7110	3	.7097	3

Table 31. Relative Ranks of Schemes by E-measure at β-levels 0.5 and 2.0 on INSPEC

Discussion

In terms of average precision and E-measure at β-level 1.0, the P-norm scheme is superior to the Paice scheme, which is in turn superior to MMM. However, at β-level 0.5, the E-measure ranks both P-norm and Paice the same, and MMM, the lowest of the three.

As on CISI at β-level 2.0, the E-measure actually ranks the Paice scheme higher than the P-norm. Though the performance measures for P-norm and for Paice are close to each other, with evidence from results obtained on two out of the three collections used, we may observe that on the average the Paice scheme performs slightly better in most cases than P-norm under the situation where recall is emphasized over precision.

Chapter 5

Conclusions

In this study, we have compared the retrieval effectiveness of P-norm, Paice, MMM and TIRS. As special cases of these retrieval schemes, we also examined the retrieval effectiveness of the classical fuzzy-set scheme. In addition, we have also obtained the average precision values for the standard Boolean scheme on all of the three collections.

Based on the results of our experimental runs on CISI, CACM and INSPEC collections, we have seen evidence for the following ranked order of the above retrieval schemes.

(1) P-norm
(2) Paice
(3) MMM
(4) TIRS
(5) Classical Fuzzy-set
(6) Standard Boolean

This ranking is based on average precision values and E-measures (with $\beta = 0.5$ or 1.0), and holds for the majority of the cases considered. However, at β-level 2.0, for which recall is deemed twice as important as precision, the E-measure seems to indicate that the Paice scheme is superior to the P-norm on both the CACM and INSPEC collections.

Both the Paice and the MMM schemes, which are variations of the classical fuzzy-set scheme, perform well on all collections. In terms of average precision, the Paice or MMM scheme shows at least an improvement of 46 percent on CISI, 84 percent on CACM and 96 percent on INSPEC over the classical fuzzy-set scheme. On the other hand, the P-norm scheme shows a 56 percent improvement over the classical fuzzy-set scheme on all collections. TIRS has an improvement of only 27 percent over the classical fuzzy-set scheme on CISI, and an improvement of 61 percent on CACM. Below we provide a summary of the percent improvements obtained by various schemes over the classical fuzzy-set scheme in terms of average precision. As seen from the summary, the P-norm, Paice or MMM scheme each gives a greater percent improvement in average precision over the classical fuzzy-set scheme on the INSPEC collection than on the CACM or CISI collection; and similarly, on CACM than CISI.

Percent Improvements Over Classical Fuzzy-set Scheme

	Test Collection		
Scheme	**CISI**	**CACM**	**INSPEC**
P-norm	56	86	106
Paice	54	84	104
MMM	46	89	96
TIRS	27	61	*NA*

The P-norm, Paice and MMM schemes generally give higher retrieval performance than the standard Boolean scheme. As shown by the summary below, except on the CISI collection, they in all cases attain more than 100 percent improvement over the standard Boolean scheme in terms of average precision.

Percent Improvements Over Standard Boolean Scheme

	Test Collection		
Scheme	**CISI**	**CACM**	**INSPEC**
P-norm	79	106	210
Paice	77	104	206
MMM	68	109	195
TIRS	46	71	*NA*

Our result that the average precision values for TIRS on CISI and CACM are relatively low as compared with the best average precision values of P-norm, Paice, and MMM, may be in part due to some shortcomings that we may have imposed on our experimental study. The use of the INNER PRODUCT function for computing the similarity between a document point and a query point in TIRS may not have been a good choice. Also, the idea of relevance ball radius of the TIRS scheme has not been fully made use of in our experimental study in the way implied by Cater & Kraft to maximize the recall-precision product.

Overall, the differences between the P-norm and the Paice schemes are marginal, and thus may not be statistically significant. The P-norm scheme performs better than the MMM scheme by only a small degree. The TIRS performance result is much lower than that of P-norm, Paice or MMM on both the CISI and CACM collections. The P-norm, Paice and MMM schemes each has a range of performance results (with variations in the associated parameter values), from which the best is selected for

comparison. However, the TIRS scheme has only a single performance result to be considered. Therefore, it may be that we have not been able to draw a very fair conclusion about TIRS in our comparison of various approaches for improving upon the standard Boolean retrieval.

The P-norm approach, being distance-based, has greater intuitive appeal than Paice, MMM or TIRS. But, its similarity computation method requires greater overhead than Paice or MMM. Irrespective of the function used for computing the similarity between a document and a query point, the TIRS approach, having to deal with the min-terms rather than just the typical Boolean query, is generally a difficult approach to implement and involves far greater computational costs than other schemes such as Paice or MMM. As we have seen from the experimental results, the P-norm scheme does not perform that much better than the MMM. Depending upon the actual cost of computations, the real overhead of the P-norm may far exceed that of the MMM. This may not justify its replacement with the MMM scheme where there is only a marginal loss in retrieval effectiveness.

Bibliography

[AHO 74] Aho, A., J. Hopcroft and J. Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Mass., 1974.

[BOOKS 80] Bookstein, Abraham. "Fuzzy Requests: An Approach to Weighted Boolean Searches." *Journal of The American Society for Information Science (JASIS),* July 1980, pp. 240-247.

[BOOKS 85] Bookstein, Abraham. "Probability and Fuzzy-Set Applications to Information Retrieval." *Annual Review of Information Science and Technology (ARIST),* edited by Martha E. Williams, Vol. 20, 1985, pp. 117-151.

[BUELL 85] Buell, Duncan A. "A Problem in Information Retrieval with Fuzzy Sets." *Journal of The American Society for Information Science (JASIS),* Nov. 1985, pp. 398-401.

