ABSTRACT OF THE DISSERTATION

AN INVESTIGATION OF THE CONDITIONS FOR EFFECTIVE DATA FUSION IN INFORMATION RETREIVAL

By KWONG BOR NG

Dissertation Director:
Professor Paul B. Kantor

Effective automation of the information retrieval task has long been an active area of research, leading to sophisticated retrieval models. With many IR schemes available, researchers have begun to investigate the benefits of combining the results of different IR schemes to improve performance, a process called “data fusion”. There are many successful data fusion experiments reported in IR literature, but there are also experiments in which the same fusion rules did not work. What is needed is a theory to tell *a priori* when one should use data fusion methods.

In this thesis we propose two conditions for effective data fusion: (1) The condition of efficacy, which states that fusion of two IR schemes with comparable performance tends to be effective; (2) The condition of dissimilarity, which states that fusion of two dissimilar IR schemes tends to be effective.

I use the IR systems participating in the TREC 4 routing task for training a model to predict the effectiveness of data fusion and the IR systems participating in the TREC 5 routing task to test that model. The model asks, “when will fusion perform better than an oracle who uses the best scheme for each pair?” I apply various statistical techniques to fit the model to the training data and use the receiver operating characteristic curve of signal detection theory to represent the power of the resulting models. Two predictive variables have been identified which predict the sign of the effectiveness of fusion of two IR schemes: (1) the ratio of precisions of the two schemes and (2) the normalized dissimilarity of the two schemes.

After training, the prediction methods predict the sign of the effectiveness of data fusion between schemes in the testing set much better than chance. For example, applying the trained models to the testing data set, when the models can predict correctly about 70% of the positive cases, they only incorrectly predict about 30% of the negative cases. The results of the experiments support the two proposed conditions for effective data fusion.

ACKNOWLEDGEMENTS

I would not be able to finish this thesis in time if it were not Paul Kantor, chair of my committee, who kept on encouraging, guiding, and helping me for my four years in the Ph.D. program. At every step of the way, he provided me with constant support, constructive criticism, and valuable suggestions. In this regard I have been more than lucky and I could not have asked for better.

Thanks to all my committee members. Their hard work and keen insights have helped shape this dissertation in important ways. Thanks to my wife, Sinyung Lo, for her warm understanding and love. Thanks to my four months old baby, Kali, for her lovely crying and smiling. Thanks also go to Soyeon Park, for her helping and assisting me in many instances. Finally, I thank my friends in Alexandria Project Laboratory, working together with them made writing this thesis a pleasure.

Dedication

To my wife Sinyung and my baby Kali.

Table of Content

LIST OF TABLES

LIST OF FIGURES

Chapter 1:

Introduction

1.1 Information Retrieval Systems and Schemes

The task of an information retrieval (IR) system is to select from a collection of information objects (e.g., documents), those that will be of interest to a user. To facilitate effective searching, an IR system uses a representation describing various attributes of all the information objects in the collection. When a user searches for documents she, herself or through an intermediary, has to formulate her information need in a format prescribed by the IR system. This format called “the query” can be free text, or a highly structured Boolean combination, or some other sophisticated expression, depending on the IR system. The IR system then matches the query with the document representations to estimate the relevancy of any document in the collection.

After examining the output of the system, the user can adjust her query (which is called "relevance feedback" to the system) to invoke another matching process. The automatic part of IR, then, can be considered as comprising three basic components:

1. document representation;
2. query formulation;
3. computation for matching of document representation and query formulation

When we fix the details of the representation, the formulation, and the computation, we have defined an IR scheme.

We use IR “system” and IR “scheme” to describe two different concepts (Kantor, 1994a; Ng *et al.*, 1997). IR system refers to the physical implementation of an IR algorithm, which can have various operational modes or various settings of parameters. Therefore, the same IR system may be used to execute different IR schemes by adjusting parameters (e.g., changing term weighting functions, or switching from ranking retrieval mode to set retrieval mode, or softening the Boolean operations, etc.) which will give rise to different outputs.

Effective automation of the information retrieval task has long been an active area of research, leading to sophisticated retrieval models for representing the information content in documents and queries and computing the similarity between the two (Kantor, 1994b). With many various IR schemes available, researchers have begun to investigate the benefits of using multiple IR schemes to improve performance.

1.2 Data Fusion

Data fusion is a relatively new concept. Generally it is an approach which combines data, evidence, or decisions coming from (or based on) various sources, of different natures, about the same set of objects, in order to increase the quality of decision making under uncertainty about the objects (Varshney, 1997).

Generally there are three level of data fusion: primary data level, attribute level, and decision level (Kantor, 1994a):

1. On the primary data level, all the information available to the detecting systems is considered together in the fusion process to make an overall estimate;
2. On the attribute level, primary signals detected from the objects by the detecting systems are processed into a set of specific attributes, and decisions about the objects are made according to an optimal decision rule based on all such attributes;
3. On the decision level, each detecting system individually makes its own partial decision about the objects, using its own data, and according to its own criteria, and a final decision is made based on these partial decisions.

Data fusion often involves multiple imperfect sensors and each of the sensors contributes its own estimation to the final decision (Viswanathan and Varshney, 1997). There must be a fusion rule or fusion scheme to combine the estimates from different sensors. For example, given a group of detectable objects, when the estimate of each sensor is reported in a single score, the fusion rule can be the sum of normalized scores; when the estimate of each sensor is reported in a format of binary decision (e.g., A and ~A), the fusion rule can be union, or intersection, of the result sets.

1.3 Data Fusion in IR

The idea of data fusion has also been explored in the area of information retrieval (IR.) The objects in question are documents in a collection and usually the decisions to be made concern their probability of relevance with respect to a particular information need or topic. Before reviewing the IR data fusion literature, we clarify some distinctions.

1. 3.1 Data Fusion Vs Database Merging and Collection Fusion

In the IR literature, the phrase “data fusion” sometimes is also used to refer to combination of search results from various databases or different document collections (in contrast to a single database or collection). This is more appropriately referred to as “database merging” or “collection fusion” (e.g., Voorhees *et al.*, 1995). In our work, “data fusion” only refers to decision making associated with multiple IR schemes and one single document collection.

1.3.2 Combination of Data Vs Combination of Queries

Another possible confusion about data fusion in IR is between "combination of queries" and "combination of data". Some researchers use the phrase "combination of queries" to refer to a combination process which takes place before queries are sent to an IR system. This approach seeks to combine different queries, formulated by different experts for the same information problem, into one combined query, which may have better retrieval performance than single interpretations (e.g., Gomez *et al*, 1988).

"Combination of data", on the other hand, takes place after outputs are produced (or after intermediate decision are made) by different IR schemes. Thus, it seeks to combine the attributes or decisions assigned to an information object (i.e., document) by several IR schemes under a fusion rule. Fusion is most powerful when the fusion rule is more effective than the best rule that can be obtained from the attributes (or decisions) assigned by any individual IR scheme.

Sometimes the fusion rule is based on pseudo-attributes instead of attributes, when the value computed by the matching scheme is determined not only by the properties of the information object itself, but also by properties of other information objects in the corpus. An example is the inverse document frequency, which may not be a constant if the collection is being renewed regularly.

In IR, combination of data has received more attention than combination of queries. However, this does not mean combination of queries is without theoretical interest. For example, in a vector space model, weighted summing of different vector representations into one combined query vector can be quite complicated, and there are not many theoretical studies of this aspect. We note that the difference between combination of queries and combination of data is not clear cut when different queries are produced automatically without human intervention, and thus might be regarded as intermediate data in the retrieval process.

In this thesis, data fusion only refers to “combination of data”, and not “combination of queries”. The essential features of this data fusion are:

1. There are multiple IR schemes;, and,
2. Combination takes place after attributes (or decision) have been assigned by each of the multiple schemes.

In particular, combination of documents retrieved by “different queries” is considered as “combination of data”, not “combination of queries”, because the data fusion takes place after documents are retrieved, not before the input (combined) query is sent to the IR system.

1.3.3 Data Fusion Approaches in IR

In IR, data fusion does not necessarily employ different IR systems, but only different IR schemes. Corresponding to the three basic components of IR, i.e., document representation, query formulation, and matching computation, research in IR has considered three approaches to data fusion:

1. multiple queries (e.g., Saracevic and Kantor, 1988);
2. multiple document representations (e.g., Katzer *et al*, 1982); and
3. multiple retrieval techniques (e.g., Belkin, Kantor, Fox & Shaw, 1995; Lee, 1997).

In each approach, some kind of combination, in contrast to using just one single query, one single document representation or one single retrieval technique, can improve retrieval effectiveness.

Usually IR schemes produce either set outputs or ranked outputs. In the cases of set output (Boolean scheme), each IR scheme produce a set of documents. Every element in the output set has the same status with respect to the estimated probability of relevancy.

Fusion rule for Boolean schemes can take the form of union of all sets, intersection of all sets, or number of occurrence of a document in all sets. For example, let the outputs of three Boolean IR schemes ***A***, ***B***, ***C*** be:

A = (a, b, c, d, e)
B = (b, c, d, e, f)
C = (c, d, e, f, g)

If the fusion rule is union of the output sets, then the result of fusion will contain the seven documents in the three sets, i.e., ***(a, b, c, d, e, f, g)***; if the fusion rule is intersection of the output sets, then the result of the fusion will contain only three documents, i.e., ***(c, d, e)***; if the fusion rule is, only those documents which have appeared twice or more in the output sets will be included in the fused set, then the result of fusion will be: ***(b, c, d, e, f)***.

Which fusion rule to prefer depends on the value of correctly retrieving a relevant document and the cost of mistakenly retrieving an irrelevant document. In this kind of data fusion, each IR scheme individually makes its own judgement about the relevancy of all documents in the collection according to its own criteria. The final decision is made based on these individual judgements. Therefore, Boolean output data fusion belongs to the decision level data fusion.

In the case of ranked outputs, one of the most common fusion rules is sum of normalized relevancy scores. Let the top five documents of a ranked output IR schemes be:

a	***6.0***
b	***3.6***
c	***3.0***
d	***2.4***
e	***1.0***

where the first column is document identification code and the second column is relevancy score assigned by the IR scheme to the document.

Let the top five documents of another ranked output IR scheme be

c	***900***
d	***600***
g	***50***
a	***-20***
f	***-100***

Since different IR schemes may have different scale of relevancy scores, in order to make the relevancy scores comparable to each other, we need a normalization procedure. The most common normalization procedure is :

$$\boxed{\textit{Normalized relevancy score} = \frac{\textit{Relevancy score} - \textit{Minimum relevancy score}}{\textit{Maximum relevancy score} - \textit{Minimum relevancy score}}}$$

Using the above normalization equation, we can compute the normalized score of each document of the two output list and re-rank them according to the sum of normalized scores. For the first list, the normalized relevancy scores of the first list are:

a	***1.00***
b	***0.52***
c	***0.40***
d	***0.28***
e	***0.00***

The normalized relevancy scores of the second list are:

c	***1.00***
d	***0.70***
g	***0.15***
a	***0.08***
f	***0.00***

The result of the data fusion by sum of normalized scores are:

c	***1.00 + 0.40 = 1.40***

a	***1.00 + 0.08 = 1.08***
d	***0.28 + 0.70 = 0.98***
b	***0.52 + 0.00 = 0.52***
g	***0.15 + 0.00 = 0.15***
e	***0.00 + 0.00 = 0.00***
f	***0.00 + 0.00 = 0.00***

If the cutoff point is still top 5 documents, then the result of data fusion is the following ranked list:

c > a > d > b > g

where " ***x > y*** " means ***x*** is above ***y*** in the rank list. We can see that the above computation assumes that the normalized relevancy score is additive across different IR schemes. It assumes the difference between 10 and 20 from one list is exactly the same as the difference between 10,010 and 10,020 from the other list if both of the two lists have the same maximum and minimum relevancy scores above the cutoff point, without considering the possibility of outliners.

In principle, the minimum score in the normalization process should be the minimum score that an IR scheme will assign to the document of the collection. However, in practice, usually people use the score of the document just before the cutoff point as minimum score (i.e., minimum score of the output lists). Therefore, the choice of the cutoff point may affect the result of the data fusion. For example, using the top five documents as cutoff point, in the above data fusion, document ***d*** is higher than document ***b*** in the fused list, but if we just use top three documents of the output lists for data fusion, document ***b*** will be higher than ***d*** in the fused list, it is because the minimum relevancy score have changed.

In the above data fusion, fusion was not performed directly on the level of relevancy scores assigned to the documents, but on the level of normalized relevancy scores. The relevancy scores are assigned to the documents by each IR scheme working independently based on its own algorithm. Therefore, if the fusion is performed directly on the level of relevancy scores, it is attribute level data fusion. When the data fusion is performed on the level of normalized relevancy scores, we may consider it as pseudo-attribute level data fusion.

The data fusion of ranked output IR schemes can be performed on other level. For example, if the fusion is not based on the relevancy scores or normalized relevancy scores, but on the rank order of the documents in the output list, then it is decision level data fusion.

The following will review some of the IR experiments in which data fusion does, or might improve performance.

1.4 Brief Reviews of Some Successful Data Fusion Experiments

Empirically, data fusion in IR works well. Many data fusion experiments have been done. While almost all of them have shown positive results, exceptions include Ng and Kantor (1996) and Ng *et al* (1997).

On the positive side, Turtle and Croft (1991) developed a multi-layer computational IR model, which can combine different document representations and different versions of a query in a consistent probabilistic framework. They implemented an information retrieval system, Inquery, based on the model, and demonstrated that combining the outputs of “free text based” formulation and Boolean formulation of the same information problem increases retrieval effectiveness.

Belkin *et al* (1993) showed that combination of different Boolean query formulations of the same information problem could lead to improvements of retrieval effectiveness.

Fox and Shaw (1994) examined various methods for combining multiple retrieval runs. They found that a method that used the sum of normalized similarity scores of five different IR schemes gave improvements over the best retrieval run.

In a joint report from Rutgers University and Virginia Tech, Belkin, Kantor, Fox and Shaw (1995) investigated the effect on retrieval performance of combination of multiple representations of TREC-2 (The Second Text REtrieval Conference) topics. Rutgers University used the Inquery system and Virginia Tech used a modified version of SMART system. Both Rutgers and Virginia Tech found that the best method of combination often led to results better than any individual retrieved set without combination.

Using the information retrieval system SMART, Lee (1995) found that different weighting schemes (with the same queries, the same document representation and the same retrieval algorithm) retrieved different “types of documents” (in terms of document length). He showed that significant improvements could be obtained by combining the retrieval results from IR schemes which differed only in the weighting of query terms.

The research results described above clearly suggest that retrieval performance can be greatly improved by using multiple IR schemes and applying data fusion to their results. It is as if each IR scheme somehow contributes its own estimates of which documents are likely to be relevant to the user's information problem, and the combined set is typically more valuable than any single IR scheme’s estimates. However, this scattered empirical success still lacks a full theoretical foundation. In particular, there seems no way to predict, *a priori*, which IR schemes to combine, or what fusion method to use, to obtain the observed improvement in retrieval effectiveness.

1.5 Organization of This Thesis

Since data fusion can improve retrieval performance without developing new retrieval principles or algorithms, and sometimes even without using another IR system, it is a potentially powerful technique that deserves further study. The organization of this thesis is as follows. In the next chapter, Chapter 2, we discuss the theoretical foundation of data fusion. In Chapter 3, we suggest two conditions for effective data fusion in IR. To test the validity of these two proposed conditions, We need a scale to measure the distance between two IR schemes outputs. We offer a rigorous definition and a set of algorithms for such a scale in Chapter 4. In Chapter 5, we discuss our experimental design, definitions of various variables, and the fusion rule of our data fusion experiments. In Chapter 6 to Chapter 10, we investigate the predicting power of different variables and different statistical methods. We train our predicting methods (which predict whether the fusion of two IR schemes will have positive effectiveness) using one set of data and then test the findings in another set of data. We report and discuss the result of the training and testing in Chapter 11 and 12. Chapter 13 is the conclusion of this thesis.

Chapter 2:

Theoretical Foundation of Data Fusion in IR

Since most of the IR data fusion experiments based their evaluations on recall and precision, we propose to classify data fusion schemes in terms of their impacts on recall and/or precision. Some theoretical research seeks to explain the power of data fusion by arguing, implicitly or explicitly, that data fusion can increase recall and/or precision. We expand this analysis in the next two sections. Then we will discuss four heuristics suggested by Kantor (1994a, 1995).

2.1 The Improved Recall Argument

In IR, recall is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection:

$$\boxed{\textbf{\textit{Recall}} = \frac{\textbf{\textit{number of relevant documents in retrieved set}}}{\textbf{\textit{number of relevant documents in collection}}}}$$

For ranked output, “number of relevant documents retrieved” can be modified to “number of relevant document before the cutoff point”.

If we call the cutoff point ***n***, the number of relevant documents retrieved, denoted by ***g***, is a function of ***n***, that is, ***g(n)***. Of course, ***g(n)*** is a non-decreasing function of ***n***.

Some IR researchers (e.g. Belkin, Cool, Croft, & Callan 1993) suggest that the combination of documents retrieved by different IR schemes will contain more relevant documents than any single retrieval scheme. In other words, they suggest that data fusion in IR can improve performance because it can improve recall -- even though some of them may not have used the word “recall”, or may even disagree with using “recall” as a performance measure for IR system.

In my opinion, all justifications of data fusion which concentrate on total number of relevant documents retrieved, without considering how “optimal” fusion can be achieved, are implicitly following this line of improved recall argument.

For example, there are analyses (e.g. Gomez, Lochbaum, & Landauer, 1990; Belkin, Cool, Croft, & Callan, 1993) suggesting that different query formulations, or document representations, or retrieval algorithms, can capture different aspects of a complex and uncertain information problem, so that they will retrieve different “kinds” of relevant documents. When the analysis concludes that the combination of the results will have more relevant documents, without considering the fact the combined set may also have even more non-relevant documents than each individual set, it is an example of the “improved recall” argument.

Since precision is one of the most important evaluation measures for IR system performance, it is almost impossible to ignore precision and just concentrate on the recall when talking about IR performance improvement. Therefore, it is not surprising that we find no pure example of the improved recall argument. The following formulation of the improved recalled argument is only for sake of analysis and discussion.

Note that in the Boolean set retrieval environment, improved precision does not necessarily leads to improved recall, but in the rank-based thresholding retrieval environment, improving recall with the same threshold after combination will also improve precision. In the latter case, the improved recall argument alone may not be enough to explain the effectiveness of data fusion. The following discussion will focus on the Boolean set retrieval environment to demonstrate the power and weakness of the improved recall argument.

The improved recall argument claims: the recall of the combined set of documents retrieved by different IR schemes is likely to be higher than that of any single set of retrieved documents.

Empirically there is strong evidence that the improved recall argument is valid, as it has been observed that different IR schemes often retrieve different documents. For example, Katzer *et al* (1982) found that, for the same information topic, when documents were retrieved by different users, or by the same user using different representation tools (i.e., controlled vocabulary indexing versus free-text vocabularies), there was little overlap among document sets. They also discovered that, for the INSPEC database, using various document representations (e.g., title., abstract) for retrieval gave similar levels of retrieval effectiveness, but retrieved quite different sets of documents. In the first Text REtrieval Conference (TREC-1), the IR systems participating in the conferences retrieved substantially different sets of documents, even though most systems performed at more or less the same level (Harman, 1993). Since different IR schemes may retrieve different documents, it seems reasonable that some "putting together" of the retrieved documents will contain more relevant documents, i.e., recall will be improved.

We can state the improved recall argument as follows:

> *The recall of the union of sets of retrieved documents is larger than or equal to the recall of any single set of retrieved documents.*

In this form, it must be valid. By combining retrieval sets we can only have more, at least as many, and never fewer, relevant documents. For the improved recall argument to be not only valid but also practical, there are two conditions that must be satisfied:

1. There is significant difference between retrieved sets;
2. The difference between retrieved sets must be due not only to the difference in the non-relevant documents in both sets, but also to a

difference of relevant documents in both sets. In other words, each set should contain a significant number of distinct relevant documents.

For example, let there be two retrieval sets ***A*** and ***B***.

$$\boldsymbol{A = (r_1, r_2, n_1, n_2, n_3)}$$
$$\boldsymbol{B = (r_1, r_2, n_2, n_4, n_5, n_6)}$$

where $\boldsymbol{r_i}$ stands for the i^th^ relevant document and $\boldsymbol{n_i}$ stands for the i^th^ non-relevant document. The difference between ***A*** and ***B*** is large in the sense that, size of ***A*** **=** 5, size of ***B*** **=** 6, but size of $\boldsymbol{(A \cap B)}$ is just 3. However, combination of ***A*** and ***B*** will not give higher recall. The set $\boldsymbol{A \cup B = (r_1, r_2, n_1, n_2, n_3, n_4, n_5, n_6)}$, has more non-relevant documents in the combination, but not more relevant documents. This is because the difference between ***A*** and ***B*** only involves non-relevant documents, not relevant documents. There are no distinct relevant documents in ***B***.

As shown in the above, empirically there were experiments demonstrating that the first condition could be satisfied, but for the second condition, it seems that there is no reported data fusion experiment demonstrating that distinct relevant documents in retrieval sets contributed significantly to the total difference. In fact, there are counter examples that demonstrate the opposite (which will be discussed below).

Although there was a suggestion (only a suggestion, not an experiment or proof) that different retrieval schemes would retrieve different documents, relevant as well as non-relevant (e.g., Belkin, Cool, Croft, & Callan, 1993), there were observations which did not agree with this suggestion (e.g., Lee 1997). In those observations, even though the second condition could not be satisfied, data fusion could still give better performance. It was argued then that data fusion worked not because of improved recall, but because of improved precision.

2.2 The Improved Precision Argument

Data fusion does not always increase recall. Turtle and Croft (1991) observed that combined outputs of different IR schemes did not give a greater number of relevant documents. They combined the results of Boolean search and probabilistic search using a weighted sum link matrix (a matrix of normalized sum of beliefs or evidences for the importance of the scheme) and found improvement in performance.

The authors originally thought that at least part of the performance improvement of data fusion was because the two IR schemes were retrieving different relevant documents. If so, the combined set would contain more relevant documents than individual sets. However, they discovered that "the documents retrieved by the Boolean queries are a subset of those retrieved by the corresponding probabilistic query." The authors pointed out that this might be because of the specific strategy the experts employed to construct the Boolean queries. It appeared that the objective in creating the Boolean queries was to capture the structural relations between concepts (structured by the Boolean operators AND, OR, and NOT) present in the natural language descriptions of the information need, not to produce high recall queries, and therefore all search terms were from the original descriptions. According to the authors, if trained searchers were asked to produce high recall Boolean queries from the natural language descriptions, they would generally use their knowledge of the subject domain and indexing practice to expand the set of terms to include synonymous and related terms, and, “it is likely that these enhanced searches would retrieve relevant documents not found by the probabilistic query”. In any case, the power of data fusion reported here had nothing to do with improved recall because all the relevant documents retrieved were present in the document set retrieved by the probabilistic queries. Data fusion improved performance because the fusion rule (weighted sum of normalized belief scores) produced a better ranking than the rankings produced by the probabilistic or Boolean queries alone.

Since data fusion does not necessary improve recall, but still can be effective, in at least some cases, the effectiveness of data fusion may be explained by arguing that what data fusion improves is not recall, but precision. In IR, precision is defined as the ratio between the number of relevant documents retrieved and the total number of documents retrieved:

$$\boxed{\textbf{\textit{Precision}} = \frac{\textbf{\textit{number of relevant documents retrieved}}}{\textbf{\textit{number of documents retrieved}}}}$$

Again, for ranked output, “number of relevant documents retrieved” can be modified to “number of relevant document before the cutoff point”. Thus the precision becomes a function, $\boldsymbol{g(n) / n}$ **,** of the cutoff point at $\boldsymbol{n}$ documents retrieved.