[BUCK 85] Buckley, C. "Implementation of the SMART Information Retrieval System." TR 85-686, Cornell University, Dept. of Computer Science, May, 1985.

[CATER 86] Cater, Steven C. "TIRS: A Topological Information Retrieval System and the Topological Paradigm: A Unification of the Major Models of Information Retrieval." *Ph.d Dissertation,* LSU Department of Computer Science, Baton Rouge, LA, December 1986.

[CATER 87] Cater, Steven C. and Donald H. Kraft. "TIRS: A Topological Information Retrieval System Satisfying the Requirements of the Waller-Kraft Wish List." *Proc. of the 10th Annual Int'l ACM-SIGIR Conference on R&D in Information Retrieval,* June 1987, pp. 171-180.

[CROFT 83] Croft, W. B. "Experiments with Representation in a Document Retrieval System." *Information Technology: Research and Development,* Vol. 2, 1983, pp. 1-21.

[FOXE 83a] 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 Dissertation,* University Microfilms Int., Ann Arbor, MI, August 1983.

[FOXE 83b] Fox, Edward A. "Some Considerations for Implementing the SMART Information Retrieval System under UNIX." TR 83-560, Cornell University, Dept. of Computer Science, September, 1983.

[FOXE 86] Fox, Edward A. and Sharat Sharan. "A Comparison of Two Methods for Soft Boolean Interpretation in Information Retrieval." Technical Report TR-86-1, Virginia Tech, Department of Computer Science, January 1986.

[FOXE 87] Fox, Eward A. "A Development of the CODER System: A Testbed for Artificial Intelligence Methods in Information Retrieval." *Information Processing and Management,* 23(4), 1987, pp. 341-347.

[FOXE 88a] Fox, Eward A, Gary L. Nunn and Whay C. Lee. "Coefficients for Combining Concept Classes in a Collection." *Proc. of the 11th Annual Int'l ACM-SIGIR Conference on R&D in Information Retrieval* June 1988, pp. 291-307.

[FOXE 88b] Fox, Edward A. and Matthew B. Koll. "Practical Enhanced Boolean Retrieval: Experience with SMART and SIRE." *Information Processing and Management,* in press for 24(3), 1988.

[GEHAN 85] Gehani Narain. *C: An Advanced Introduction.* Computer Science Press, Rockville, MD, 1985.

[HARTER 86] Harter, Stephen P. Harter. *Online Information Retrieval: Concepts, Principles, and Techniques.* Academic Press, Inc. 1986.

[KANDEL 86] Kandel, Abraham. *Fuzzy Mathematical Techniques with Applications.* Addison-Wesley Publishing Co., 1986.

[MYERS 86] Myers, Raymond H. *Classical and Modern Regression with Applications.* PWS Publishers, Duxbury Press, 1986.

[PAICE 84] Paice, C. P. "Soft Evaluation of Boolean Search Queries in Information Retrieval Systems." *Information Technology, Res. Dev. Applications,* Vol. 3 No. 1, January 1984, pp. 33-42.

[RIJSB 79] van Rijsbergen, C. J. *Information Retrieval.* Second Edition, Butterworth & Co (Publishers) Ltd, 88 Kingway, London WC2B6AB, UK. 1979.

[SAS 85] SAS Institute Inc. *SAS User's Guide: Statistics Version 5 Edition.* Cary, North Carolina, USA 1985.

[SALT 83a] Salton, Gerald, Edward A. Fox and Harry Wu. "Extended Boolean Information Retrieval." *Communications of the ACM,* Vol. 26, No. 12, pp. 1022-1036, Dec 1983.

[SALT 83b] Salton, Gerald and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Computer Science Series, McGraw-Hill, Inc., 1983.

[SALT 86] Salton, Gerald. "Another Look at Automatic Text-Retrieval Systems." *Comm. of the ACM,* Vol 29, No. 7, July 1986, pp. 648-656.

[SMITH 87] Smithson, Michael. *Fuzzy Set Analysis for Behavioral and Social Sciences.* Recent Research in Psychology Series, Springer-Verlag, NY, 1987.

[SWETS 69] Swets, J. A. "Effectiveness of Information Retrieval Methods." *American Documentation,* Vol. 20, No. 1, January 1969, pp. 72-89.

[TONG 85] Tong, Richard M. and Daniel Shapiro. "Experimental Investigations of Uncertainty in a Rule-Based System for Information Retrieval." *Int'l Journal Man-Machine Studies,* Vol. 22, 1985, pp. 265-282.

[TONG 87] Tong, R. M., R. A. Clifford, G. J. Crowe, and P. R. Douglas. "Conceptual Legal Document Retrieval Using the RUBRIC System." *Proc., First Int'l Conf. on AI and Law (ACM),* Boston, MA, May 1987.

[WONG 86] Wong S. K. M., W. Ziarko, V. V. Raghavan and P. C. N. Wong. "On Extending the Vector Space Model for Boolean Query Processing." *Proc. of the 9th Annual Int'l ACM-SIGIR Conference on R&D in Information retrieval* June 1986, pp. 171-185.

[YAGER 80] Yager, R. Y. "On A General Class of Fuzzy Connectives." *Fuzzy Sets and Systems,* North-Holland Publishing Co., Vol. 4. No. 3, pp. 235-242.

[YAGER 87] Yager, R. Y. "A Note on Weighted Queries in Information Retrieval Systems." *Journal of The American Society for Information Science (JASIS),* January 1987, pp. 240-247.

[ZADEH 65] Zadeh, L. A. "Fuzzy Sets." Information and Control, Vol. 8, pp. 338-353.

The two page vita has been removed from the scanned document. Page 1 of 2

The two page vita has been removed from the scanned document. Page 2 of 2