There were observations (e.g., Saracevic & Kantor, 1988; Lee, 1997) suggesting that relevant documents are more likely to be retrieved by different IR schemes than are non-relevant documents. In other words, different IR schemes will retrieve similar sets of relevant documents but different sets of non-relevant documents. This argument contradicts the improved recall argument, but it also makes sense because different IR schemes are much more likely to agree correctly than to agree in error. This property, if true, can be used to improve precision by any fusion method which gives more weight to common documents than non-common documents. In fact, it has been accepted as truth or taken for granted in some parts of the IR literature (e.g., Voorhees and Harman 1997). We call this line of reasoning “the improved precision argument”.

In theory, using the improved precision argument for data fusion can be quite simple, or quite complicated. It depends on whether or not we consider the attributes (or decisions) assigned by different IR schemes to each document to be stochastically identical and independent distribution (i.e., the sources of difference are not deterministic). For example, in a set retrieval environment (i.e., binary decision: relevant or non-relevant), if we suppose that the decisions of multiple IR schemes are stochastically identical and independent, we may use a simple symmetrical fusion rule (e.g., how many times a document appears in the retrieved sets) for data fusion. Although the assumption of stochastically identical and independent distribution appears not to have been tested, researchers tend to use symmetrical rules for data fusion.

There is some research that lends support to the improved precision argument. For example, Saracevic and Kantor (Saracevic & Kantor 1988), in a study of cognitive decisions and “human -- IR systems” interaction involved in information seeking and retrieving, asked different experts to construct Boolean queries based on the same description of information problem and then submitted the queries to an online information retrieval system (Dialog). They found that the overlap in selection of search terms by different searchers was low, and these different query formulations retrieved many more different documents than common documents. In 469 cases (58.6%) of all documents retrieved and in 471 cases (58.9%) of relevant or partially relevant documents retrieved the degree of overlap was between 0% and 5%. They also observed that the odds that a document is relevant increase monotonically with the number of different retrieval sets in which it appears.

This observation suggests that common documents are more likely to be relevant documents than are distinct or unique documents in the retrieval sets. If the combining method is designed to favor the documents retrieved by more retrieval runs, the combined set will have higher precision. On this observation, one may drop the improved recall argument and still be able to explain (at least partly) the power of data fusion when the fusion method gives more weight to common documents than distinct documents in the retrieval sets.

Lee explored the power of data fusion along this line. In his investigation, Lee (1997) explicitly questioned the validity of the improved recall argument. He investigated whether combined retrieval sets really contain more relevant documents than each individual retrieval set by examining the overlap among relevant documents retrieved by different runs and the overlap among non-relevant documents retrieved by different runs. Lee selected six retrieval results from the TREC-3 (Harman, 1995) adhoc task and calculated the pair-wise overlap ratio defined as:

$$\boxed{\textbf{\textit{Rel doc overlap ratio}} = \frac{\textbf{\textit{no. of common rel docs}} \times \textbf{\textit{2}}}{\textbf{\textit{sum of no. of rel docs retrieved in two runs}}}}$$

$$\boxed{\textbf{\textit{Non-rel doc overlap ratio}} = \frac{\textbf{\textit{no. of common non-rel docs}} \times \textbf{\textit{2}}}{\textbf{\textit{sum of no. of non-rel docs retrieved in two runs}}}}$$

He observed that the relevant document overlap ratio is more than twice the non-relevant document overlap ratio. The average of 15 pairwise comparisons is 0.7822, and 0.3519, for relevant document overlap ratio and non-rlelvant document overlap ratio respectively (computed based on data from Lee's paper).

He called this “the unequal overlap property” and argued that data fusion schemes should be based on the unequal overlap property.

While these observations support the assertion that common documents in different retrieval sets are more likely to be relevant than non-common documents, it has not been demonstrated that generally distinct relevant documents play no significant role in data fusion. If we combine the retrieval sets by only including common documents, precision may be improved a lot, but it is necessary that the recall will drop, perhaps significantly. In addition, in a ranked output environment, overlap of retrieved sets is matter of degree, depending on the cutting point or threshold of retrieval.

2.3 Four Heuristics offering Theoretical Foundations of Data Fusion

One of the main issues in data fusion is to decide how data are to be combined to generate an overall ranking. It is unclear how the best subset of the combination of all sets of retrieved documents can be selected and ranked. Fusion strategies which are “union-like” reflect the improved recall argument, fusion strategies which are “intersection-like” reflect the improved precision argument, but we don’t have a well established theoretical framework as the foundation of the improved recall argument or the improved precision argument. Kantor (1994a, 1995) has suggested four possible approaches to theoretical justification of data fusion in IR:

1. Bayesian argument: If different kinds of evidence support the same

preference, then their simultaneous appearance supports it even more. Therefore, the more sources of evidence are available, the more accurate the judgment of the probability of relevance of a document to the query will be. This argument can be considered as another formulation of the signal processing argument (see also Belkin, Kantor, Fox and Shaw, 1995; Varshney, 1997) in which a combination of independent signals can be expected to have better operating characteristics.

2. Imprecision of language argument: Language is imprecise. Different characterization of the same information problem can only be considered as imprecise descriptions scattering around the precise or perfect description in some abstract "concept space". Therefore, overlap of different imprecise descriptions is more likely to "contain" a precise description than is any single imprecise description.
3. Vector distribution argument: This may be considered a translation of imprecision of language argument into geometric model. The rationale is, if several representative vectors are chosen from a random distribution, their mean is more likely to be closer to the center of the distribution than any one of them. This idea is pursued in detail in Kantor (1995), and applied to some empirical results.
4. Expansion of the parameter space argument: When the possible choices for an optimization problem expand, the optimum may get better, but can never get worse. The paradox is, when there are more choices, the probability of

getting the optimum may decrease. The question becomes: how can we choose wisely to eliminate the non-optimal choices?

All the above arguments are heuristic. They are not proofs. More work is needed to develop a general theory of data fusion in IR.

Chapter 3:

Conditions for Effective Data Fusion

No well developed general theory of data fusion in IR is available yet. Part of the reasons may be that we don’t have enough studies of the basic phenomena of data fusion. We know that data fusion does not always give better performance. For example, Fox and Shaw (1994), besides successful experiments, reported some unsuccessful experiments in which data fusion did not improve performance. Ng and Kantor (1996) and Ng *et al* (1997) also reported their unsuccessful data fusion experiments. However, it appears that there are no conclusive discussions of why data fusion sometimes fails. When we don’t have a clear understanding of the necessary conditions for data fusion, even unsuccessful experiments may tell us more about the possible directions to investigate.

3.1 Unsuccessful Data Fusion Experiments

According to the improved precision argument, different IR schemes will retrieve similar sets of relevant documents but different sets of non-relevant documents, i.e., two different IR schemes are much more likely to agree correctly than to agree in error. For two schemes to agree on a document's relevance, one interpretation is that the suitably normalized relevance scores assigned to it are approximately equal. Geometrically this means that the document is placed near the principal diagonal of the positive quadrant in the ordinate systems of normalized relevance scores of the two IR schemes.

To test this idea, Ng and Kantor (1996) ranked documents in decreasing order of variance of the normalized relevance scores (with respect to a description of the same information need) assigned to them by two different IR schemes. The first scheme was called the "word-based approach". It was a ranking retrieval method with word frequency statistics and the cosine rule applied to estimate the similarity between the document vector and query vector. The second scheme was called the "5-gram-based approach". It employed the same strategy except that instead of using the word directly, it used overlapping 5-grams to represent the collection text and the query text. The performance of this variance-based fusion scheme was worse.

In another experiment reported in the same paper, data fusion using the sum of normalized relevance scores of the two IR schemes also did not improve performance. We suggested that both results might be due to the fact that the two IR schemes were operationally too similar to each other. Not only did both schemes retrieve similar set of documents, they also assigned more or less the same ranks for the documents they retrieved. This means that the two schemes are not providing independent information about the retrievable documents. Therefore combining data by summing normalized relevance scores did not give more information about the documents.

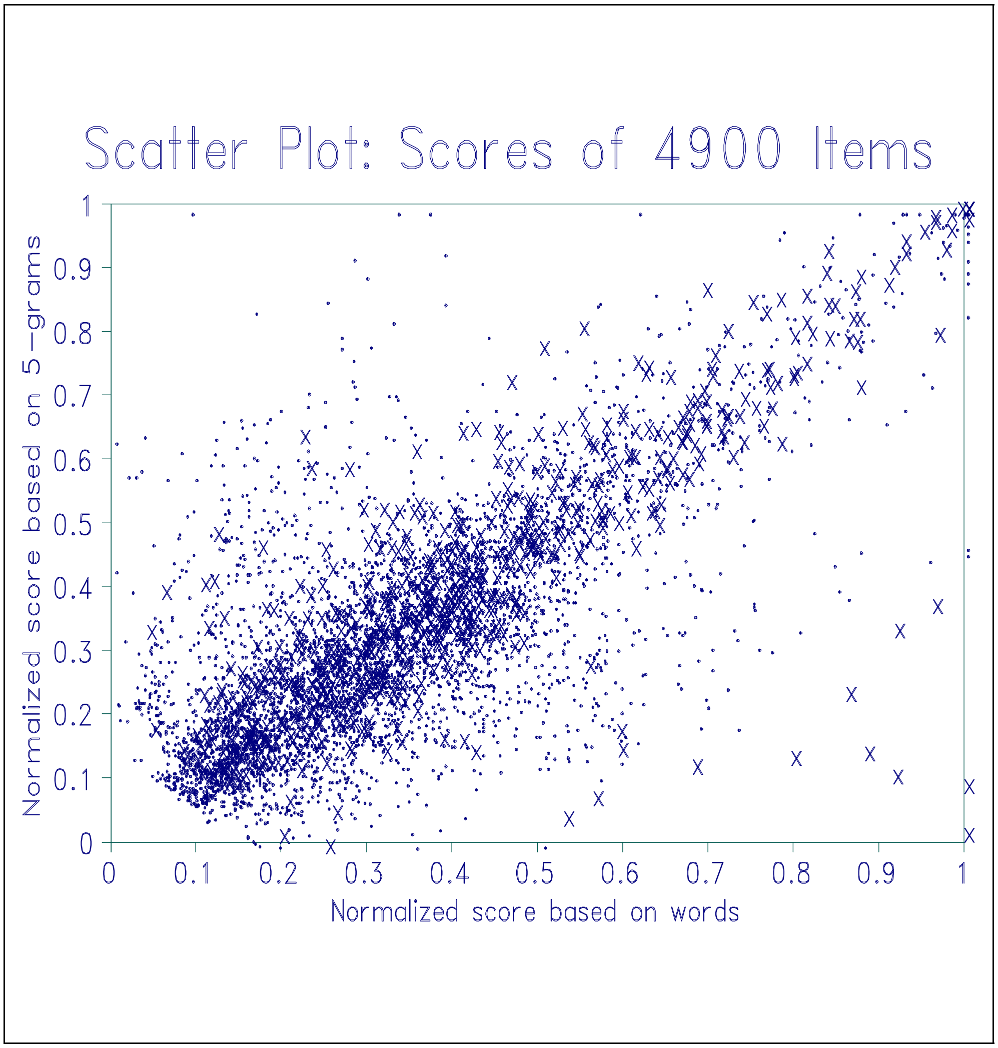

Figure 3.1: Situation in which data fusion will not be expected to work well. Scatter plot of the scores assigned to the top 100 documents for each of 49 adhoc topics by the word-based approach and the 5-gram-based approach. , "x" represents a relevant document and "." represents a non-relevant document.

This similarity might have also affected the variance experiment because the variation of the variances might be too small for any practical purpose. *Figure 3.1* (previous page) is the scatter plot of the normalized scores of the top 100 documents retrieved by both schemes for 49 topics. The points are quite concentrated along the diagonal. This is precisely the situation in which data fusion will not be expected to work.

If two IR schemes are operationally identical, no fusion rule can give a better performance because all the information that can be found in the outputs are the same.

The above interpretation found some support from Fox and Shaw (1994). In their data fusion experiments, Fox and Shaw found that combining two runs of the same type, either both vector queries or p-norm queries, showed little improvement over the individual runs, and performed worse than the best of the two runs in many instances. They suggested it might be because the two schemes were too similar. On the other hand, in their experiments, combining any one of the vector queries with any one of the p-norm queries always showed an improvement.

Another unsuccessful data fusion experiment was reported by Ng, *et al.* (1997). Using the TREC 5 documents collection, we combined the data from two very different IR schemes. The first IR scheme was called ***Scheme KB***. It was based on discriminative term-stems: The appearance of each term-stem in all the judged documents was assumed to follow a Poisson distribution, and the parameter of that distribution was estimated and used to rank the word-stems for their ability to discriminate between relevant and non-relevant documents. The second IR scheme was called ***Scheme DL***. It was based on the compressibility of test documents with respect to a dictionary built from all the judged relevant documents. It used the LZW compression model (Welch, 1984) to calculate the entropies of the test documents conditioned on a topic statement plus the documents judged relevant to the topic. These entropies were used to yield probabilities that form the basis for ranking the relevance of new documents for each topic. Even though ***Scheme KB*** was operationally very different from ***Scheme DL***, data fusion based on these two schemes did not improve performance. We conjectured that this was because the compression based scheme was a very poor IR scheme.

I have participated in two of the above experiments and the results made me aware of the importance of exploring the conditions for effective data fusion. If there is more than one sensor, they can act together to confirm an event. They can also reduce any ambiguity surrounding the event. Therefore, in theory, data fusion should give better performance. Why, then, does data fusion sometimes work well but sometimes not? Are there rules that govern data fusion's performance? From the two failed experiments, empirically one can suggest two conditions for effective data fusion:

1. Condition of Efficacy: fusion of two IR schemes with comparable performance tends to be effective;

2. Condition of Dissimilarity: fusion of two dissimilar IR schemes tends to be effective.

The rationale of these two conditions for effective data fusion is easy to understand intuitively, but we must remember that intuition is not proof.

3.2 Condition of Efficacy

Efficacy is a matter of degree. If the performance of an IR scheme is very poor but better than randomly selecting documents, in theory it still can offer extra evidence for data fusion, and improve performance. Compared to another scheme, a poor IR scheme may retrieve a few distinct relevant documents to increase recall, or retrieve a few common relevant documents to confirm the possible relevance of those documents to increase precision.

In addition, there is always the possibility of tuning to make a better fusion: when the individual systems each have their thresholds reset to take account of the existence of each other, inferior sensors can still contribute to the fusion. For example, Bartell, Cottrell and Belew (1994) combined the output of two IR schemes, one based on query terms and the other based on query phrases. The performance of the phrase-based scheme was much lower than the term-based scheme, but after optimization tuning, which actually weighted the phrase-based output slightly higher than the term-based output (the optimized phrase weight is found to be 0.738 and the term weight is 0.675), data fusion improved performance. This demonstrates that, after sophisticated tuning, an inferior IR scheme can contribute to data fusion and improve performance.

However, when there are no training data available, data fusion between a capable IR scheme and a very inferior IR scheme may make the improvement difficult to detect. It may also impose more constraints on the fusion method, needed to eliminate the non-relevant documents retrieved.

3.3 Condition of Dissimilarity

Dissimilarity is also a matter of degree. Unless the two IR schemes produce identical outputs (same documents with same ranks), very similar outputs should still be able to offer some extra evidence to improve performance.

However, practically, similar outputs usually make the additional information difficult to detect. As mentioned above, for those data fusion experiments which did not get better performance (e.g., Ng and Kantor, 1996; Fox and Shaw, 1994), the fusion rule used (sum of normalized relevance scores) in fact was effective in other data fusion experiments, but apparently performance was not improved because the outputs were too similar.

3.4 Caveat Concerning The Condition of Efficacy

There is one caveat about the condition of efficacy. It states that the fusion of comparable IR schemes tends to be effective. It does not exclude the possibility that data fusion between two very incapable IR schemes can be effective.

Theoretically, if data fusion improves performance, the improvement of the data fusion between two very incapable IR schemes (poor scheme vs. poor scheme fusion) may not be as difficult to detect as the data fusion between a capable IR scheme and a very inferior IR scheme (good scheme vs. poor scheme fusion). The baselines for comparison of these two fusions are very different. As for the former case (i.e., poor scheme vs. poor scheme fusion), the improvement is measured against an incapable IR scheme, while for the latter case (i.e., good scheme vs poor scheme fusion), the improvement is measured against a capable IR scheme. In fact, for those unsuccessful data fusion experiments we mentioned above in Section 3.1, they were done between capable IR scheme and very incapable IR scheme, not two very incapable IR schemes.

The above described two possible conditions for effective data fusion in IR. Before investigating further and running experiments to test the validity of these two conditions, we need two measures to represent efficacy and dissimilarity. For efficacy, there are already some well established measures for IR schemes, e.g., precision and recall. We don’t need to develop a new measure for it. However, for schemes dissimilarity, we need a rigorous definition of inter-IR-schemes distance in order to establish such a measure. The next chapter offers this definition.

Chapter 4:

Rigorous Definition and Implementation Algorithm of Inter-IR-Scheme Dissimilarity

Comparison between IR set output and IR ranked output is not commonly attempted. It may be because there are difficulties with finding a control variable fair to both systems (Keen 1992). However, it is not theoretically impossible. The famous Cranfield Tests provided a classic method to compare ranked output and set output (see Cleverdon ,1967, p.616). Since most of sophisticated IR schemes produce ranked outputs, in my study, we will only focus on the dissimilarity between ranked output lists.

4.1 Distance between two ranked lists

To quantify the dissimilarity between two IR schemes, we need an objective measure. Since our concern is the operational performance, not the underlying mechanism or algorithm, instead of defining the distance between IR schemes, we can measure the distance between their outputs.

Mathematically, a ranked list of ***N*** items can be broken down into ***½N(N –1)*** ordered pairs. From such a set of ordered pairs, only one ranked list can be reconstructed. For example, for a ranked list with 3 items ***A***, ***B***, ***C*** such that

$$\boldsymbol{A} > \boldsymbol{B} > \boldsymbol{C}$$

(I use “ > ” to represent hierarchical relationship in the ranked list), the list generates ½ ×3×(3–1) = 3 ranked pairs, i.e.,

$$\begin{gathered} \boldsymbol{A} > \boldsymbol{B} \\ \boldsymbol{A} > \boldsymbol{C} \\ \boldsymbol{B} > \boldsymbol{C} \end{gathered}$$

From these 3 ranked pairs, we can re-construct exactly one ranked list which contains all the 3 items, i.e., the original ranked list, ***A* > *B* > *C***.

In other words, the ranked list and the set of ranked pairs are alternative ways of representing the ordering among the elements.

Of course, there could be inconsistent pair ranks, such as

$$\begin{gathered} \boldsymbol{A} > \boldsymbol{B} \\ \boldsymbol{B} > \boldsymbol{C} \\ \boldsymbol{C} > \boldsymbol{A} \end{gathered}$$

Those situations do not occur in IR ranked output lists. Therefore, when comparing two ranked lists, instead of comparing two ranked lists directly, we can compare the ordered pairs determined by each list.

If we have two ranked lists which contain the same elements, we can break down these two lists into two set of ranked pairs. Each set should contain the same number of pairs and every pair appearing in one set will also be a pair in the other set (but may be in different order.) For example, for two ranked lists:

$$\boldsymbol{A} > \boldsymbol{B} > \boldsymbol{C}$$
$$\boldsymbol{B} > \boldsymbol{A} > \boldsymbol{C}$$

each generates three ranked pairs. For the first list, they are

$$\boldsymbol{A} > \boldsymbol{B}$$
$$\boldsymbol{A} > \boldsymbol{C}$$
$$\boldsymbol{B} > \boldsymbol{C}$$

For the second list, they are

$$\boldsymbol{B} > \boldsymbol{A}$$
$$\boldsymbol{A} > \boldsymbol{C}$$
$$\boldsymbol{B} > \boldsymbol{C}$$

Every pair in the first set is also a pair in the second set. The first of the three pairs is out of order with respect to each other ($\boldsymbol{A} > \boldsymbol{B}$ vs. $\boldsymbol{B} > \boldsymbol{A}$), while the second and third pairs are in the same order.

This representation of ranked lists can be used to calculate the distance between two ranked lists (Kemeny, 1964). If the two ranked lists have the same elements but the elements are arranged in a different order, we can represent each list in terms of ranked pairs and count the number of out-of-order pairs between them.

For example, for the two lists

A > ***B*** > ***C***
B > ***A*** > ***C***

there is one out-of-order pair (i.e., in the first list, ***A* > *B***, while in the second list, ***B* > *A***). We call this count the non-normalized distance between the two lists. (In this case, 1). Based on the number of out of order pairs, we can measure the distance between the two lists.

4.2 Scores of Out-of-Order Pairs for Non Common Items

In an IR ranked output environment, if we use two different IR schemes to retrieve relevant documents in a collection with respect to a query, we will have two ranked output lists. We can use the number of out-of-order pairs to measure the dissimilarity between the ranked outputs of two IR schemes.

When the collection is huge, it is not likely for an IR system to offer the user a full output list. Therefore, we will only compare the top portion of the ranked output lists, not full lists, of different IR schemes. In this case, it is likely that the documents in the two lists are not exactly the same.

When the two ranked lists have different elements, we may encounter two new situations:

1. For a given pair in one list, only one element is present in the other list;
2. For a given pair in one list, neither element is in the other list.

For a pair from one list such that only one of the documents is in the other list, logically the order of the pair in the other list can be easily determined because the missing document must be in the lower part of the other list, below the cutting point.

For those pairs of which both documents only appear in one list but not in the other, they can be either out-of-order or not, with equal probability (because there are as many permutations of the list in which the pair is in order as there are with it out of order.) Therefore we treat the out-of-order scores for those pairs as "0.5" (Kantor, Ng and Hull, 1998).

4.3 Topic-Wise vs Scheme-Wise Dissimilarity

With a scale to measure the distance between the ranked output lists produced by two IR schemes, we can develop two measures:

(1) <u>topic-wise dissimilarity</u>, which measures the distance between different IR schemes topic by topic. That means, for different topics, the distance between two IR schemes may be different according to the topic involved.

(2) <u>scheme-wise dissimilarity</u>, which measures the distance between different IR schemes by summing or averaging the topic-wise dissimilarity scores of various topics. That means, the distance between two IR schemes is a composite scale based on all the topic-wise dissimilarities of the two

schemes. In theory, the more topics involved in constructing the scheme-wise dissimilarity scale, the more reliable is the scale for general comparison purpose.

4.4 Normalization of Inter-IR-Scheme Dissimilarity

For those IR schemes which have larger cutoff points (i.e., have more documents in their ranked output lists), their distances between each other and between some other IR schemes tend to be higher for those IR schemes which have smaller cutoff points. This is because, for an IR scheme with cutoff point $\boldsymbol{N}$, there are $\boldsymbol{\frac{1}{2}\,(N \times (N\text{-}1))}$ pairs of items (documents) in its output list. The larger the value of $\boldsymbol{N}$, the more likely there will be a higher counts of out-of-order pairs when comparing with another ranked output lists. This will make the inter-IR-scheme dissimilarity scores non-comparable for schemes with different cutoff points.

In order to normalize the scores, I need a normalization factor. The normalization factor I choose is the maximum possible out-of-order scores a pair of IR schemes can get. Two IR schemes will have the maximum possible out-of-order scores when there is no common document in their output lists.

For two disjointed output lists, the total number of out-of-order scores will consist of two portion. The first portion is the number of out-of-order pairs with one item from one list and the other item from the other list. It is equal to $\boldsymbol{N_1 \times N_2}$ where $\boldsymbol{N_1}$ and $\boldsymbol{N_2}$ are the cutoff points for the first and second IR schemes respectively.

The second portion is the number of out-of-order pairs with both of the two items from the same output lists. There will be $\frac{1}{2} (N_1 \times (N_1 - 1))$ pairs for the first list and $\frac{1}{2} (N_2 \times (N_2 - 1))$ for the second list. Since we consider the out-of-order scores of these kind of pair to be “0.5” (see section 4.2), the second portion is equal to half of

$$\boxed{\frac{1}{2} (N_1 \times (N_1 - 1)) + \frac{1}{2} (N_2 \times (N_2 - 1))}$$

Therefore, the normalization factor becomes:

$$\boxed{\textit{normalization factor} = N_1 \; N_2 + \frac{1}{2} \left[\frac{1}{2} N_1 \, (N_1 - 1) + \frac{1}{2} N_2 \, (N_2 - 1) \right]}$$

The normalized inter-IR-scheme dissimilarity is the out-of-order pairs scores divided by its normalization factor.

4.5 Algorithm for Calculating Inter-IR-Scheme Dissimilarity

Let N be the cutoff point of a ranked output list produced by an IR scheme. For each output list, there are $\frac{1}{2} N \times (N - 1)$ pairs of documents. When the cutoff points are large, e.g., 1,000, there will be ½(1,000 × 999) = 499,500 pairs of documents in each output list. In other words, there are 499,500 comparison to compute the number of out- of-other pairs.

When the cutoff point is as high as 1,000, it is highly likely for the two lists to have a lot of unique documents. When there are unique documents in the two output lists, there will be many more than 499,500 comparisons. For the extreme case of two disjoint lists, we will need to go through 1,000 × (2,000 – 1) = 1,999,000 comparisons to compute the number of out-of-order pairs.

In our experiments, we use two data set (see next chapter, Chapter 5). Each data set has more than ten thousand cases. For each case, we have to go through about five hundred thousand to two millions comparison to compute the number of out-of-order pairs. If we use the number of out-of-order pair to measure the distance between two IR schemes output, we will need an algorithm to compute the number of out-of-order pairs efficiently.

Representing the number of out-of-order pairs by ***z*,**, ***z*** can be divided into five components:

$$\boxed{\boldsymbol{z = z_1 + z_2 + z_3 + z_4 + z_5}}$$

where

$\boldsymbol{z_1}$ is the out-of-order scores for pairs with one element in the intersection of the two lists, and one element unique to the list generated by the first IR scheme;

z_2 is the out-of-order scores for pairs with one element in the intersection of the two lists, and one element unique to the list generated by the second IR scheme;

z_3 is the out-of-order scores for pairs with both elements in the intersection;

z_4 is the out-of-order scores for pairs with one element unique to the first IR scheme and the other unique to the second IR scheme;

z_5 is the out-of-order scores for pairs with both elements appearing in only one of the two lists.

We develop a set of algorithms to compute these five components effectively.

4.5.1 Formulae for the Components z_1, z_2, z_4, and z_5

In this section, we discuss the algorithms we used in our computer program to calculate z_1, z_2, z_4, and z_5 effectively (Kantor, Ng & Hull, 1998). The calculation of z_3 requires a recursive function which recursively divides the common elements into two sets and then calculates the out-of-order scores for each of these sets iteratively. Details of the algorithms of computing z_3 will be discussed in the next section, section 4.5.2.

z_1 and z_2 are the out-of-order scores for pairs with one element in the intersection of the two lists, and one element unique to the list generated by the other IR scheme. The algorithm of calculateing z_1 and z_2 is the same. In the following, we only discuss the algorithm of calculating z_1**.**

For ***list 1***, the total number of out-of-order pairs between a common element and all the other unique elements in ***list 1*** is equal to the number of unique documents in ***list 1*** above that common document. This is because for those unique documents that are below the common document, the orders are the same as ***list 2***.

The number of unique documents in ***list 1*** above a common document ***c*** is equal to one less of the rank of ***c*** minus the number of common documents above it. Therefore the number of out-of-order pairs between ***c*** and all unique documents (denoted by ***OOP(c)***) is:

$$\boxed{\boldsymbol{OOP(c)} \quad = \quad \text{rank of } \boldsymbol{c} \quad - \quad \text{number of common documents above } \boldsymbol{c} \; - 1}$$

Let the rank of the k^{th} common document in ***list 1*** be c_k . Since There are $k - 1$ common documents above c_k, there are $c_k - (k -1) - 1 = c_k - k$ unique documents above c_k. In other words, $OOP(c_k) = c_k - k$. Summing up the $OOP(c_k)$ for all the ***m*** common documents, we have:

$$\begin{aligned}\sum_{k=1}^{k=m} OOP\ (c_k) &= \sum_{k=1}^{k=m} (c_k - k) \\ &= \sum_{k=1}^{k=m} c_r - \sum_{k=1}^{k=m} k \\ &= \sum_{k=1}^{k=m} c_r - \frac{m(m+1)}{2}\end{aligned}$$

In other words, $z_1 = \sum_{k=1}^{k=m} c_r - \frac{m(m+1)}{2}$.

When we wrote our computer program for calculating ***z***, the equation we used for z_1 was different from above. We used the following equation:

$$z_1 = \frac{1}{2}(N_1)(N_1 + 1) - \frac{1}{2}(m)(m+1) - R_1$$

where ***m*** is the number of common documents retrieved by both schemes, R_1 is sum of ranks of documents that are retrieved by the first IR scheme but not the second.

Since sum of the rank of common elements plus the sum of rank of unique elements is equal to sum of the rank of all the elements in ***list 1***, i.e., ,

$$\boxed{R_1 + \sum_{r=1}^{r=m} c_r = \frac{N_1(N_1 + 1)}{2}}$$

we can replace $\sum_{r=1}^{r=m} c_r$ by $\frac{N_1(N_1+1)}{2} - R_1$. Therefore the the equation we used is mathematically equvalent to $z_1 = \sum_{k=1}^{k=m} c_r - \frac{m(m+1)}{2}$.

The equation of z_2 is:

$$\boxed{z_2 = \frac{1}{2}(N_2)(N_2 + 1) - \frac{1}{2}(m)(m+1) - R_2}$$

where R_2 is sum of the ranks of documents that are retrieved by the second IR scheme but not the first. The proof of the formula for z_2 is the same as z_1.

z_4 is the out-of-order scores for pairs with one element unique to the first IR scheme and the other unique to the second IR scheme, the calculation is straight forward:

$$\boxed{z_4 = (N_1 - m)\,(N_2 - m)}$$

z_5 is the out-of-order scores for pairs with both elements appearing in only one of the two lists. Since the number of unique elements in each list is equal to ***N – m***, and we consider the out-of-other score of this kind of pairs is “0.5”, the out-of-order score is sum of the number of possible combinations times “0.5”, i.e.,

$$\boxed{z_5 = \frac{1}{2}\left[\frac{1}{2}(N_1 - m - 1)(N_1 - m) + \frac{1}{2}(N_2 - m - 1)(N_2 - m)\right]}$$

4.5.2 Details of the Iterative Algorithm for z_3

The algorithm for z_3 is as follow: For two lists ***S*** and ***L*** with the same elements, we can calculate the out-of-order score by first splitting the list ***L*** into two halves, ***T*** and ***B***, where ***T*** is the top half of the list and ***B*** is the bottom half (It is not necessary for ***T*** and ***B*** to be exactly equal.) Now there are only two kinds of pairs:

- Condition 1: Between split lists: one item from ***T***, the other from ***B***
- Condition 2: Within split lists: both items from the same spit list.

Let r_i be the rank of the i-th element of ***T*** in ***S***, and let ***t*** be the size of ***T***. For the first condition (condition of between split lists), the number of out-of-order pairs can be calculated by the following formula which is similar to Wilcoxon rank-sum test:

$$\boxed{\textit{Out of order scores} = \sum_{i=1}^{t} r_i - \frac{t\,(t+1)}{2}}$$

For the second condition (condition of within split lists), I can again form two split lists. The number of out-of-order pairs for each of them can be calculated by splitting the list into top and bottom again (e.g., split T into top-T and bottom-T; split B into top-B and bottom-B) such that for each of them, there are only two conditions of out-of-order:

- Condition 1: Between split lists
- Condition 2: Within split lists

For a new condition 2, I can split the newly split lists again. This recursive splitting can be looped until the lists are un-splittable (i.e., there is only one item in each list.)

Now that we have a rigorous definition of inter-IR-scheme dissimilarity and a set of formulae to calculate this dissimilarity, I can use this measure for my experiments to test the validity of the two conditions for effective data fusion in IR.

Chapter 5:

Research Questions, Data, and Rule of Combination

5.1 Research Questions

In the above, two possible conditions, efficacy and dissimilarity, are proposed as the practical conditions for effective data fusion in IR:

1. <u>Condition of Efficacy:</u> Fusion of two IR schemes with comparable performance tends to be effective;
2. <u>Condition of Dissimilarity:</u> Fusion of two dissimilar IR schemes tends to be effective.

This thesis will investigate the influences of schemes efficacy and inter-IR-scheme dissimilarity on the effectiveness of data fusion. Our research questions are:

1. Are these two conditions good criteria for predicting the effectiveness of data fusion in IR, and if so, how should we implement this idea?
2. How powerful are the IR schemes efficacy and inter-IR-scheme dissimilarity in discriminating effective data fusion from non effective

data fusion in IR?

Our conjecture is: the effectiveness of data fusion is a function of the efficacy of each individual IR scheme, and the dissimilarity between the schemes.

To measure the efficacy of an IR scheme, one can use precision, recall, or some combination of precision and recall, we use precision at the 100th document as the measure for IR scheme efficacy.

To measure the dissimilarity between two IR schemes, we use the normalized inter-IR-schemes dissimilarity (as discussed in Chapter 4) to represent the distance between the two scheme.

5.2 Data

We use the output lists of the IR schemes produced for the routing tasks of the fourth and fifth Text REtrieval Conferences (i.e., TREC 4 and TREC 5, see Harman, 1996; Voorhees and Harman ,1997) as my raw data for investigating, training, and testing of the two condition for effective data fusion.

The Text REtrieval Conferences have been run as workshops for participating groups to discuss their IR systems results on the retrieval tasks done using the TREC collections (about 750,000 to 1,000,000 documents). They were sponsored by National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of the TIPSTER Text Program.

Until most recently, each of the TREC conferences has centered around two main tasks, one was called the routing task, and the other was called the adhoc task. In the routing task, the same questions (topics) were asked, but new documents were being evaluated. These searches are similar to those required by news clipping services and library profiling systems. In the adhoc task, new questions were asked against a static set of documents. The task is similar to the way that a researcher might use a library.

To accomplish the routing task and adhoc task, each of the participating IR schemes produce a list of the 1,000 documents for each topic, ranked according to decreasing scores of relevance as estimated by that IR scheme. The schemes were evaluated based on recall (relative recall) and precision. We use these ranked output lists as input for data fusion.

Since the output list of each of the participating IR schemes contains not only the ranked order of the 1000 documents retrieved, but also the relevancy scores assigned by the scheme, there is enough information for the data fusion to be done on the attribute level (by assigned relevancy scores) as well as decision level (by assigned ranked order). In our experiment, we use sum of normalized relevance scores as our rule of data fusion, so the fusion is on a pseudo-attribute level (see Chapter 1, section 1.3).

For TREC 4 routing task, there were a total of 28 sets of results, with 26 of them based on runs for all the full data set available. We only use these 26 set of results. Each set of result had 50 topics, so totally we have $\frac{1}{2} \times 26 \times 25 \times 50 =$

16,250 possible cases of data fusion between two IR schemes.

For TREC 5 routing task, there were a total of 26 sets of results, with 23 of them based on runs for the full data set available. We only use these 23 set of results. There were originally 50 topics, so each set of result included 50 topics. However, after cleaning the data and running some preliminary tests, we found that there were 5 topics (topics 68, 125, 237, 240 and 243) for which no one participating scheme has retrieved any relevant documents. we checked this finding with the judged document lists provided by NIST. According to the judged document list, there were no relevant documents for these five topics (only non-relevant documents). Therefore, for TREC 5, totally we have $\frac{1}{2} \times 23 \times 22 \times 45 =$ 11385 possible data fusion between two IR schemes.

5.3 Rule of Combination

Since there are many successful data fusion experiments using the sum of normalized relevancy scores as the fusion rule (e.g., Fox & Shaw, 1993; Fox & Shaw, 1994; Belkin, Kantor, Fox & Shaw, 1995), in our data fusion experiments, we use the sum of normalized relevancy scores as the fusion rule for combination.

We use all the 1,000 documents of the output lists for fusion. We rank the sum or normalized relevance scores such that the larger the sum of normalized relevancy score of a document, the higher its rank on the fused list. The normalized relevancy score is calculated according to the following formula:

$$\textbf{\textit{normalized relevancy score}} = \frac{\textbf{\textit{original relevancy score}} - \textbf{\textit{minimum relevancy score}}}{\textbf{\textit{maximum relevancy score}} - \textbf{\textit{minimum relevancy score}}}$$

where minimum relevancy score is the lowest relevancy score within the top 1,000 documents of the IR scheme for the topic in concern and maximum relevancy score is the highest relevancy score within the top 1,000 documents of the IR scheme for the same topic.

Chapter 6:

Methodology, Experimental Design, and Reliability of the Dissimilarity Measure

We employ quantitative research methods to investigate the relationship between scheme efficacy, inter-scheme dissimilarity, and data fusion effectiveness. The statistical methods that employed included linear discriminant analysis, multiple regression analysis, logistic regression analysis, and one non-parametric and empirical method. Signal detection theory and receiver operating characteristic curve (Egan, 1975) are used to investigate the detection rate and false alarm rate for different discriminant scores generated by the statistical analyses.

In the following sections, we discuss the general idea of the experimental design and offer the definitions of the variables that will be used in the analyses.

6.1 Effectiveness of Data Fusion

Let the fusion of two IR schemes $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ by fusion rule $\boldsymbol{f}$ be denoted as $\boldsymbol{S_1 f S_2}$.

Let the performance of an IR scheme S be $P(S)$. $P(S)$ can be recall, or precision, or some other performance measure in IR. For a topic-wise comparison, we can define the effectiveness of data fusion $E\ (S_1 f\, S_2)$ as:

$$\boxed{E_o\ (S_1\ f\ S_2) = \frac{P\ (\ S_1 f\ S_2) - max\{P(S_1), P(S_2)\}}{max\{P(S_1), P(S_2)\}}}$$

The above formula can be used for both topic-wise comparison and scheme-wise comparison. For a scheme-wise comparison, we can replace the topic performance scores by average performance scores over all topics. For this investigation, we concentrate on the topic-wise comparison.

In the above definition, $E_o\ (\ S_1\ f\ S_2\ )$ is positive only if $S_1\ f\ S_2$ is better than the best scheme. Therefore, $E_o\ (\ S_1\ f\ S_2\ )$ can be interpreted as the percent improvement over the best scheme.

There are other possible definitions of $E\ (\ S_1\ f\ S_2\ )$. For example:

$$\boxed{E_u\ (\ S_1\ f\ S_2\ ) \quad = \quad \frac{P\ (\ S_1\ f\ S_2\ ) \ - \ \frac{1}{2}\ [\ P(S_1\ ) + P(S_2\ )\ ]}{\frac{1}{2}\ [\ P(S_1\ ) + P(S_2\ )\ ]}}$$

As with the original definition, $\boldsymbol{E_u (S_1 f S_2)}$ can be used for topic-wise comparison as well as scheme-wise comparison.

$\boldsymbol{E_u (S_1 f S_2)}$ is the effectiveness in comparison with an uninformed decision maker, while $\boldsymbol{E_o (S_1 f S_2)}$ is effectiveness in comparison with an oracle. An uninformed decision maker would simply choose at random between the two schemes, with an expected performance:

$$\boxed{\boldsymbol{\frac{1}{2}[P(S_1) + P(S_2)]}}$$

On the other hand, an oracular decision maker would always choose the better of the two systems, with a resulting performance $\boldsymbol{max \{P(S_1), P(S_2)\}}$. In this investigation, we use the more stringent $\boldsymbol{E_o}$ as our performance measure.

Mathematically the hypothesis of this thesis is:

$$\boxed{\boldsymbol{E (S_1 f S_2) = F (P(S_1), P(S_2), d(S_1 , S_2))}}$$

where $\boldsymbol{d(S_1 , S_2)}$ is the normalized inter-IR-scheme dissimilarity between $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$,i.e., the normalized $\boldsymbol{z}$, with minimum value as 0 (i.e., $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are identical in terms of the elements and ranked order included in the two lists) and maximum value as 1 (i.e., there is no one single common element in the two lists).Where $\boldsymbol{F}$ is a function. In this work, we explore certain form of the function. The null hypothesis, there is no such a function.

6.2 Variables

Accordingly, our analyses have one dependent variable and three independent variables. The dependent variable can be $E\ (S_1\ f\ S_2\ )$ directly, or some transformation of $E\ (S_1\ f\ S_2\ )$. For example, a binary transformation like,

$$y = 0 \text{ if } E\ (S_1\ f\ S_2\ ) <= 0$$
$$y = 1 \text{ if } E\ (S_1\ f\ S_2\ ) > 0$$

where y represents whether the data fusion improves performance, not how much improvement can the data fusion gain.

There will be three independent variables:

1. $P\ (S_1\ )$: performance, or efficacy, of IR scheme 1, i.e., precision at 100 documents of IR scheme 1;
2. $P\ (S_2\ )$: performance, or efficacy, of IR scheme 2, i.e., precision at 100 documents of IR scheme 2;
3. $d\ (S_1,S_2\ )$: inter-scheme dissimilarity between S_1 and S_2 , i.e., normalized z.

From the TREC 4 and TREC 5 routing task data, we generate all the cases with the four variables, denoted by y^i, x^i_1 , x^i_2 and x^i_3 for the i^{th} case, where y^i is the effectiveness of data fusion (or some transformation of it), $x^i{}_1$ and $x^i{}_2$ are the efficacy of IR schemes 1 and 2 respectively for a topic, and $\boldsymbol{x^{i}_{3}}$ is the inter-scheme dissimilarity for that topic.

We try to demonstrate that $\boldsymbol{y}$ is a function of $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ and $\boldsymbol{x_3}$. We approach this problem as a machine-learning problem – to extrapolate from the effectiveness of data fusion of the IR schemes participated in TREC 4 routing task data the effectiveness for data fusion of the IR schemes participated in TREC 5 routing task.

6.3 Reliability of the Dissimilarity Measure

Before using the normalized dissimilarity (number of out-of-order pairs between the rank output lists of two IR schemes, divided by the maximum possible number of out-or-order pairs) to measure the inter-IR-scheme dissimilarity, we want to test its reliability to make sure that it is a consistent measure across different types of information problems. The method we used appeals to the internal consistency of the inter-IR-schemes dissimilarity scores across different topics.

First we constructed a matrix of topics by IR scheme pairs (*Table 6.1*). The cell entries of the matrix are the normalized dissimilarity between the two IR scheme outputs (row) with respect to a topic (column).

In the *Table 6.1*, ***scheme (i,j)*** stands for the pair of IR schemes ***i*** and ***j*** and ***d(i,j,t)*** stands for the normalized ***z*** of ***scheme(i,j)*** computed using their output lists for topic ***t***. Given such a matrix, we can compute an estimate of reliability based on observed correlations or covariances of the topics with each other.

	topic 1	topic 2	...	topic t	...
Scheme(1,1)	***d(1,1,1)***	**...**	**...**	**...**	**...**
Scheme (1,2)	***d(1,2,1)***	**...**	**...**	**...**	**...**
...	**...**	**...**	**...**	**...**	**...**
scheme (i,j)	***d(i,j,1)***	***d(i,j,2)***	**...**	***d(i,j,t)***	**...**
...	**...**	**...**	**...**	**...**	**...**

Table 6.1: Matrix of topics by system pairs. Cell entries are distances between IR schemes outputs (row) with respect to a topic (column).

For any two IR systems ***i*** and ***j***, if the calculation of inter-scheme dissimilarity is reliable, we should get a highly positively correlated inter-scheme dissimilarity scores when we use other completely different (or partly different) set of topics. If ***d(i,j,t)*** for different ***t*** are not positively correlated with each other, we have no reason to believe they are positively correlated with other possible topics we may have selected. In this case, we do not expect to see a positive relationship between this test (i.e., average sum of ***z*** for this particular set of topics) and other similar tests (i.e., using other set of topics).

This above reasoning is very similar to the calculation Cronbach's Alpha (Carmines & Zeller, 1979; Dunn 1989). The only difference is, Cronbach Alpha usually is used to measure the internal consistency of a composite scale, while in our analysis here, we have no intention to combine different value of normalized dissimilarity to form a composite measure.

Cronbach’s Alpha has several interpretations. It can be viewed as the correlation between test or scale and all other possible tests or scales containing the same number of items, which could be constructed from a hypothetical universe of items that measure the characteristic of interest. Cronbach’s Alpha can be computed using the following formula:

$$\boxed{\textbf{\textit{Cronbach}}\ \alpha = \frac{\textbf{\textit{no. of items}} \times \dfrac{\textbf{\textit{averaged covariance}}}{\textbf{\textit{averaged variance}}}}{\textbf{\textit{1 + (no. of items –1)}} \times \dfrac{\textbf{\textit{averaged covariance}}}{\textbf{\textit{averaged variance}}}}}$$

which, translating into our case, is equal to:

$$\boxed{\frac{\textbf{\textit{number of topics}} \times \textbf{\textit{variance ratio}}}{\textbf{\textit{1 + (number of topics - 1)}} \times \textbf{\textit{variance ratio}}}}$$

where ***variance ratio*** is the ratio of ***averaged covariance between topics*** to ***averaged variance of the topics***. We can see from the above formula that Cronbach's Alpha is based on the internal consistency of the normalized ***z*** scores. The higher this Cronbach Alpha, the more reliable the normalized ***z*** is.

Because of the fact that in our experiments we will use TREC 4 data for training and TREC 5 data for testing, we are not going to use TREC 5 data to test the reliability of ***z*** but only limited to the TREC 4 data.

For the routing task, there were 26 systems and 50 topics. The Cronbach Alpha is 0.9938. Removing the normalized ***z*** scores with respect to any one topic cannot increase the value of alpha (*Table 6.2,*). This value is remarkably high. In other words, the dissimilarity measure is very reliable, with little dependence on the topics submitted to the system.

	Variance	**Alpha**
	if topic	**if topic**
	deleted	**deleted**
TOPIC_AA	113.7639	0.9937
TOPIC_AB	112.9049	0.9936
TOPIC_AC	114.1261	0.9938
TOPIC_AD	113.1696	0.9937
TOPIC_AE	112.4377	0.9937
TOPIC_AF	112.9014	0.9937
TOPIC_AG	114.5902	0.9938
TOPIC_AH	113.0302	0.9937
TOPIC_AI	113.1392	0.9938
TOPIC_AJ	111.9168	0.9937
TOPIC_AK	112.769	0.9936
TOPIC_AL	112.5939	0.9936
TOPIC_AM	112.9835	0.9937
TOPIC_AN	113.1641	0.9938
TOPIC_AO	113.1996	0.9938
TOPIC_AP	112.5325	0.9937
TOPIC_AQ	113.3018	0.9938
TOPIC_AR	112.4524	0.9937
TOPIC_AS	112.3402	0.9939
TOPIC_AT	113.8767	0.9937
TOPIC_AU	113.3633	0.9937
TOPIC_AV	112.2765	0.9936
TOPIC_AW	113.0764	0.9937
TOPIC_AX	114.2227	0.9938
TOPIC_AY	113.4078	0.9938

	Variance	**Alpha**
	if topic	**if topic**
	deleted	**deleted**
TOPIC_AZ	113.0966	0.9937
TOPIC_BA	114.0286	0.9938
TOPIC_BB	113.4403	0.9937
TOPIC_BC	113.2203	0.9936
TOPIC_BD	113.3349	0.9937
TOPIC_BE	113.4194	0.9937
TOPIC_BF	113.8436	0.9938
TOPIC_BG	114.4264	0.9937
TOPIC_BH	113.6741	0.9938
TOPIC_BI	112.8828	0.9937
TOPIC_BJ	112.6756	0.9937
TOPIC_BK	112.6801	0.9937
TOPIC_BL	113.1965	0.9937
TOPIC_BM	114.8024	0.9938
TOPIC_BN	112.7825	0.9937
TOPIC_BO	113.0274	0.9937
TOPIC_BP	114.092	0.9938
TOPIC_BQ	113.9856	0.9937
TOPIC_BR	113.6662	0.9938
TOPIC_BS	113.8674	0.9937
TOPIC_BT	113.6201	0.9937
TOPIC_BU	113.2336	0.9937
TOPIC_BV	114.492	0.9938
TOPIC_BW	114.2047	0.9938
TOPIC_BX	113.6581	0.9937

Number of Cases = 325 Reliability Coefficients Alpha = .9938

Table 6.2 Result of reliability analysis for the normalized z scores.

Chapter 7:

Discriminant Analysis Using

Linear Combinations of The Three Independent Variables

To begin our exploration, we test the two conditions for effective data fusion in IR by linear discriminant analysis, using the statistical program SPSS for Windows version 6.1.

From the output lists submitted by all the IR schemes participating in TREC 4 routing tasks, we generate 16,500 fused lists by the following procedures.

1. We rank the documents retrieved by the two IR schemes according to the sum of normalized relevancy scores and keep the top one thousand documents,
2. We compare the top 100 documents of the fused list with the list of relevant documents provided by TREC 4 and compute the precision at the 100th document of the fused list, i.e., $\boldsymbol{P(S_1\, f\, S_2)}$, where $\boldsymbol{f}$ represents the simple symmetric fusion rule.

3. We calculate the relative improvement of the data fusion, $E_o\ (S_1\ f\ S_2)$, by the formula:

$$\boxed{E_o\ (\ S_1\ f\ S_2\ ) \quad = \quad \frac{P\ (\ S_1\ f\ S_2\ ) - max\ \{\ P\ (\ S_1\ ),\ P\ (\ S_2\ )\ \}}{max\ \{\ P\ (\ S_1\ ),\ P\ (\ S_2\ )\ \}}}$$

4. We divide the all the cases into two groups: cases that with positive $E_o(S_1 f S_2)$ belong the positive group, and cases that with negative $E_o(S_1 f S_2)$ belong to the negative group. We drop those cases with $E_o(\ S_1\ f\ S_2\ ) = 0$ from analysis.

Thus the positive group contains all the cases where symmetrical data fusion by sum of normalized relevancy scores improves precision at the 100th document; the negative group contains all the cases that the precision at the 100th document of the list produced by the symmetrical data fusion is worse than the better scheme of the two original IR schemes. All comparisons are against an oracle who chooses the better system from pair, for each topic.

7.1 Conceptual Framework of Applying Linear Discriminant Analysis

We seek a linear combination of the three independent variables (i.e., precision at 100th document of the two IR schemes outputs, and the normalized z score of the two lists) as the basis for assigning cases into the two groups:

$$\boxed{\boldsymbol{D = \beta_0 + \beta_1 P(S_1) + \beta_2 P(S_2) + \beta_3\ d(S_1, S_2)}}$$

the coefficients $\boldsymbol{\beta}$ of the linear equation will be chosen so that the values of $\boldsymbol{D}$ differ as much as possible between the groups.

If we represent all the positive cases and negative cases by points in a three dimensional space with axes defined by $\boldsymbol{P(S_1)}$, $\boldsymbol{P(S_2)}$, and $\boldsymbol{d(S_1 , S_2)}$, we can move a plane perpendicular to the line defined by vector $\boldsymbol{(\beta_1 , \beta_2 , \beta_3)}$. That plane serves as the classification plane with cases on one side of the plane classified as positive cases and cases on the other side of the plane classified as negative cases (for details, see Section 7.4). Moving the plane along the line will vary the classification of positive cases and negative cases. This idea can be implemented by using the discriminant scores $\boldsymbol{D}$ of all the cases to plot an ROC (receiver operating character, see below) curve and examine the relation between detection rate and false alarm rate (Egan, 1975).

7.2 One Way ANOVA Associated with the Linear Discriminant Function

We examine the effectiveness of the simplest possible data fusion scheme, combination of normalized relevance scores, with equal weights. We find that this naïve fusion approach performs better than the uninformed decision maker (who picks either scheme with equal probability) in 11,785 cases out of 16,250 cases, and even beats an oracle (for the discussion of the oracle performance measure, $\boldsymbol{E_o}$, and uninformed performance measure, $\boldsymbol{E_u}$, see Chapter 6) who uses the best of the two schemes in every case, in some 3,623 of the cases.

	Positive improvement	**Negative improvement**	**Total valid cases**
Oracle	3,623	9,171	12,794
Uninformed	11,785	3,150	14,395

Table 7.1: Number of positive cases and negative cases using different effectiveness measures. (Cases with zero improvement —i.e., fusion equal to better scheme -- are dropped)

Although the three independent variables may be interrelated and we will employ statistical techniques that incorporate these dependencies, it is often helpful to begin analyzing the differences between the positive group and negative group by examining univariate statistics.

We begin with significance tests for the equality of group means for each variable using One-Way ANOVA (the $\boldsymbol{F}$ statistic). The results are summarized in *Table 7.2,* where $\boldsymbol{P\ (S_{low})}$ is the precision at the 100th document of the poorer scheme of the two IR schemes for this topic, $\boldsymbol{P\ (S_{high})}$ is the precision at the 100th document of the better scheme, and $\boldsymbol{d\ (S_{low}\ ,\ S_{high})}$ is the normalized $\boldsymbol{z}$ score between the output lists of the two schemes.

	F	Degrees of Freedom	Significance
$P(S_{low})$	792.192	1, 12792	0.000
$P(S_{high})$	9.566	1, 12792	0.002
$d(S_{low}, S_{high})$	156.211	1, 12792	0.000

Table 7.2: Tests of equality of group means for the training data set

The F statistics of $P(S_{low})$ and $d(S_{low}, S_{high})$ are very significant, indicating that precision at the 100th document of the poorer scheme and the normalized z scores may be good discriminative variable to differentiate the positive group from the negative group. The F statistic of the variable of precision at the 100th document of the better scheme is also very significant, but comparatively speaking it may not be as discriminative as $P(S_{low})$ and $d(S_{low}, S_{high})$.

7.3 Results of the Linear Discriminant Analysis

We ask how well can we predict the effectiveness of data fusion using linear discriminants based on a combinations of the variables $P(S_1)$, $P(S_2)$ and $d(S_1, S_2)$. In this pilot test, the predictive power of the three variable is gratifyingly good. For example, assuming equal prior probability for positive and negative effectiveness, 69.1% (i.e., (6,069 + 2,775) / (9,171 + 3,623)) of the cases are correctly classified using linear discriminant analysis.

		Predicted Group Membership		**Total**
		Negative Improvement	**Positive Improvement**	
Actual Group Membershp				
Negative Improvement	**Count**	6069	3103	9171
Positive Improvement		848	2775	3623
Actual Group Membership				
Negative Improvement	**Percent**	66.2%	33.8%	100.0%
Positive Improvement		23.4%	76.6%	100.0%

Table 7.3: Using ***P*** (S_{low}), ***P*** (S_{high}), ***d*** (S_{low}, S_{high}) *as predictor variables, and 0.5 as prior probability for linear discriminant analysis, 69.1% of original grouped cases correctly classified. (Cases with zero improvement are dropped from analysis.)*

The corresponding formula of the discriminant score is as follow:

$$D = -7.877 \times P\,(S_{low}) + 4.907 \times P\,(S_{high}) - 2.115 \times d\,(S_{low}, S_{high}) + 1.052$$

Since the ranges of $P\,(S_{low})$, $P\,(S_{high})$ and $d\,(S_{low}, S_{high})$ are the same (from 0 to 1), the coefficients in the above equation are comparable without standardization. It appears that $P\,(S_{low})$ contributes most to the discriminant scores. This is not very surprising because, with the same number of degree of freedom, the ***F*** statistic of $\boldsymbol{P\ (S_{low})}$ is 792, while for $\boldsymbol{d\ (S_{low}\ ,\ S_{high})}$ it is 156, and for $\boldsymbol{P\ (S_{high})}$ comparatively it is much smaller, just 9.6.

7.4 Using ROC Curve to Represent Predictive Power

The above results are best summarized graphically, in terms of the ROC curve (receiver operating characteristic curve, see e.g., Egan 1975; Kantor 1988) which show the chance of correctly "detecting" improvement, as a function of the number of false alarms. The following is a brief explanation of ROC curve and two of its basic concepts, i.e., detection rate and false alarm rate.

Suppose we want to set up a simple a simple burglar alarm for a house by putting two cans near the entrance door, with one can on the top of the other. In principle, whenever a thief breaks in, s/he will set off the alarm, i.e., the two cans will fall. However, sometimes it does not work, let's say in average it only works in seven times out of ten. On the other hand, when there are trucks passing by the road in front of the house, sometimes they will also set off the alarm, that is, the two cans will fall, let's say in average three out of ten times when a truck passes by the cans will fall. Using the terminology of receiver operating characteristic, the detection rate of this system is 70%, while the false alarm rate is 30%.

If we use three cans instead of two, it will be more difficult for a thief to break-ins without setting off the alarm, let's say now the alarm works on eight of ten break-in; on the other hand, when a truck passes by, six time out of ten it will also set off the alarm. In other words, the detection rate of the system rises from 70% up to 80%, while the false alarm rate rises from 30% to 60%.

The more cans we use, the higher the detection rate and false alarm rate. If we balance ten cans instead of three, it is impossible for a thief to break in without setting off the alarm, however, whenever a truck passes, it will also set off the alarm without fail. In other word, the detection rate rises to 100%, and the false alarm rate also rises to 100%. If we don't put any cans near the door, the detection rate is 0, and also the false alarm rate.

Plotting all the pairs of detection rate and false alarm rate on a graph, with detection rate as the vertical axis and false alarm rate as the horizontal axis, we will get a concave curve, with false alarm rate represented as a function of detection rate. The curve will be concave because, in the beginning, it is easier for a thief to setoff the alarm than for a truck to setoff the alarm, at the end, it is easier for a truck to setoff the alarm than a thief.

Translating the above metaphor to our situation, the burglar alarm is the discriminant function, which seeks to predict whether an IR-schemes-combination is effective or non-effective. When the function *correctly* predicts that a combination will increase precision at the 100th document (compared to the best of the two IR schemes), we consider it as a detection. When the function *incorrectly* predicts an improvement, we consider it as false alarm.

The number of cans corresponds to the threshold value of the discriminant score. The more cans, the higher the detection rate, and also the false alarm rate. In our case here, we predict that all IR-schemes-combinations with discriminant score less than the threshold will have effectiveness of data fusion greater zero. The higher the discriminant score we use as cut off point, the higher the detection rate, and also the false alarm rate.

It is much easier to visualize the relationship between the discriminant function and the predictive variables with respect to the ROC curve when we have two predictive variables instead of three, as the latter would required a three-dimensional display. In the following, we use a two-dimensional display to explain the relationship using just two predictive variables.

Suppose we have only two predictive variables. We can represent all the positive and negative cases of data fusion in a two dimensional plane of the predictive variables by **+** and ***o*** respectively (*Figure 7.1*, Next page). When we move the line corresponding to constant value of the discriminant function upward and perpendicular to itself across the plane of the two variable, we will have more and more cases under the line. The increase in detection rate and false alarm rate on the ROC curve corresponds to moving the line in such a direction and labeling all the cases under the line as positive cases.

If we label all the cases under the line as positive cases, when we move the line upward, the detection rate will become higher and higher, as will the false alarm rate. For example, the detection rate of region A in *Figure 7.1* is 40%

(number of + signs in region ***A*** divided by the total number of + signs in the plane) and false alarm rate of region ***A*** is 10% (number of – signs in region ***A*** divided by total number of – signs in the plane.)

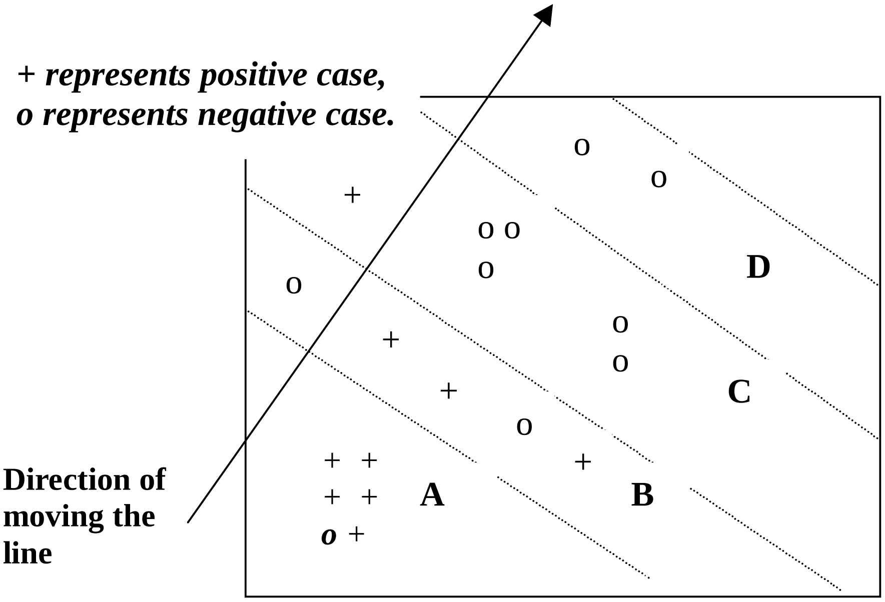
+ represents positive case,
o represents negative case.
o
o
+
o o
o
D
o
o
o
+
+
C
o
+ +
+
+ +
A
B
o +
Direction of moving the line

Figure 7.1: Moving of the discriminant line on the plane of two predictive variables. The doted lines represent the line corresponding to a constant value of the discriminant function.

As the line is moved to cover region ***B***, the detection rate rises to 80% (number of + signs in region ***A*** and ***B***, divided by total number of + sign in the plane) and the false alarm rises to 30% (number of – signs in region ***A*** and ***B***, divided by the total number of – signs in the plane.) The corresponding ROC curve is in *Figure 7.2* (Next Page).

Generally, for each strip of the plane added, the ratio of the increase in detection rate to the increase in false alarm rate is smaller and smaller. making the

ROC curve concave in shape. That is because, in the ideal case, with an optimal discriminant function, the ratio of the number of positive cases to the number of the negative cases in each added strip will be smaller and smaller. For example, the ratio in region ***A*** of *Figure 7.2* is 5, in region ***B*** it drops to 1.5, in region ***C*** it drops to 0.4, and in region ***D*** it drops to 0.

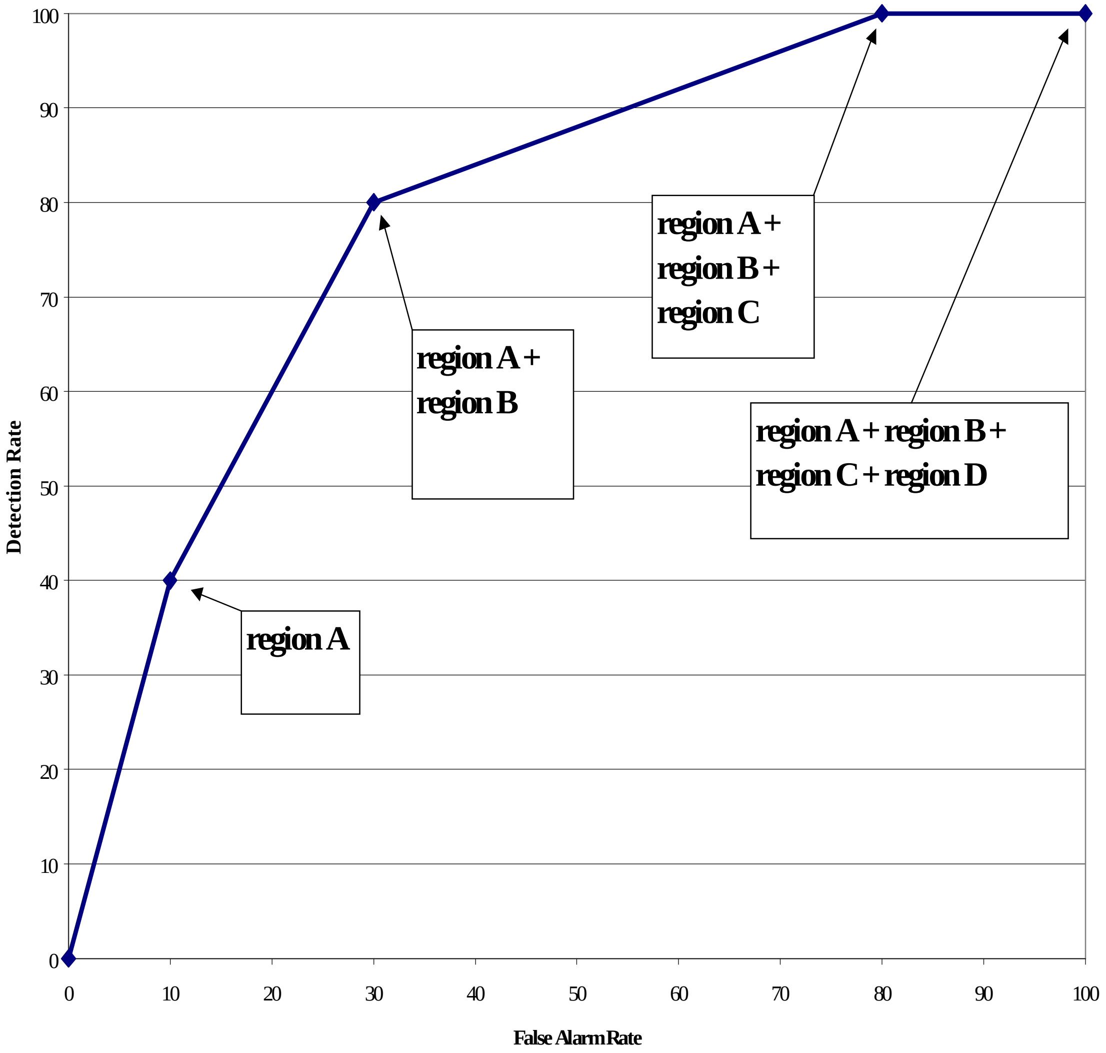

Figure 7.2: ROC curve corresponding to moving the discriminant line in Figure 7.1

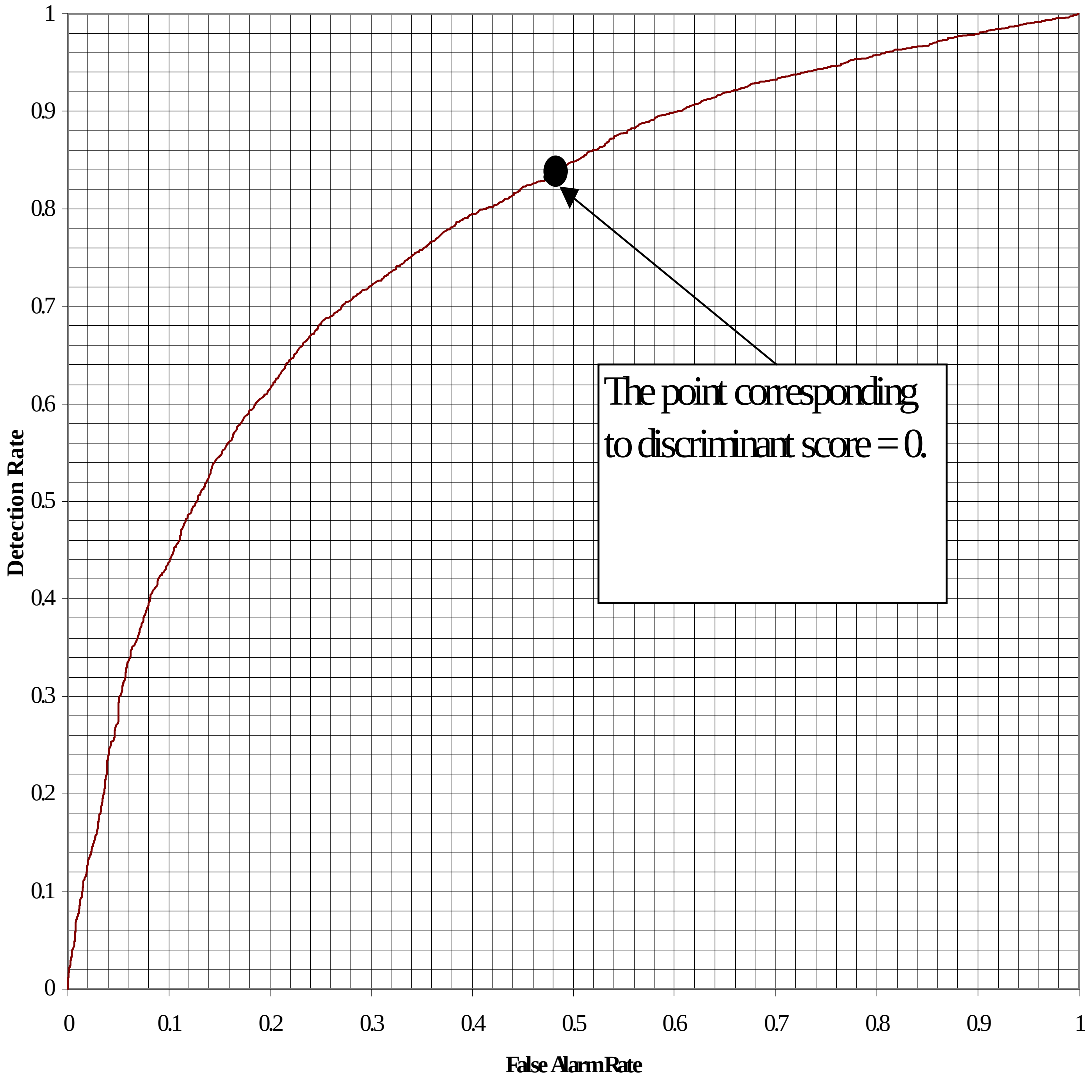

Figure 7.3: ROC Curve for discriminant scores of the training data set from 3 Independent Variables: (1) Precision at 100 th document of the poorer IR scheme; (2) Precision at 100 th document of the better scheme, and (3) Normalized inter-IR-Schemes Dissimilarity.

If we sort the discriminant scores numerically in descending order, we can plot an ROC curve to represent the detection rate and each level of different false alarm rate (*Figure 7.3*) .

In *Figure 7.3* (previous page), every point on the curve represents a possible cutoff point to discriminate between the positive group and the negative group. It has an associated detection rate and false alarm rate. The associated detection rate of a point is the ratio of the number of IR-schemes-combination that would be correctly classified, using that point as a cutoff, to the total number of combinations in the positive group. The associated false alarm rate of a point is the ratio of the number of IR-schemes-combinations that would be incorrectly classified as member of positive group, using that point as a cutoff, to the total number of combinations in the negative group.

The closer an ROC curve comes to the ideal point (100% detection, 0% false alarm rate), the better the performance of the predictor(s). The ROC curve of the three predictive variables in *Figure 7.3* shows that the discriminant score may be a good classification index in the sense that, as the threshold for predicting effectiveness is varied smoothly, the curve rises in a concave fashion and the detection rate is sometimes substatially higher then false alarm rate. We see in *Figure 7.3* (previous page) that the linear discriminant formed using the efficacies and the dissimilarity, achieves a detection rate of 60%, at a false alarm rate of only about 18%. In other words, with this threshold setting it would label 60% of the 3,623 cases showing improvement correctly, and would label only 18% of the cases with no improvement, incorrectly. As the detection rate rises to 80%, the false alarm rate rises more rapidly, to 40%, half of the detection rate. This ROC curve demonstrates that the two proposed conditions for effective data fusion are statistically valid predictors for effectiveness of data fusion in IR.

7.5 Discriminant Analysis using Other Discriminant Functions

In the above analysis, the discriminant function is based on a linear combination of the three independent variables. It will be interesting to see and compare the ROC curves produced by other discriminant functions using just one, or two but not all three of the proposed independent variables.

We have done six additional discriminant analyses based on six different combinations of the three independent variables, they are:

$$D_1 = \beta_{01} + \beta_{low\_1} P(S_{low})$$

$$D_2 = \beta_{02} + \beta_{high\_2} P(S_{high})$$

$$D_3 = \beta_{03} + \beta_{33} \; d(S_{low}, S_{high})$$

$$D_4 = \beta_{04} + \beta_{low\_4} P(S_{low}) + \beta_{high\_4} P(S_{high})$$

$$D_5 = \beta_{05} + \beta_{low\_5} P(S_{low}) + \beta_{35} \; d(S_{low}, S_{high})$$

$$D_6 = \beta_{06} + \beta_{high\_6} P(S_{high}) + \beta_{36} \; d(S_{low}, S_{high})$$

If the ROC curve for one function lies above that for another, the method corresponding to the higher curve is preferred, no matter what the preferences or utility schemes of the user or decision maker. *Figure 7.4* (next page) shows the ROC curves for predicting when fusion will do even better than the Oracle.

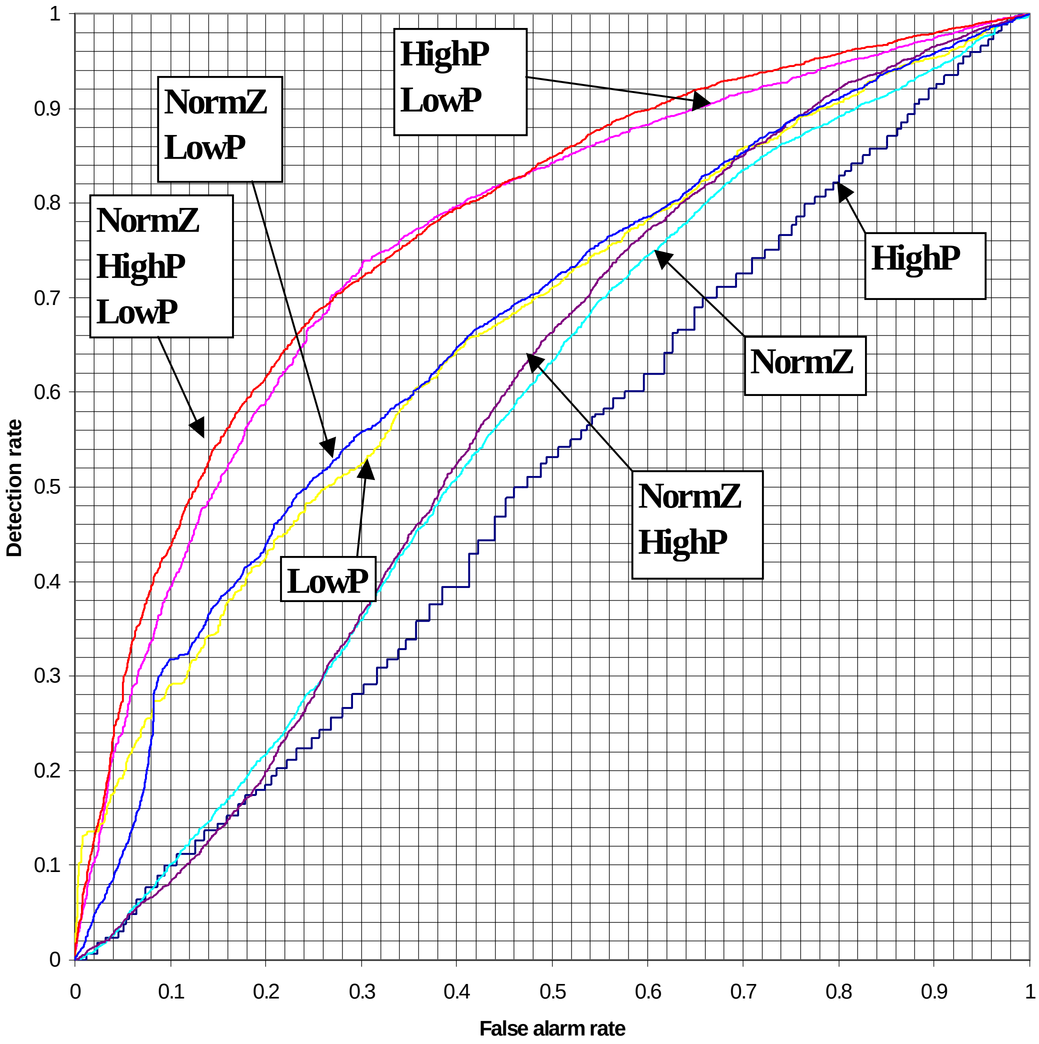

Figure 7.4: ROC curves of seven discriminant functions for the training data set. ***HighP*** *and* ***LowP*** *are the precision rate of two IR schemes with better precision and poorer precision at the 100th document respectively;* ***NormZ*** *is the normalized dissimilarity between the two IR schemes output lists.*

There are seven curves in *Figure 7.4*, corresponding to the seven sets of predictors. Each predictor is either a single independent variable, or a linear combination of the independent variables. ***High P*** and ***Low P*** stand for the efficacy (precision at the 100th document) of an underlying scheme which has higher performance and lower performance respectively; while ***Norm Z*** stands for normalized ***z*** scores. We see that the two curves which use

1. Efficacies of two IR schemes (i.e., ***Low P, High P***), and
2. Efficacies of two IR schemes and normalized dissimilarity (i.e., ***Low P, High P, Norm Z***)

achieve dominant performance. The largest difference is in the vicinity of a 20% to 40% false alarm rate, where detection increases 18% (from about 42% to 60% and 62% to 80%) compared to the third and fourth curves (i.e., ***Low P*** curve, and ***Low P + Norm Z*** curve). In most of the cases, using the combination of the efficacy measures (i.e., ***Low P*** and ***High P***) and dissimilarity measure (i.e., ***Norm Z***) can achieve better detection rate than only using the efficacy measures.

However, in the range of 71% - 80% detection rate, using only the efficacy measures can achieve a very slightly higher detection rate than using the combination of the efficacy measure and dissimilarity measure. Of course this results from the fact that each discriminant analysis is trying to do the best it can overall. If we asked the scheme with the three variables to do its best in the 71%-80% detection region it will beat the scheme with only using the efficacy measure.

Most significantly, using only the dissimilarity measure cannot predict the effectiveness very well, concerning the negative experience of Ng and Kantor

(1996).

These results, obtained without any effort to tune or optimize the fusion rule, leave us optimistic about being able to predict when data fusion will be effective in IR. Thus, again, the conditions of efficacy and dissimilarity appears to be valid criteria for predicting effectiveness of data fusion.

Chapter 8:

Discriminant Power of the Efficacy of Constituent IR Schemes

8.1 Contributions of Different Efficacy Levels of the IR Schemes to the Overall ROC Curve

From the previous test, we found that the ROC curve of the discriminant function of the linear combination of $\boldsymbol{P_{low}}$, $\boldsymbol{P_{high}}$, and normalized $\boldsymbol{z}$ has the best discriminative power compared with other six combinations. Now we investigate if all the cases have the same contribution to the detective power of this ROC curve.

In our discussion of the condition of efficacy (Chapter 3), we suggested that because of different baselines for measuring relative improvement, it is more difficult to detect the improvement of data fusion between a capable IR scheme and a very incapable IR scheme. Accordingly we can expect the contribution of cases of different combination of efficacy levels to the overall ROC curve will be different, with the contribution from cases of combination between a capable IR scheme and a very incapable IR scheme be the least. This idea can be tested by isolating the three groups of cases from all the cases composing the overall ROC curve:

1. High-Low Group consists of those cases that are results of data fusion between one capable IR scheme and one incapable IR scheme;
2. Low-Low Group consists of those cases that are the results of data fusion between two incapable IR scheme;
3. High-High Group consists of those cases that are the results of data fusion between two capable IR schemes.

Using the original discriminant score from the discriminant function of the linear combination of $\boldsymbol{P_{low}}$, $\boldsymbol{P_{high}}$, and normalized $\boldsymbol{z}$, we plot the three ROC curves formed by the High-High group, the Low-Low group, and the High-Low group separately (*Figure 8.1,* next page).

In *Figure 8.1*, the High-High group consists of cases that are the results of data fusion between two IR schemes which have precision at the 100th document higher than or equal to 60%; the Low-Low group consists of cases that are the results of data fusion between two IR schemes which have precision at the 100th document lower than or equal to 40%; the High-Low group consists of cases that are the result of data fusion between an IR scheme with precision at 100th document higher than or equal to 60% and another IR scheme with precision at 100^{th} document lower than or equal to 40%. The All group consists of all cases.

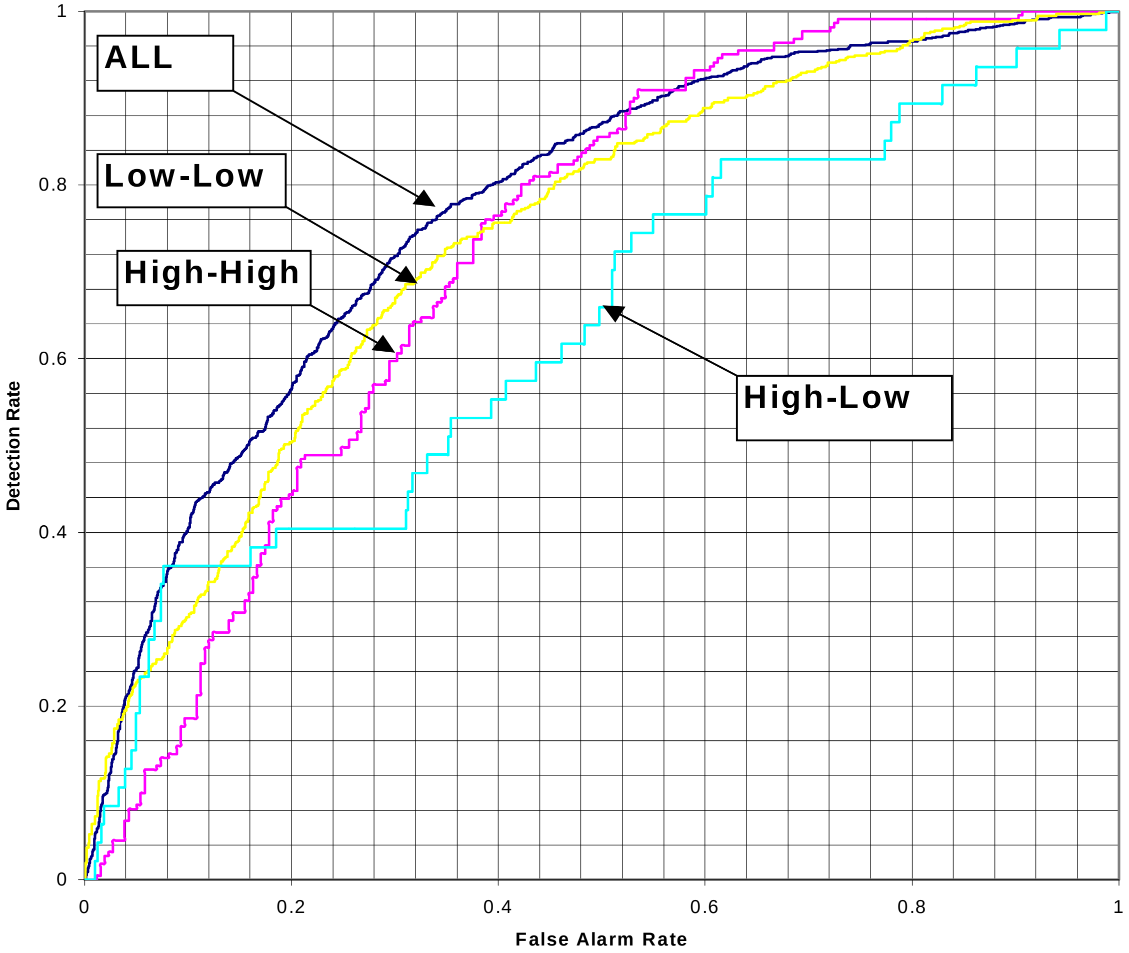

Figure 8.1: The ROC Curves of 4 groups: (1) group consists of all cases; (2) high-high group; (3) low-low group; and (4) high-low group, for the training data set.

The ROC curve of the High-Low group is much lower than the other three ROC curves. In fact, it is very close to the diagonal of the graph, indicating that the false alarm rate is almost always the same as the detection rate. Its predictive power is very low, for example, at the detection rate of 80%, the other three curves only have false alarm rate from approximately 40% to 45%, but the ROC curve of High-Low group is about 75%. This graph supports the our suggestion that practically it is more difficult to detect the improvement of data fusion between a capable IR scheme and a very incapable IR scheme.

8.2 The Discriminative Power of the Ratio of Precisions

There are two interesting observations from *Figure 8.1*. The first one is, the ROC curve of the High-High group does not do better than the original ROC curve (the ROC curve of all cases) until the point at about 88% detection rate and 52% false alarm rate. In other words, using any discriminant scores below the discrminant scores corresponding to the point of 88% detection rate and 52% false alarm rate as cutoff point, the ratio of positive cases to negative cases for all cases are always higher than the ratio of positive cases to negative cases for the High-High cases.

The second interesting observation is, the ROC curve of the Low-Low group is pretty good, in fact it is better than the ROC curve of the High-High group until the point at about 67% detection rate and 38% false alarm rate. In other words, using any discriminant scores below the discriminant score corresponding the point of 67% detectoin rate and 38% false alarm rate as cutoff point, the ratio of positive cases to negative cases for Low-Low cases are always higher than the ratio of positive cases to negative cases for the High-High cases.

These two observations, with the diagonal ROC curve for the High-Low group, lead us to investigate whether the ratio between the two variables $\boldsymbol{P_{low}}$ and $\boldsymbol{P_{high}}$ may have better discriminative power than a linear combination of them.

The discriminative power of the ratio of precisions can be visualized by plotting two scatter plots of $\boldsymbol{P_{low}}$ against $\boldsymbol{P_{high}}$ (*Figure 8.2* and *Figure 8.3*).

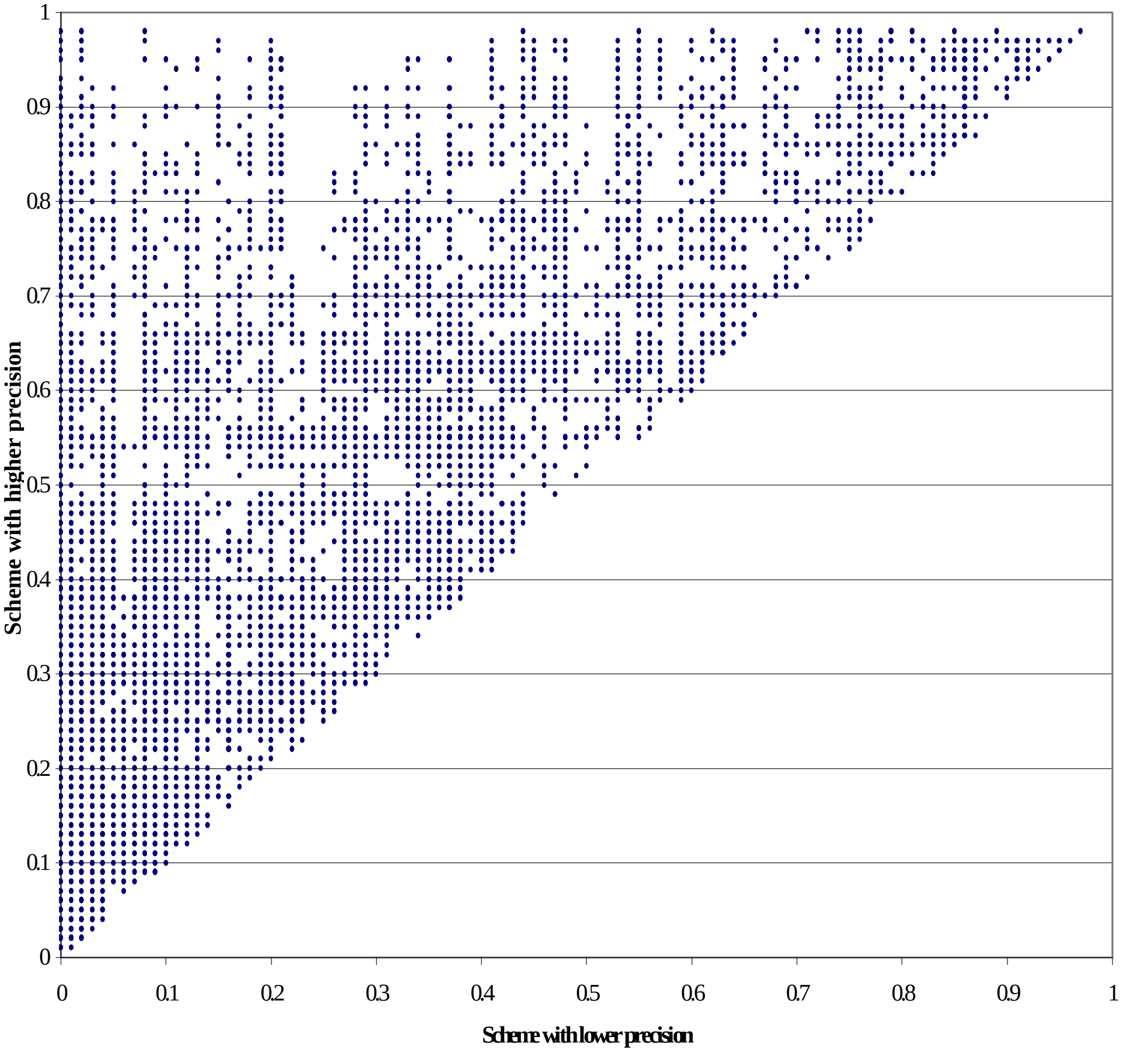

Figure 8.2: Scatter plot of precision at 100th document of two IR schemes for all negative cases of the training data set. X-axis is precision at 100th document of the poorer scheme and Y-axis is the precision at 100th document of the better scheme.

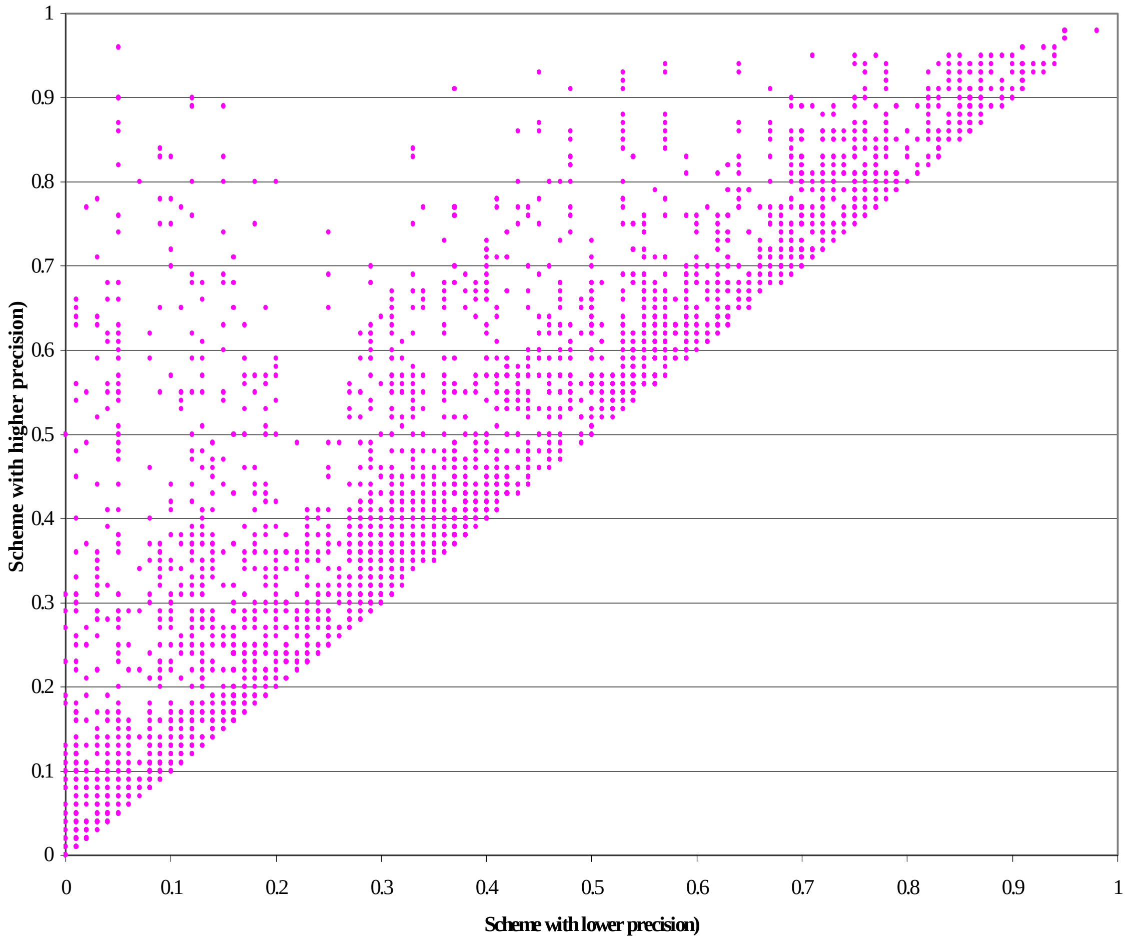

Figure 8.3: Scatter plot of precision at 100th document of two IR schemes for all positive cases of the training data set. X-axis is precision at 100th document of the poorer scheme and Y-axis is the precision at 100th document of the better scheme.

Comparing *Figure 8.2* and *Figure 8.3*, one can see that although for the negative cases, there is no clear distinguishable pattern, for the positive cases, the cases tend to concentrate at the diagonal, i.e., ratio of precisions ($\boldsymbol{P_{low} / P_{high}}$) ≈ 1 .

To investigate which one of the following two functions will have a better discriminative power:

1. The discriminant function based on a linear combination of $\boldsymbol{P_{low}}$, $\boldsymbol{P_{high}}$ and normalized ***z*,** or
2. The discriminant function based on a linear combination of $\boldsymbol{P_{low} / P_{high}}$ and normalized $\boldsymbol{z}$,

we can compare the eigenvalues of the two discriminant functions. The eigenvalue of a discriminant function is the ratio of the between-groups sums of squares to within-group sums of squares. Therefore, a larger eigenvalue is associated with a better discrminant function.

The eigenvalue of the discriminant function based on a linear combination of $\boldsymbol{P_{low}}$, $\boldsymbol{P_{high}}$ and normalized $\boldsymbol{z}$ is “0.393”, while the eigenvalue of the discriminant function based on a linear combination of $\boldsymbol{P_{low} / P_{high}}$ and normalized $\boldsymbol{z}$ is “0.483”. It is clear that the ratio of precision may be a better discriminant variable. *Figure 8.4* (next page) is the comparison between the two corresponding ROC curves, more details of the discriminant function based on a linear combination of $\boldsymbol{P_{low} / P_{high}}$ and normalized $\boldsymbol{z}$ will be provided in Chapter 10.

In *Figure 8.4*, the solid curve uses the combination based on the ratio of precision of the two IR schemes and the normalized dissimilarity, and achieves dominant performance compared to the light curve which uses the combination based on the separate precision of the two IR schemes and the normalized dissimilarity. The largest different is in the vicinity of a 30% false alarm rate, where detection increases from about 70% to 80%, a relative improvement of some 14%. This is particularly striking since it is accomplished with a **reduction** in the number of independent variables.

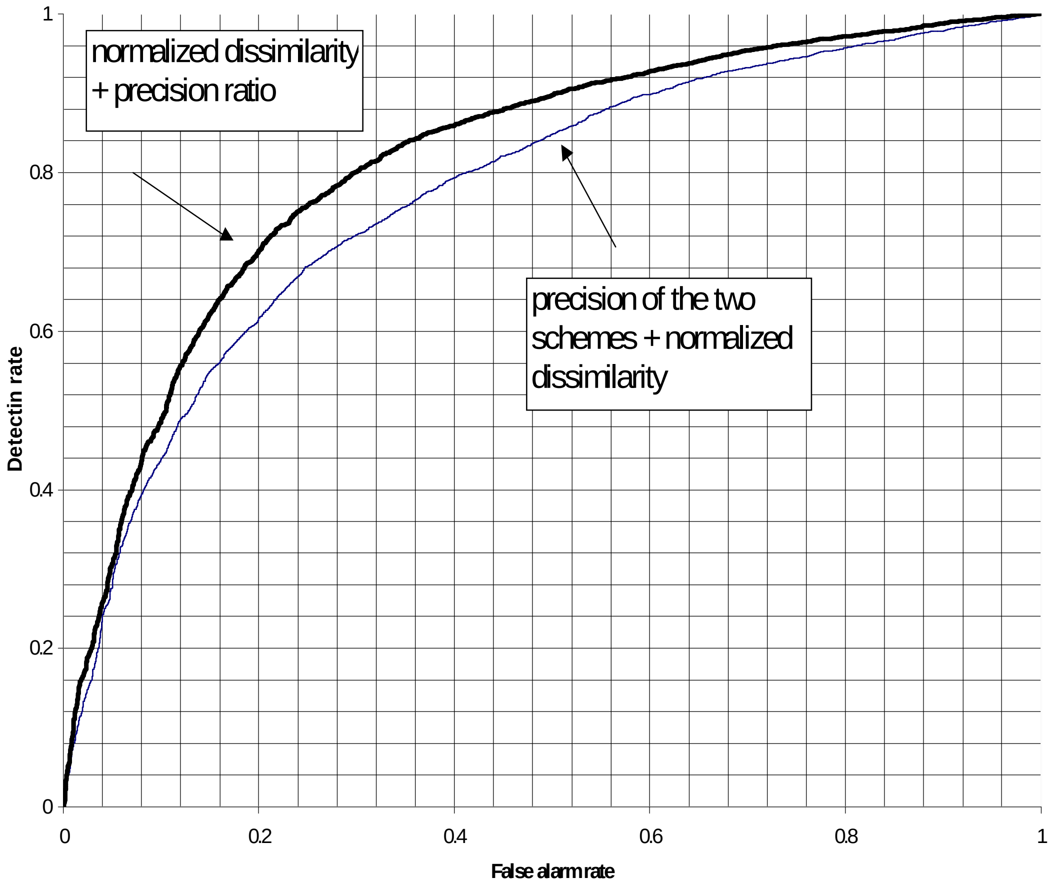

Figure 8.4 Comparison of the ROC curves, of the training dat set, based on two different discrimininant functions.

According to the above analysis, the ratio of precision is a better discriminant variable. Therefore, in the training stage of our experiment, we use ratio of precision as the variable to investigate how the condition of efficacy, in combination with the condition of dissimilarity, affects the effectiveness of data fusion.

Chapter 9:

Exploratory Analysis of Ratio of Precisions, Normalized Dissimilarity, and Data Fusion Effectiveness

9.1 Examinations using Scatter Plots of the Variables

In the previous Chapters, two variables have been identified as independent variables with good discriminating power for predicting the effectiveness of data fusion using a simple symmetrical fusion rule between two IR schemes. They are:

(1) the normalized dissimilarity of the two IR schemes, and

(2) the ratio of precision at the 100th document of the two IR schemes.

Using the training data set (i.e., all output lists produced by the 26 IR schemes that have participated the TREC 4 routing tasks), when we combined all the output lists topic by topic for the 50 TREC 4 routing topics using the simple symmetrical fusion rule, there are 16,250 combinations. Out of these 16,250 fused cases, 3,623 cases have precision at the 100th document higher than the best of the two IR schemes ("positive cases"), 9,171 cases have precision at the 100th document lower than the best of the two IR schemes ("negative cases"). If we plot a scatter plot of normalized dissimilarity vs. ratio of precision for these 3,623 and 9,171 cases, we can visually look for some relationship between normalized dissimilarity and ration of precision which will tell us something about how to combine these two independent variables.

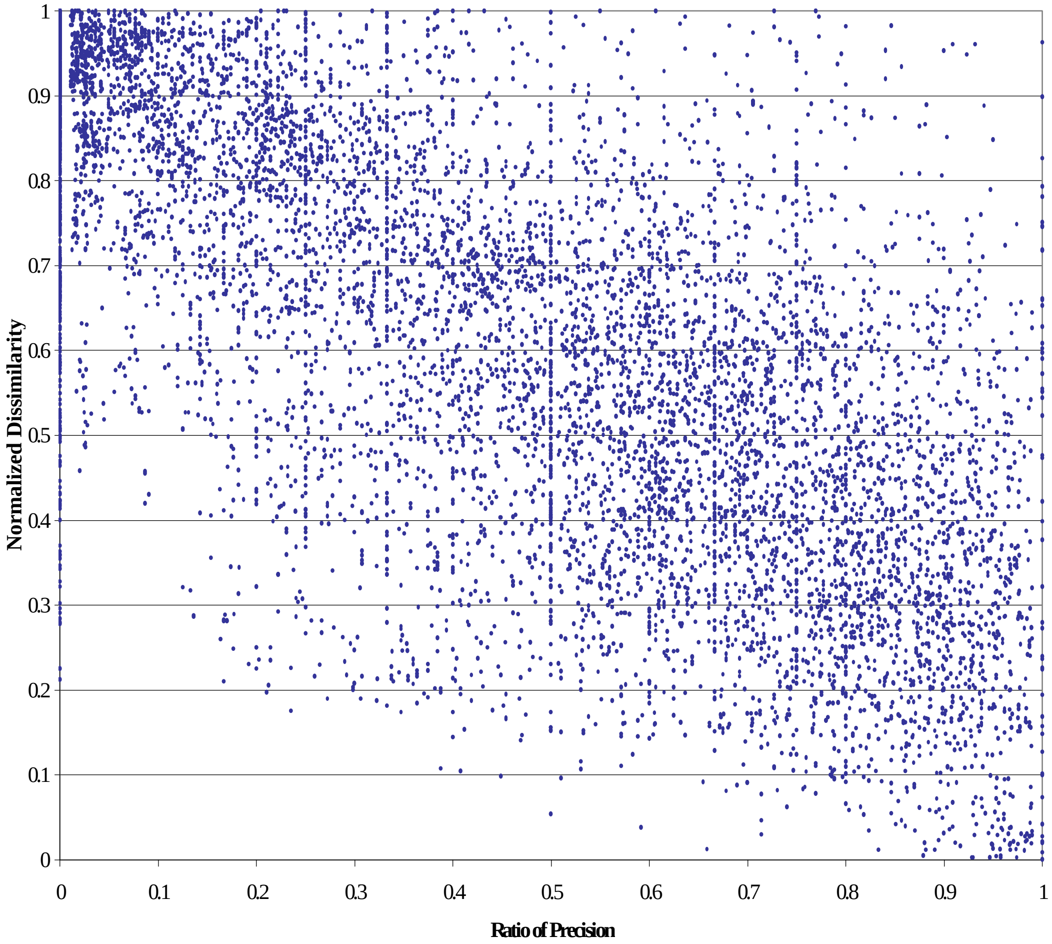

Figure 9. 1 : Scatter plot of ratio of precisions vs. normalized dissimilarity for negative cases of the training data set.

Figure 9.1 is the scatter plot of precision ratio vs. normalized dissimilarity for the negative cases. From *Figure 9.1*, we can see that there are very few negative cases in the region of small precision ratio and small normalized dissimilarity. For example, there is no case in the region of ***(ratio of precision < 0.2, normalized dissimilarity < 0.2)*** , and there are only about 10 cases in the region of ***(ratio of precision < 0.4, normalized dissimilarity < 0.4)***. There are also comparatively very few cases in the region of high precision ratio and high normalized dissimilarity (i.e., top right corner of the graph). In fact, it appears that the dots representing all negative cases tend to scatter evenly around the line defined by the equation:

Normalized Dissimilarity + Ratio of Precision = 1

We see from *Figure 9.1* that for the negative cases, when precision ratio approaches 1, the normalized dissimilarity approaches 0; when the ratio of precision approaches 0, the normalized dissimilarity approaches 1. That means, for the negative cases:

(1) When the performance (i.e., precision at the 100th document) of the two IR schemes are more or less the same, their output lists are more similar to each other, and vice versa;

(2) When the performance (i.e., precision at the 100th document) of one IR scheme is much better than the other, their output lists are more dissimilar to each other, and vice versa.

Intuitively the above observations are not completely surprising because when two IR output lists contain similar documents with similar ranking, one will expect to see similar performance (i.e., high ratio of precision). On the other hand, when two IR output lists are dissimilar in terms of common documents and ranking among documents (i.e., high normalized dissimilarity), one will expect to see different performance.

However, let’s see if these observations about the negative cases are also true in the positive cases. *Figure 9.2* (next page) is the scatter plot of the ratio of precision vs. normalized dissimilarity for the positive cases.

From *Figure 9.2*, as in *Figure 9.1*, we also see that there are very few positive cases in the region of small precision ratio and small normalized dissimilarity. For example, again, there is no case in the region of ***(precision rate < 0.2, normalized dissimilarity < 0.2)***, and there are only 2 cases in the region of ***(precision rate < 0.4, normalized dissimilarity < 0.4)***.

It is also true that there are comparatively fewer cases in the region of high precision ratio and high normalized dissimilarity. However, there are many more dots in that region for the positive cases than for the negative cases. In fact, the dots representing all positive cases do not tend to scatter evenly around a straight line, but concentrate approximately in the region of ***(ratio of precision > 0.8).***

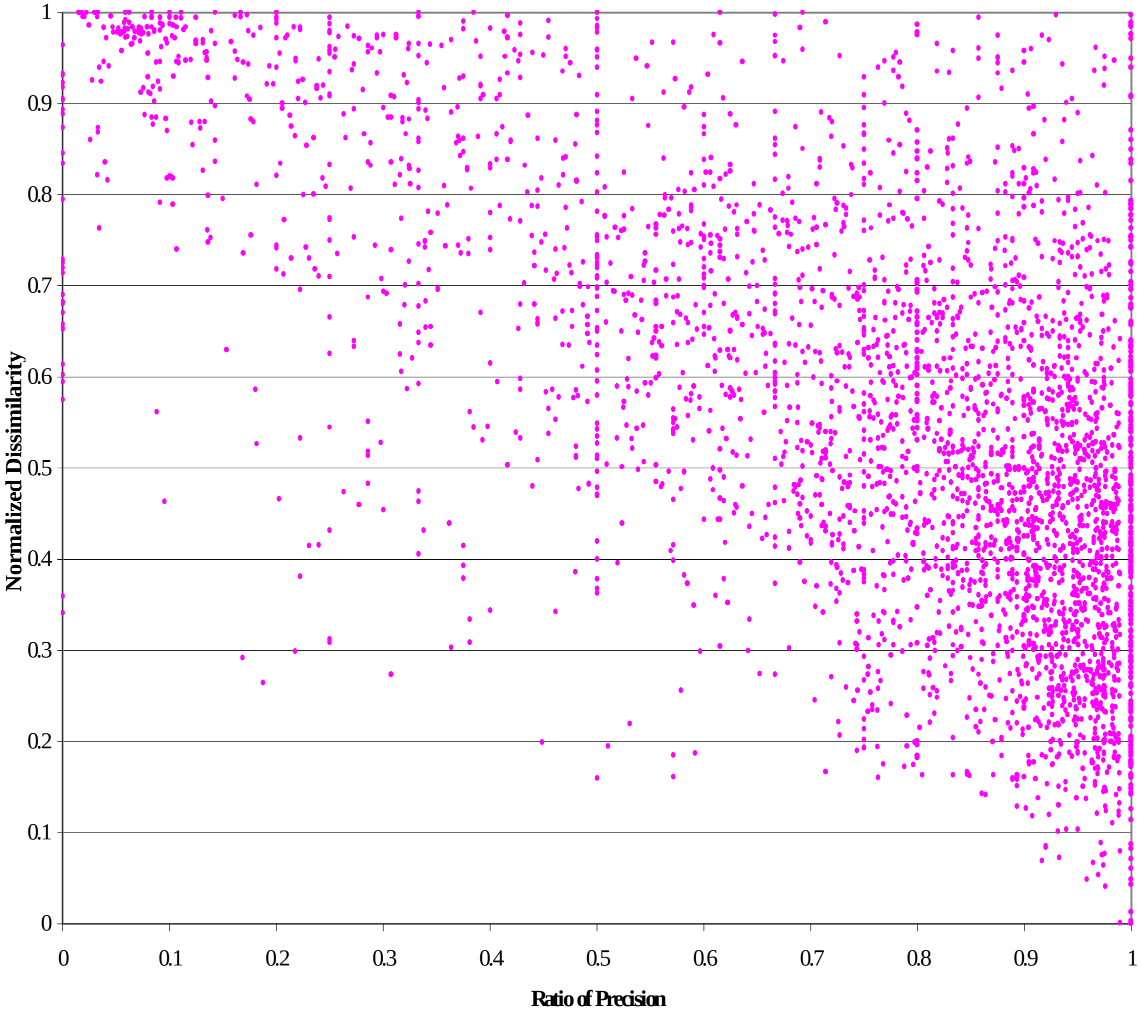

Figure 9. 2 : Scatter plot of precision ratio vs. normalized dissimilarity for positive cases of the training data set.

In addition, while the distribution of dots in the scatter plot of negative cases is quite symmetrical around the straight line

Normalized Dissimilarity + Precision Ratio = 1

the distribution of dots in the scatter plot of the positive case is concentrated in the region above the same straight line. That means, compared to the negative cases, the majority of the positive cases are in the region of higher precision ratio and higher normalized dissimilarity; and there are comparatively very few cases in the region of low precision ratio and low normalized dissimilarity. This observation supports the two proposed conditions for effective data fusion in IR — low similarity and comparable precision.

9.2 Frequency Distributions of Subgroups

We can “bin” the cases into 100 square regions by choosing 10 ranges of normalized dissimilarity and 10 ranges of ratio of precision. Comparing the frequency distributions in these bins gives another picture of how normalized dissimilarity relates to ratio of precisions in the two situations.

Figure 9.3 (next page) is the frequency for the negative cases by normalized dissimilarity and ratio of precisions. The cases are grouped into 100 small regions. The height of the bar at each region represents the number of cases occurring in that region.

From *Figure 9.3*, one can also see that the cases are scatter more or less evenly around the negative diagonal. In addition, there are more than 1,000 cases located in the region of ***(0 <= ratio of precisions < 0.1, 0.9 <= normalized dissimilarity < 0.1)***, while the other 99 regions have lower frequencies. These

1,000 plus cases are from combinations of a comparatively much more capable IR scheme and a much less capable IR scheme (i.e., very low ratio of precisions), with very dissimilar outputs (i.e., very high normalized dissimilarity).

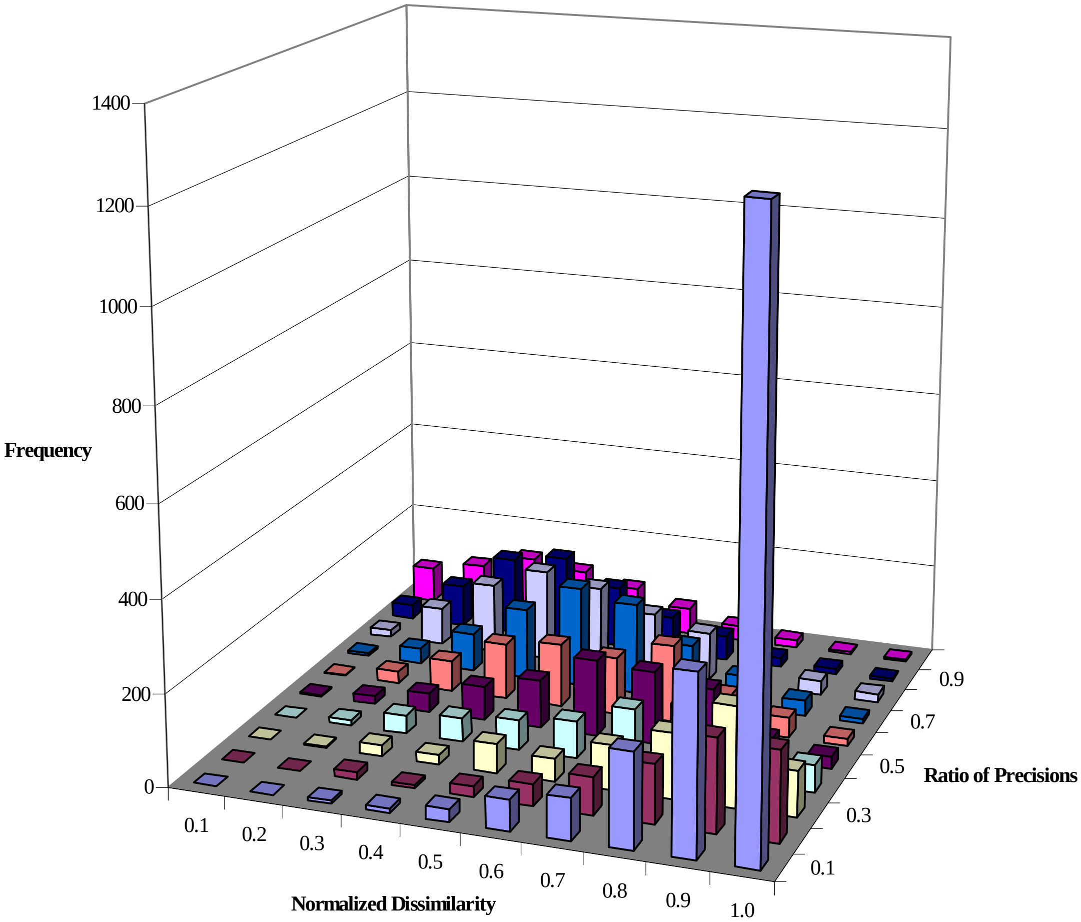

Figure 9.3: Frequency distribution of negative cases, of the training data set, by normalized dissimilarity and ratio of precisions.

The above observation seems to suggest that very dissimilar IR schemes do not necessary lead to effective data fusion. However, it does not imply that combination of very high dissimilarity IR schemes is more likely to have negative effectiveness of data fusion. Empirically the probability depends on the ratio of positive cases to negative cases in the region of very high normalized dissimilarity, not on the ratio of the number of negative cases in that region to the total number of negative cases.

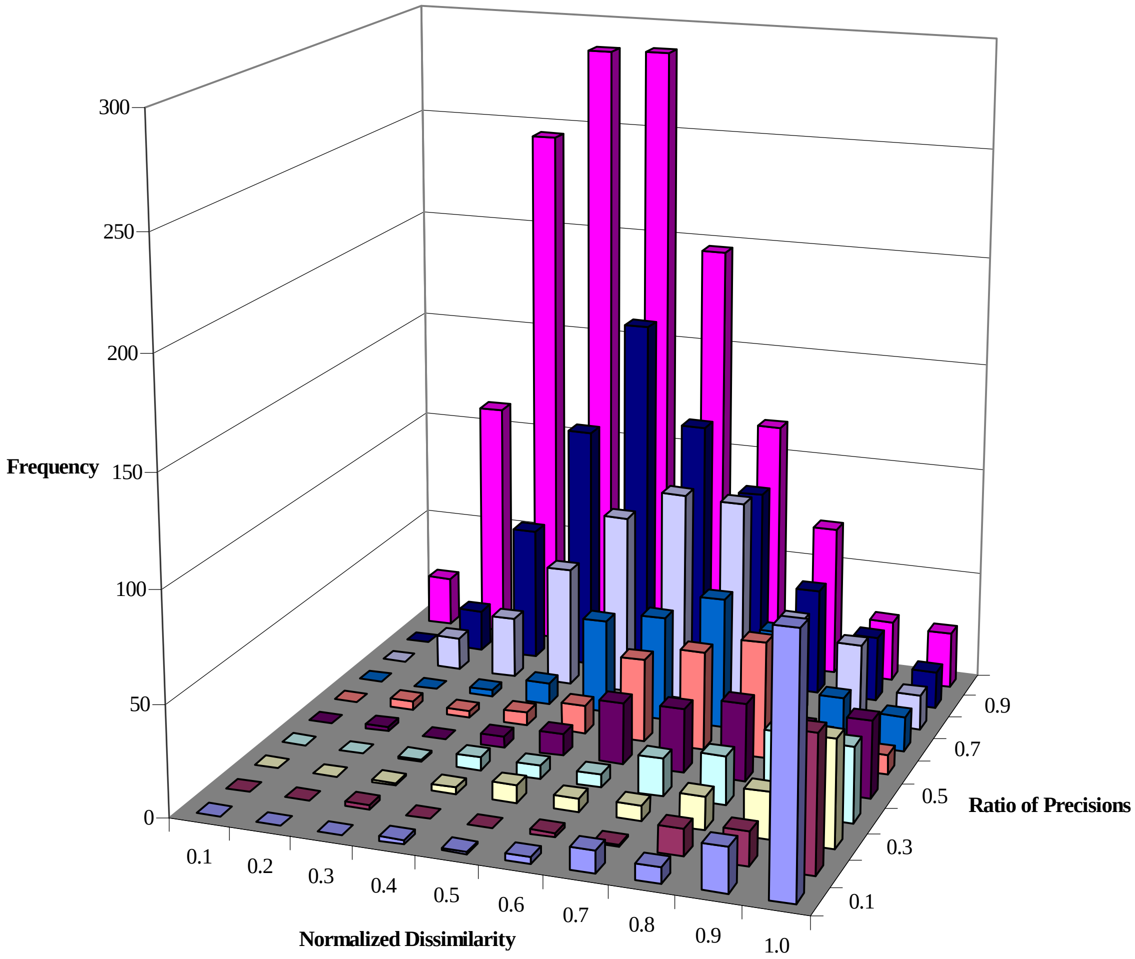

Figure 9.4: Frequency distribution of positive cases, of the training data set, by normalized dissimilarity and ratio of precisions.

Now let us examine the frequency distribution of the positive cases. *Figure 9.4* is the frequency distribution by normalized dissimilarity and ratio of precisions for the positive cases grouped into 100 bins. In sharp contrast to the negative cases, the highest frequency bars of the positive cases are not located in the lowest precision ratio regions ***(ratio of precisions* < *0.1)*** but in the highest precision ratio regions ***(ratio of precisions* > *0.9)***. This observation reflects the discriminant power of the precision ratio variable.

In addition, the highest bars are not in the region of highest dissimilarity but in the region of medium dissimilarity. It seems that, in the region of high ratio of precisions, combination of IR schemes with medium dissimilarity is more likely to have effective data fusion than IR schemes with very high dissimilarity.

9.3 Examination of the Region of High Ratio of Precisions

From *Figure 9.2* and *Figure 9.4*, we see that many positive cases are approximately located in the region of high ratio of precision, for example, the region with ratio of precision than 0.8. Let's examine that region more carefully.

Although overall, there are many more negative cases than positive cases, in the region of ***(ratio of precisions* > *0.8)***, positive cases are more common than negative cases. In this region, there are 2,114 positive cases, 1,518 negative cases, and 1,027 cases with effectiveness of data fusion equal to 0. For these 2,114 + 1,518 + 1,027 = 4,659 cases, we plot a scatter plot of the data fusion effectiveness (i.e., $\boldsymbol{E_o}$, see Chapter 6 Section 6.1) vs. normalized dissimilarity (*Figure 9.5*, next page).

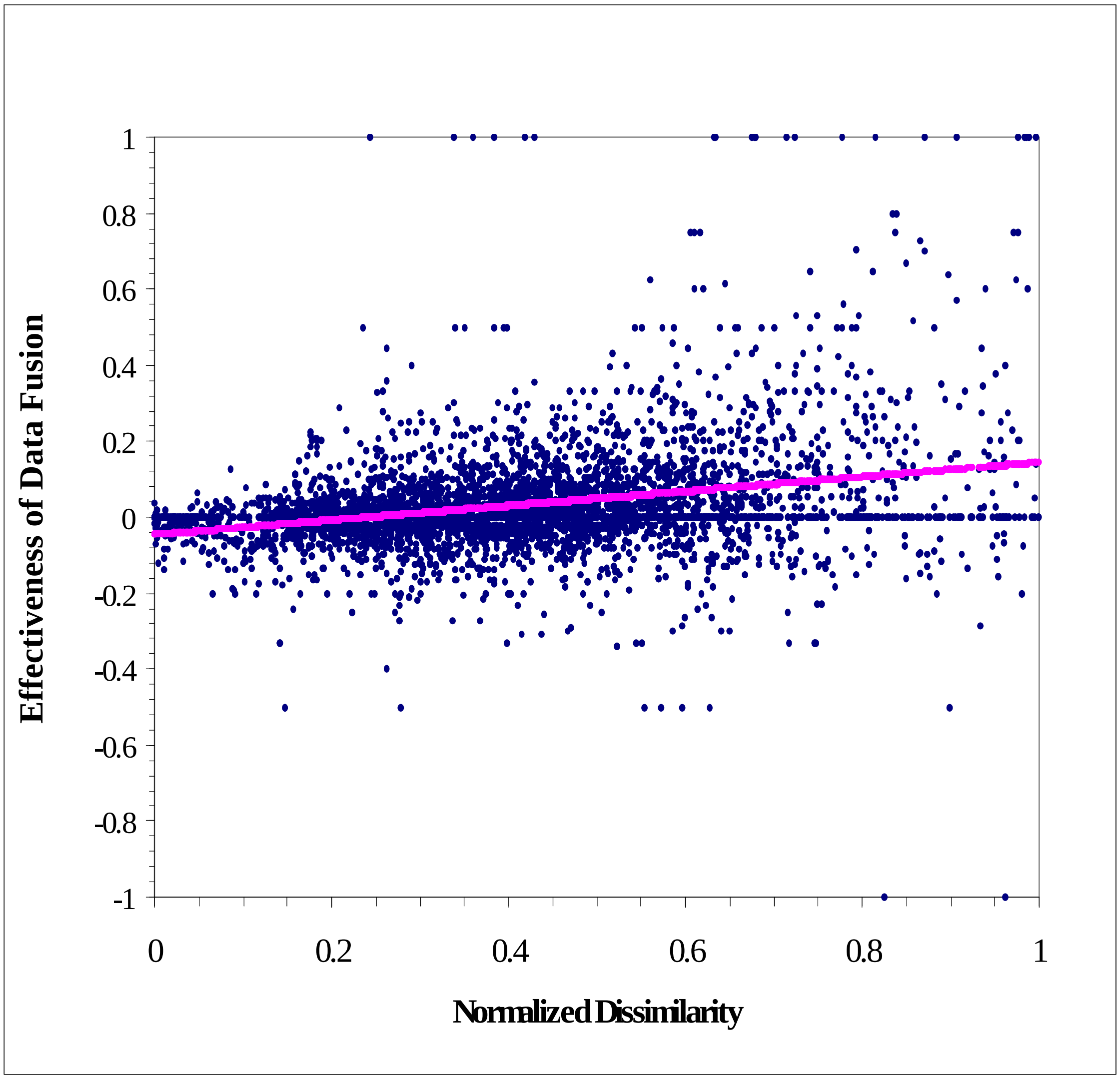

Figure 9.5: Scatter plot of normalized dissimilarity vs. effectiveness of data fusion for cases with ratio of precisions greater than 0.8. The straight line is the linear regression of normalized dissimilarity on effectiveness in this region.

From *Figure 9.5*, if there were no regression line, it would not be easy for us to see whether effectiveness is related to normalized dissimilarity. With the regression line, we can see that normalized dissimilarity is positively correlated with effectiveness.

The slope of the regression line is 0.2 (with a standard error of 0.01), indicates that when normalized dissimilarity increases by 10%, on the average, effectiveness of data fusion will increase 2% . However, there is a lot of variation.

As stated in Chapter 6 Section 6.1, effectiveness of data fusion is positive only if $S_1 \; f \; S_2$ is better than the best scheme because we use the percent improvement over the best scheme to measure the effectiveness. In other words, when the normalized dissimilarity of two IR schemes increases by 10%, using simple symmetrical data fusion, on the average, we can expect about 2% improvement compared to the best of the two schemes.

We can use the ratio between the estimated slope and its standard error to test the null hypothesis that there is no linear relationship between normalized dissimilarity and effectiveness of data fusion in this region. The distribution of this ratio is the Student's t distribution with $N - 2$ degree of freedom, where N is equal to the number of cases. With 4659 cases, the degree of freedom is 4657, the t statistic is 0.2/0.01 = 20 and the P-value is 1.03×10^{76} . (P-value is the smallest level of significance at which null hypothesis can be rejected.)

Even though a slope of 0.2 is not large It is larger than the slope if we use all 16250 cases of training data in the regression analysis. Using all the cases in the regression analysis, the slope is 0.17, with a standard error of 0.05. In other words, in the region of high ratio of precisions, increase in normalized dissimilarity will cause more increase in effectiveness.

We can also test the null hypothesis of there is no linear relationship between effectiveness of data fusion and normalized dissimilarity for the linear regression analysis of all 16250 cases. In the analysis normalized dissimilarity of all cases regressed on effectiveness of data fusion, the ratio between the estimated slope (0.17) and its standard error (0.05) is 3.4. With 16248 degree of freedom, the P-value of ***t*** = 3.4 is 0.0003, still very significant. In other words, it is justifiable for us to assume, even normalized dissimilarity does not varies too much with the effectiveness of data fusion, statistically they are positively related to each other, and in average, when the normalized dissimilarity of two IR schemes increases by 0.1, using simple symmetrical data fusion, we can expect about 1.7 % improvement compared to the best of the two schemes.

Chapter 10:

Comparison of Three Statistical Analysis Methods for Predicting Effectiveness of Data Fusion in IR

In the previous chapters two variables, ratio of precision and normalized dissimilarity, have been identified as predictive variables for the effectiveness of data fusion. In this chapter, we investigate how the training data set and these two variables to predict the effectiveness of data fusion in a test data set. The training data set produced 16,250 fused document number lists. Each list contains 1,000 document numbers. The lists were generated from all pairs of outputs of twenty-six IR schemes for fifty TREC 4 routing topics. The simple (symmetrical, addition, and non-analytical) fusion rule was used in generating the 16,250 lists.

We have applied three statistical analysis techniques on the training data. They are:

1. linear discriminant analysis;
2. multiple linear regression analysis;
3. logistic regression analysis

We used the SPSS for Windows program (version 6.1) in my statistical analyses.

We want to see which of these three statistical analysis techniques is most useful in assigning a classification score to each of the 16,250 output lists. The classification scores will be used to generate ROC curves. The statistical analysis technique which produces the ROC with highest detection rate at each false alarm rate will be used to predict the effectiveness of data fusion on the testing data set.

10.1 Linear Discriminant Analysis

We have used Linear discriminant analysis in previous chapters (i.e., Chapter 7 and Chapter 8), so in here we will not discuss the details of the conceptual framework and justification of applying this method on our data set but just give brief summary when necessary.

We seek a linear combination of the independent variables ***r*** (ratio of precision) and ***z*** (normalized dissimilarity) that will discriminate between the ***a priori*** defined groups in the training data set in such a way that the ratio of ***between-group sum of squares*** to the ***within-group sum of squares*** is maximum. Any other linear combination of ***r*** and ***z*** will have a smaller ratio.

Using the strictest measure $\boldsymbol{E_o}$ (this compares fusion to an oracle, see Chapter 6) as the measure for the effectiveness of data fusion, we divide the all the cases into two groups. To do this, we replace the numerical value of $\boldsymbol{E_o}$ by its sign.

Thus, the positive group contains all the cases in which symmetrical data fusion by sum of normalized relevancy scores improve precision at the 100th document; the negative group contains all the cases in which the precision at the 100th document of the fused list produced by the symmetrical data fusion is worse than the better scheme of the two original IR schemes. We drop those cases with $\boldsymbol{E_o}$ equal to zero from analysis. There are 12,794 cases left in analysis.

We begin our analysis with significance tests for the equality group means for each of the two variables using One-Way ANOVA (***F*** statistic). The results are summarized in *Table 10.1*, where ***r*** represents ratio of precision, ***z*** represents normalized dissimilarity.

	F	Degree of Freedom	Significance
z	156.211	1, 12792	0.000
r	2678.608	1, 12792	0.000

Table 10.1: Tests of equality of group means for the training data set

The ***F*** statistics of ***r*** and ***z*** are very high and very significant, indicating that it is unlikely for positive group to have the same means, for either valuable, on the discriminant function as the negative group. In other words, ***r*** and ***z*** can be used individually as discriminative variables. In addition, the ***F*** statistic for ***r*** is much higher than ***z***, suggesting that precision ratio may have much higher discriminative power than normalized dissimilarity.

The formula for the discriminant score is

$$\boxed{\textbf{\textit{Discriminant Score}} = \mathbf{3.195\, z + 4.643\, r - 4.36}}$$

Larger values of this scale are more like to yield effective data fusion.

The percentage of cases classified correctly is often taken as an index of the effectiveness of the discriminant function. When evaluating this measure, we can compare the observed misclassification rate to that expected by chance alone. If the probability of being in the positive group is the same as in the negative group (i.e., 0.5), 72.9% of original cases can be correctly classified by this discriminant function.

If we use a different prior probability, e.g., using the observed number in positive group and the observed number in the negative group, 78.4% of original cases can be correctly classified.

Since the discriminant scores will be used to plot a ROC curve, the above two cutting points are just convenient examples to demonstrate the power of the function, we can examine the detection rate and false alarm rate for all cutting points using the ROC curve.

In the above discriminant function, the ranges of ***r*** and ***z*** are the same, with minimum equal to zero and maximum equal to one, therefore the coefficients are comparable without standardization. It appears that the ratio of precision contributes more to the discriminant scores than does normalized dissimilarity.

When the positive group is considerably smaller than the negative group (i.e., 3,623 positive cases vs. 9,171 negative cases), a high correct classification rate can occur even when most of the positive group cases are mis-classified. The positive group is, however, of particular interest. Therefore, inspecting the ROC curve will tell us much more about the discrminant power of this function.

Using the above result, we plot a ROC curve (*Figure 10.1*, next page) by sorting the discriminant scores numerically in descending order. Every point on the curve represents a possible cutoff point to discriminate between positive group and negative group, associated with a detection rate and false alarm rate. The closer an ROC curve comes to the ideal point (100% detection, 0% false alarm rate) the better the performance of the predictor(s).

The ROC curve in *Figure 10.1* shows that, as the threshold for predicting the sign of effectiveness of data fusion is varied smoothly, the curve rises in a concave fashion such that the detection rate is always higher then false alarm rate. We see in *Figure 10.1* that the linear discriminant formed achieves a detection rate of 60%, at a false alarm rate of only about 14%. In other words, with this threshold setting it would label 60% of the cases showing improvement correctly, and would label only 14% of the cases with no improvement incorrectly.

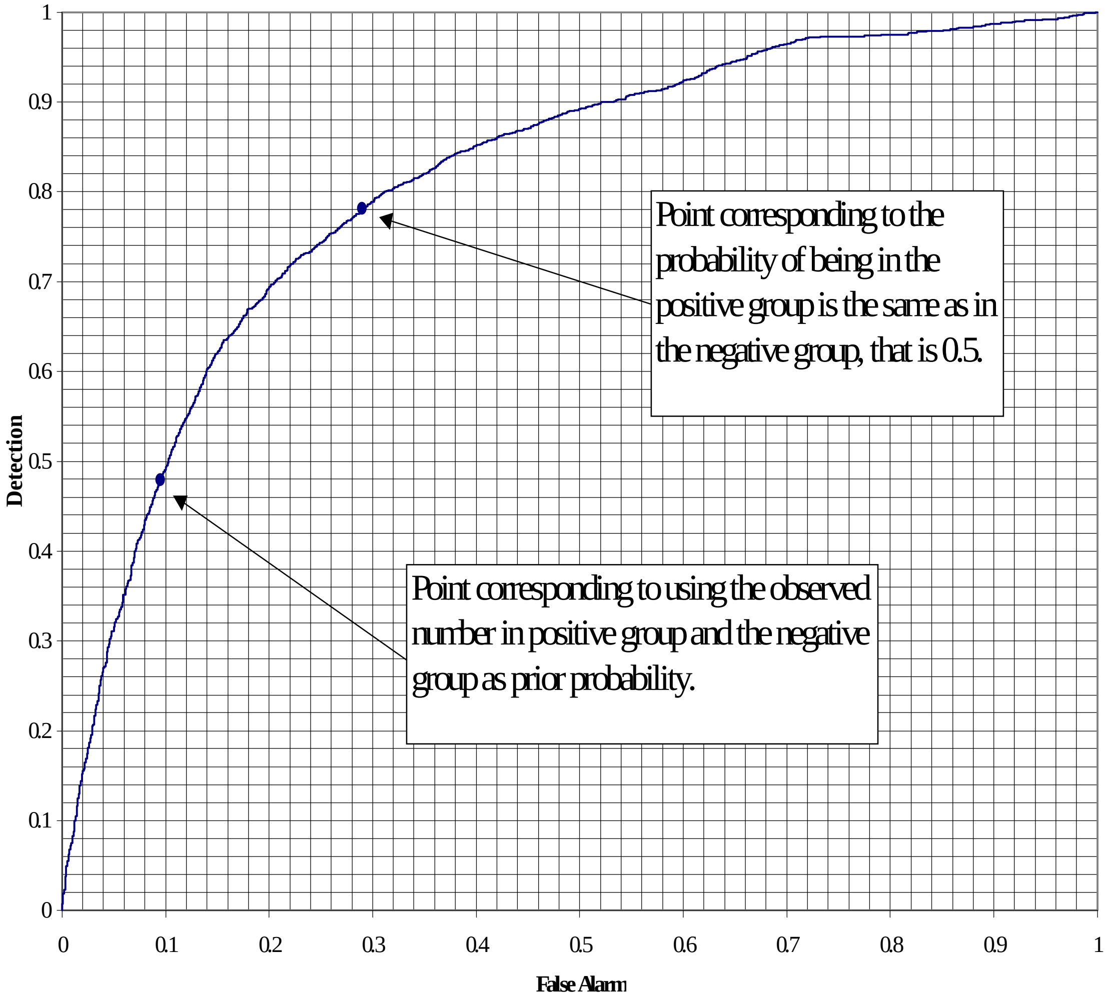

Figure 10.1: ROC curve of the tranining data set, using discriminant scores from the discriminant function $\boldsymbol{D} = 3.195\, \boldsymbol{z} + 4.643\, \boldsymbol{r} + 4.36$

When the detection rate rises to 80%, the false alarm rate is just about 31%, less than half of the detection rate. The linear discriminant function seems to be quite powerful, comparing to classification with chance alone, in predicting when the effectiveness of data fusion will be positive and when it will be negative.

10.2 Multiple Linear Regression Analysis

Two-group linear discriminant analysis is closely related to multiple linear regression analysis. If the binary grouping variable is considered the dependent variable and the predictor variables are the independent variables, the multiple regression coefficients estimated by OLS method (ordinary least squares method) are proportional to the discriminant function coefficients. Therefore it will not give us more information to run a multiple regression analysis on the same set of data and binary dependent variable. However, in our case here, the dependent variable is not necessary binary.

In the above analysis, we have only included the cases where the effectiveness of data fusion differs from zero. Cases are excluded when their effectiveness of data fusion are exactly equal to zero (i.e., the fused list has the same precision at 100th document as the better of the two IR scheme.) In addition, we did not use the magnitude of the effectiveness of data fusion. we only used the sign of effectiveness in our analysis. Much information was ignored. In this section, we use multiple linear regression to see if it will give us a better ROC curve.

In applying multiple linear regression analysis, first, we want to test whether the dependent variable, $\boldsymbol{E_o}$ (effectiveness of data fusion), is linearly related to the two independent variables, ***r*** and ***z*** (ratio of precision and normalized dissimilarity), and then calculate the strength of the linear relationship.

Second, we want to use the predicted values of the effectiveness of data fusion (by the regression equation) to plot an ROC curve to investigate the relationship between detection rate and false alarm rate and to compare this ROC curve with the ROC curve produced by linear discriminant analysis.

The application of multiple linear regression analysis here is primarily concerned with estimating and/or predicting the mean value of the effectiveness of data fusion on the basis of normalized dissimilarity and precision ratio. It postulates that the conditional mean of the effectiveness of data fusion is a linear function of normalized dissimilarity and precision ratio such that

$$\boxed{\boldsymbol{E_o} = B_0 + B_r \boldsymbol{r} + B_z \boldsymbol{z}}$$

Where the $\boldsymbol{B_0}$ is the intercept, $\boldsymbol{B_r}$ and $\boldsymbol{B_z}$ are the partial slopes of ***r*** and ***z*** in the regression equation.

To estimate the value of $\boldsymbol{B_0}$, $\boldsymbol{B_r}$ and $\boldsymbol{B_z}$ from the data set. We use the OLS method which results in a line that minimizes the sum of the square of the difference between the predicted value and the observed value of $\boldsymbol{E_o}$. In the 16,250 cases of the training data set, there are 495 cases where the precision at the 100th document for both IR schemes is zero, and $\boldsymbol{E_o}$ could not be defined. I eliminate those cases from analysis, using only the remaining 15,755 cases.

The regression equation estimated by the OLS method is:

$$\boxed{\boldsymbol{E_o = 0.408r + 0.164z - 0.418}}$$

According to this equation, when

(1) $\boldsymbol{z}$ is approximately one, or exactly equal to one ($\boldsymbol{z}$ = 0 when two output lists are identical), and

(2) $\boldsymbol{r}$ is approximately one, or exactly equal to one ($\boldsymbol{r}$ = 0 when the precision at the 100th document of the poorer IR scheme is zero),

then, in average, the precision at 100th document of the fused list produced by data fusion using simple symmetrical rule will be worse than the better scheme by approximately 0.418 – 0.408 = 10 % .

This means, when the output lists of the two IR schemes are highly similar, but have the same precision at the 100th document, on average, data fusion using simple symmetrical fusion rule will improve performance by about 15.4%.

The above can find empirical supports from, Ng and Kantor (1996), and Ng, ***et al*** (1997). They reported that in their data fusion experiments, using a simple symmetrical data fusion rule, the precision at the 100th document of the fused lists produced by combining capable IR schemes with incapable IR schemes were much worse that the precision at 100th document of the capable IR schemes.

According to the regression equation, with the same normalized dissimilarity, on average, an increase in 10% of precision ratio will produce about 4.08% increase in data fusion effectiveness (it is impossible to have increase in one unit, i.e., 100%, of precision ratio). And, with the same precision ratio, increase in 10% of normalized dissimilarity will produce 1.6% increase in effectiveness of data fusion (again, it is impossible to have increase in one unit, i.e., 100% , of normalized dissimilarity). The equation suggests that, on average, both precision and ratio and normalized dissimilarity are positively related with effectiveness of data fusion.

We can use the ratio between the estimated coefficients and their standard errors to test the relationship between the dependent variable and the independent variables:

> ***Null Hypothesis H_0 :*** There is no linear relationship between (1) the effectiveness of data fusion and (2) ratio of precision and normalized dissimilarity.
>
> ***Alternative Hypothesis H_1 :*** Effectiveness of data fusion is linearly related to normalized dissimilarity and ratio of precision.

There are 15,755 cases in the analysis, so the distribution of this ratio is the Student's ***t*** distribution with 15,755 – 2 = 15,753 degree of freedom (*Table 10.2*). From *Table 10.2,* we can reject the null hypothesis at the significance level of 0.000.

	Coefficients	**standard error**	**Student's *t***	**Significance**
Constant	-0.418	0.009	-46.039	0.000
Normalized dissimilarity	0.164	0.010	16.965	0.000
precision ratio	0.408	0.007	56.119	0.000

Table 10.2 Significance of the coefficients of the regression equation for the training data.

To test the goodness of fit of the linear model, a common measure is $\boldsymbol{R^2}$, or the coefficient of determination, which is the square of the correlation coefficient between the observed value of the effectiveness of data fusion and the predicted value of effectiveness of data fusion.

$\boldsymbol{R^2}$ can also be understood as how much better we can predict the effectiveness of data fusion from the normalized dissimilarity and precision ratio than we could predict the effectiveness of data fusion without the information about the normalized dissimilarity and precision ratio (i.e., using the mean as the predicted value). The $\boldsymbol{R^2}$ of the regression analysis is "0.204". That means, using the regression equation reduces the sum of squared errors of prediction by more than 20%.

Now, we sort the predicted effectiveness of data fusion in descending order.

At each point in the sorted list, we can calculate the detection rate and false alarm rate using that point as the cutoff for prediction. Then we use the detection rate and false alarm rate of all the cases to plot an ROC curve. We compare this ROC curve with the ROC curve of the dsicriminant function (*Figure 10.2*).

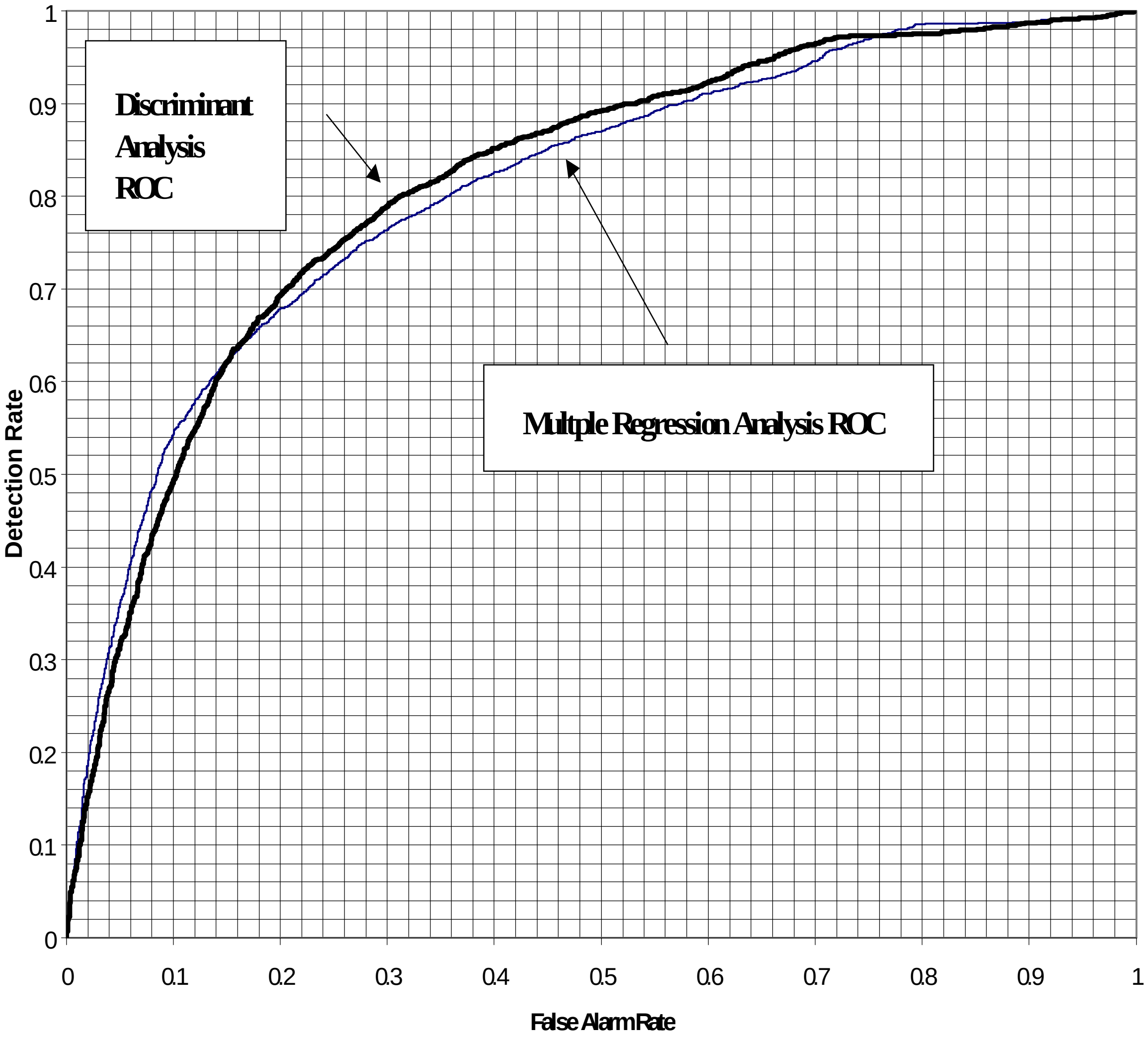

Figure 10.2 *ROC curves of the training data set, using multiple linear regression analysis and linear discriminant analysis.*

From *Figure 10.2* (previous page), we see that the ROC curves of multiple regression analysis and discriminant analysis inter-cross each other a few times. The most prominent crossing point is at the co-ordinate of 60% detection rate and 14% false alarm rate. Below that point, the ROC curve of multiple regression analysis is better than the ROC curve of discriminant analysis, above than point, the ROC of discriminant analysis is better than the ROC curve of multiple regression until the false alarm rate is about 75% and detection rate is about 97.5%. Generally the different between the two ROC curves is about 0% – 2% of detection rate, and it never exceeds 2% of detection rate. It seems that the predictive power of the two methods are similar to each other.

The multiple regression we have done here is a linear model. After using the multiple regression, we try to use polynomial regression to see if we can build a statistically significant second-order model of the form:

$$\boxed{\boldsymbol{E_o} = \beta_{\boldsymbol{o}} + \beta_1 \boldsymbol{r}^{2} + \beta_2 \boldsymbol{z}^{2} + \beta_3 \boldsymbol{rz} + \beta_{\boldsymbol{4}} \boldsymbol{r} + \beta_{\boldsymbol{5}} \boldsymbol{z}}$$

The result is, only the coefficients of normalized dissimilarity and ratio of precisions have a ***P***-value of 0.000, all the others parameters (square of normalized dissimilarity, square of ratio of precisions, and the product of normalized dissimilarity and ratio of precisions) have ***P***-value higher than 0.05 (*Table 10.3*):

Parameters	**Coefficients**	***P*-value**
z	.285	.000
z^2	-0.08.791	.066
r	.572	.000
r2	-0.05211	.099
rz	0.05991	.215

Table 10.3: Coefficients and P-value of the polynomial regresson for the training data st.

10.3 Logistic Regression Analysis

In the previous session, we applied multiple regression analysis and discriminant analysis to the training data set. It seems that even though multiple regression analysis has used more data (i.e., 15,755 cases) and more information about the dependent variable (i.e., not just the sign but also the magnitude of the effectiveness of data fusion), it does not offer a generally better ROC curve, but more or less the same operating characteristic as discriminant analysis. It looks reasonable to use discriminant function on the test data set (fused list generated by combining the output lists of all IR schemes participating in TREC 5 routing task, using the same simple symmetrical fusion rule: sum of normalized relevancy scores). However, there are some inherent limitations associated with discriminant analysis, which lead us to consider one more method: logistic regression analysis.

For linear discriminant analysis, two assumptions must be met for the prediction rule to be optimal. The first is the assumption of multivariate normality of independent variables. The second is the assumption of equal variance-covariance matrices in the two groups. The logistic regression model requires far fewer assumptions than discriminant analysis; and even when the assumptions required for discriminant analysis are satisfied, logistic regression still performs well (Hosmer and Lemeshow, 1989).

In the following, we will use logistic regression to investigate the discriminative power of the two independent variables. As in the linear discriminant analysis, we eliminate from the training set those cases with data fusion effectiveness equal to zero.

Let $P(E_o > 0)$ represents the probability that the simple symmetrical data fusion between the ranked output lists of two IR schemes (S_1 , S_2) will be better than the best of S_1 and S_2 (i.e., positive data fusion effectiveness). Let $Odds(E_o > 0)$ denotes the odds of positive data fusion effectiveness, that is:

$$\boxed{Odds(E_o > 0) = \frac{P\ (E_o > 0)}{1-\ P\ (E_o > 0)}}$$

Unlike $P(E_o > 0)$, $Odds(E_o > 0)$ has no fixed maximum value, but like the probability, it has a minimum value of zero. The logit of E, $\log_e(Odds(E_o > 0))$, becomes negative and increasingly large in absolute value as the odds decrease from 1 to 0, and becomes increasingly large in the positive direction as the odds increase from 1 to infinity. If we use the natural logarithm of $\boldsymbol{Odds\ (E_o > 0)}$ as dependent variable, the equation for the relationship between the dependent variable and the independent variables then becomes:

$$\boxed{\boldsymbol{log_e\ (Odd\ (E_o > 0) = B_0 + B_1\, r + B_2\, z}}$$

Converting $\boldsymbol{log_e\ (Odd\ (E_o > 0))}$ back to the odds by exponentiation results in the equation

$$\boxed{\boldsymbol{Odds\,(E_o > 0) = e^{B_0 + B_1 r + B_2 z}}}$$

Converting the odds back to the $\boldsymbol{P(E_o > 0)}$ by

$$\boxed{\boldsymbol{P(E_o > 0) = \frac{Odds(E_o > 0)}{1 + Odds(E_o > 0)}}}$$

Then we have a function of probability of positive data fusion. We can use the training data to estimate the coefficients of the above logistic regression equation, $\boldsymbol{B_o}$, $\boldsymbol{B_1}$, and $\boldsymbol{B_2}$.

In the previous multiple linear regression analysis, the parameters of the model were estimated by the OLS (ordinary least square) method. OLS method selects regression coefficients that result in the smallest sums of squared distance between the observed effectiveness of data fusion and the predicted effectiveness of data fusion. In logistic regression analysis, the parameters of the model are estimated using the MLE method (maximum-likelihood method) provided by the SPSS statistical package (Norusis 1994).

For the MLE method, the coefficients that make the observed effectiveness of data fusion most "likely" are selected (i.e., the highest probability of the observed results, given the parameter estimates.)

Just as the sum of squared errors is the criterion for selecting parameters in the multiple regression model, the ***log likelihood*** is the criterion for selecting parameters in the logistic regression model. When using the sign or direction (i.e., positive or negative) of effectiveness of data fusion as dependent variable, the log likelihood is equal to:

$$\boxed{N_{Eo=1} \log_e [P(E_o > 0)] + (N - N_{Eo>0}) \log_e [1 - P(E_o > 0)]}$$

Where N is total the number of cases, and $N_{Eo>0}$ is the number of positive cases (Menard 1995). In the above equation, initially $P(E_o > 0)$ is equal to the total number of positive cases divided by the total number of cases. After four iteration of estimation, the change of ***log likelihood*** decreased less than 0.01% and the estimation terminated. The logistic regression equation is as follow:

$$\boxed{\boldsymbol{Odds(E_o > 0) = e^{\,6.4485\; -\; 5.6068\, r\; -\; 3.6956\, z}}}$$

When multiplied by –2, the ***log likelihood*** has approximately a χ^2 (chi square) distribution. We can use the ***–2 log likelihood*** to test the null hypothesis that the coefficients for ***r*** and ***z*** are zero. The difference between the ***–2 log likelihood*** for the model with only a constant and the ***–2 log likelihood*** for the current model is 3368.315, with 2 degree of freedom, and the significance is 0.0000.

Another way to determine how well the logistic model performs is to see how well the model classifies the observed data. If we classify a case with $\boldsymbol{P(E_o > 0)}$ > 0.5 as positive case, the overall classification accuracy for the logistic regression is 78.57% , 0.17% better than the discriminant analysis classification accuracy when the latter uses the ratio of positive cases to the negative cases as prior probability.

With the logistic regression equation, we can estimate the probability of positive effectiveness of data fusion for every case and use the probability to plot a ROC curve (*Figure 10.3*, next page.)

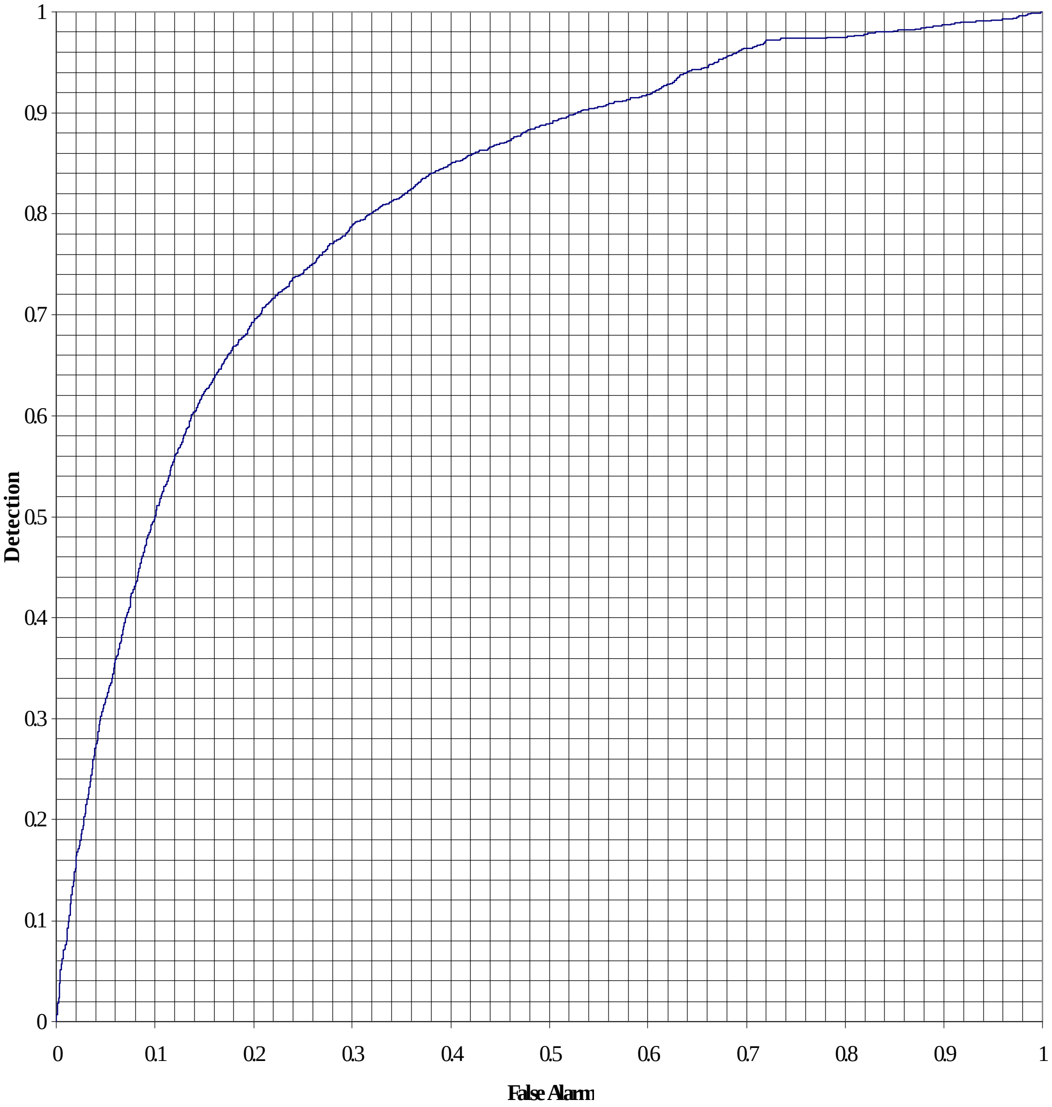

Figure 10.3 ROC curve (of the training data set) using probabilities of positive fusion estimated by logistic regression.

The ROC curve of probability of positive data fusion estimated by logistic regression is almost identical to the ROC curve in of linear discriminant analysis (*Figure 10.1*.) If we put the two ROC curves in one figure, they almost overlap with each other and are visually indistinguishable (*Figure 10.4*, next page).

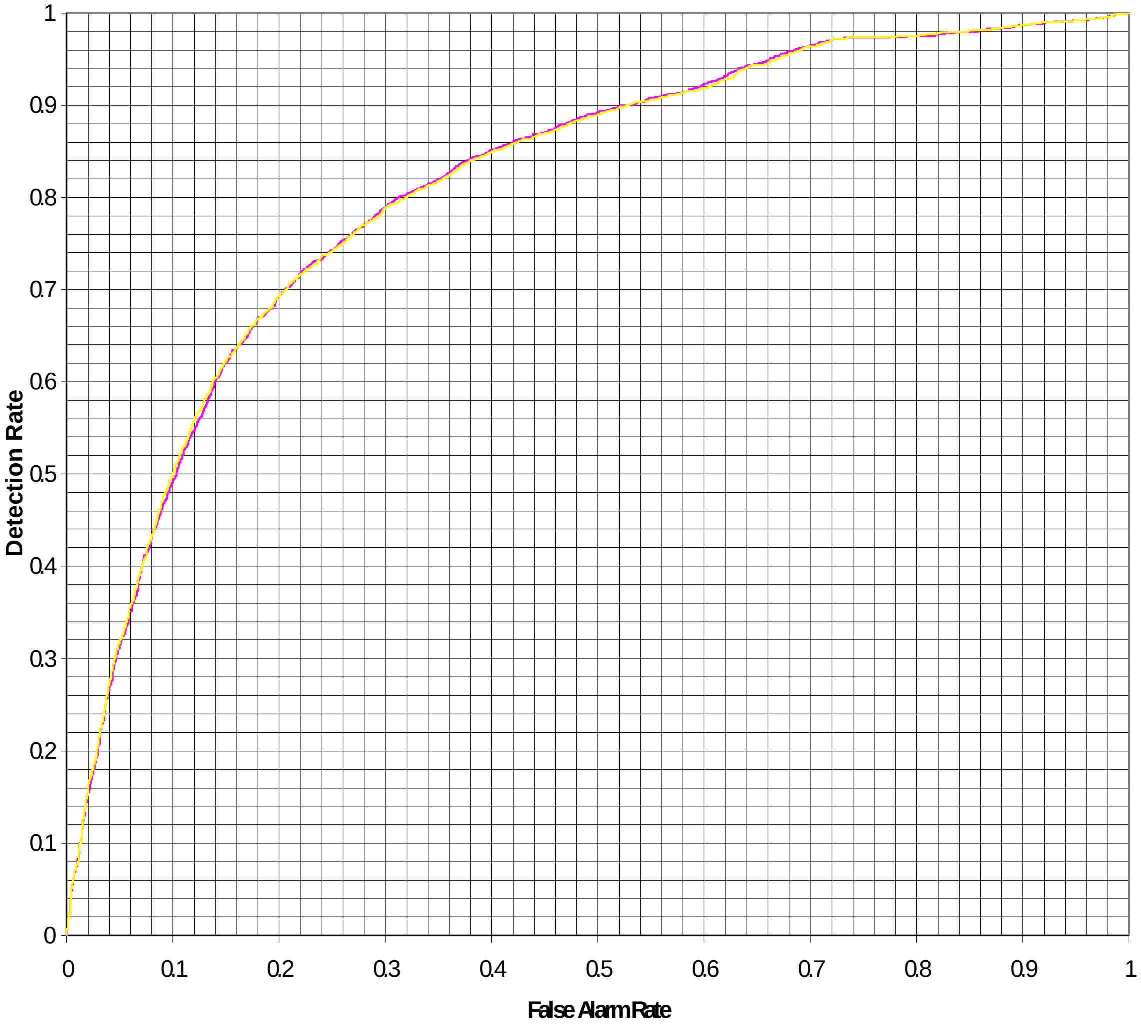

Figure 10.4 ROC curves (of the training data set) using probabilities of positive fusion estimated by logistic regression analysis (darker line) and discriminant scores of linear discriminant analysis (lighter line). The two curves almost totally overlap with each other and cannot be visually distinguishable.

Going back to the original data, we can see that, with the same false alarm rate, sometimes the detection rate of logistic regression is better than that of discriminant analysis, and sometimes the detection rate of discriminant analysis is better than that of logistic regression. For example, when the false alarm rate is 0.060, the detection rate of discriminant analysis is 0.351, while the detection rate of logistic regression is 0.352, 0.1% higher than that of discriminant analysis; when the false alarm is 0.653, the detection rate of logistic regression is 0.944, while the detection rate of discriminant analysis is 0.946, 0.2% higher than that of logistic regression.

10.4 Comparison of the Three Methods

When applied to our training data, it seems that the classification powers of the discriminant analysis and logistic regression are very close to each other. This is not surprising because the discriminant function is

$$\boxed{\textit{Discriminant Score} = 4.643\, r + 3.195\, z - 4.36}$$

so the ratio of the coefficients of normalized dissimilarity and precision ratio is 3.195/4.643 = 0.6881; while the logistic regression equation is:

$$\boxed{\frac{P(E_o > 0)}{1 - P(E_o > 0)} = e^{6.4485 - 5.6068\, r - 3.6956\, z}}$$

so the ratio of the coefficients of normalized dissimilarity and precision ratio –3.6956/–5.6068 = 0.6591. In other words, the two methods give more or less the same relative weight to the two independent variables in their classification computation. Comparing to the multiple regression analysis, which has the equation:

$$\boxed{E_o = 0.408r + 0.164z - 0.418}$$

and the ratio of the coefficients 0.164/0.408 = 0.4020, both logistic regression and discriminant analysis comparatively give more weight to normalized dissimilarity (*Table 11.2.*)

	Multiple regression	**Discriminiant analysis**	**Logistic regression**
Normalized dissimilarity	0.164	3.195	–3.6956
Ration of Precisions	0.408	4.643	–5.6068
Ratio of coefficients	0.4020	0.6888	0.6591

Table 11.2 relative weight of the coefficients of the predictive variables assigned by different methods

The ratio of the coefficients determines the slope of the line the vector of classification method in the two dimensional space of normalized dissimilarity and precision ratio (for detailed, see next chapter, Section 11.4), so it is not surprising that the ROC curves of discriminant analysis and logistic regression are very close to each other.

The purpose of comparing the ROC curves of the three statistical analyses techniques is to select a method which will give us higher detection. Since the detection rate of the ROC curve of multiple regression is lower than that of logistic regression and discriminant analysis after detection rate exceeds 60%, we decided not to apply the multiple regression equation on the testing data.

Now we will apply the discriminant function and the logistic regression equation estimated from the training data on the testing data.

Chapter 11

Training and Testing: The Predictive Power of the Two Parametric Analysis Methods and One Non-Parametric Analysis Method

11.1 Testing Data

In TREC 5, there were twenty six IR schemes participating in the routing task. Twenty three of them used all the collection documents available, the rest of them only used a subset of the collection (Voorhees & Harman, 1997). We use the output lists of these twenty three schemes as testing data.

There were fifty topics in TREC 5 routing task, therefore we have in all 23×50 = 1,150 output lists. Each output list contains 1,000 documents in ranked order of assigned relevancy scores. From these 1,150 lists, there are ½(23×22×50) = 16,250 pairs for data fusion. As it turns out, for topics 68, 125, 237, 240 and 243, there are no relevant documents in the testing collection. That means, for these five topic no matter how good or how bad a data fusion rule is, there will not be any decrease or increase in precision at the 100th document. We eliminated these five topics from my testing data, so finally we have ½(23×22×45) = 11,385 cases.

For each pair of information retrieval schemes ($\boldsymbol{S_1}$ **,** $\boldsymbol{S_2}$):

1. We divided the precision at the 100th document of the comparatively poorer scheme by the precision at the 100th document of the better scheme to get the ratio of precisions ($\boldsymbol{r}$). If the two schemes have the same precision at the 100th document, the precision ratio is one;
2. We compute the normalized dissimilarity ($\boldsymbol{z}$) between the two IR schemes using the ranked order of all the 1,000 documents of each list);
3. We rank the documents retrieved by the two IR schemes according to the sum of normalized relevancy scores and keep the top one thousand documents,
4. We compare the top 100 documents of the fused list with the list of elevant documents provided by TREC 5 to get the precision at the 100th document, of the fused list, i.e., $\boldsymbol{P(S_1\ f\ S_2)}$, where $\boldsymbol{f}$ represents the simple symmetric fusion rule.
5. We calculate the relative improvement of the data fusion against an oracle, $\boldsymbol{E_o\ (S_1\ f\ S_2)}$, by the formula:

$$\boxed{E_o\,(S_1\ f\ S_2) = \frac{P\,(\,S_1 f\ S_2) - max\{P(S_1), P(S_2)\,\}}{max\{P(S_1), P(S_2)\,\}}}$$

11.2 Predictive Power of Two Parametric Methods: Discriminant Function and Logistic Regression Equation

We apply the discriminant function estimated from the training data set to predict the discriminant scores of the cases in the testing data set, and then use the predicted discriminant scores to plot an ROC curve. We also apply the logistic regression equation estimated from the training data set to the testing data set to predict the probability that the effectiveness of data fusion is greater than zero. In this case, we use the predicted probability to plot an ROC curve.

Figure 11.1 shows the ROC curves using the two predicting methods. As in the training data set, the two ROC curves almost completely overlap with each other and are visually not distinguishable. That means that the predictive power of these two methods are practically the same.

From *Figure 11.1,* we can see that when the detection rate is below about 75%, the predictive power of either of these two ROC curves is much better than predicting without any information about normalized dissimilarity and precision ratio.

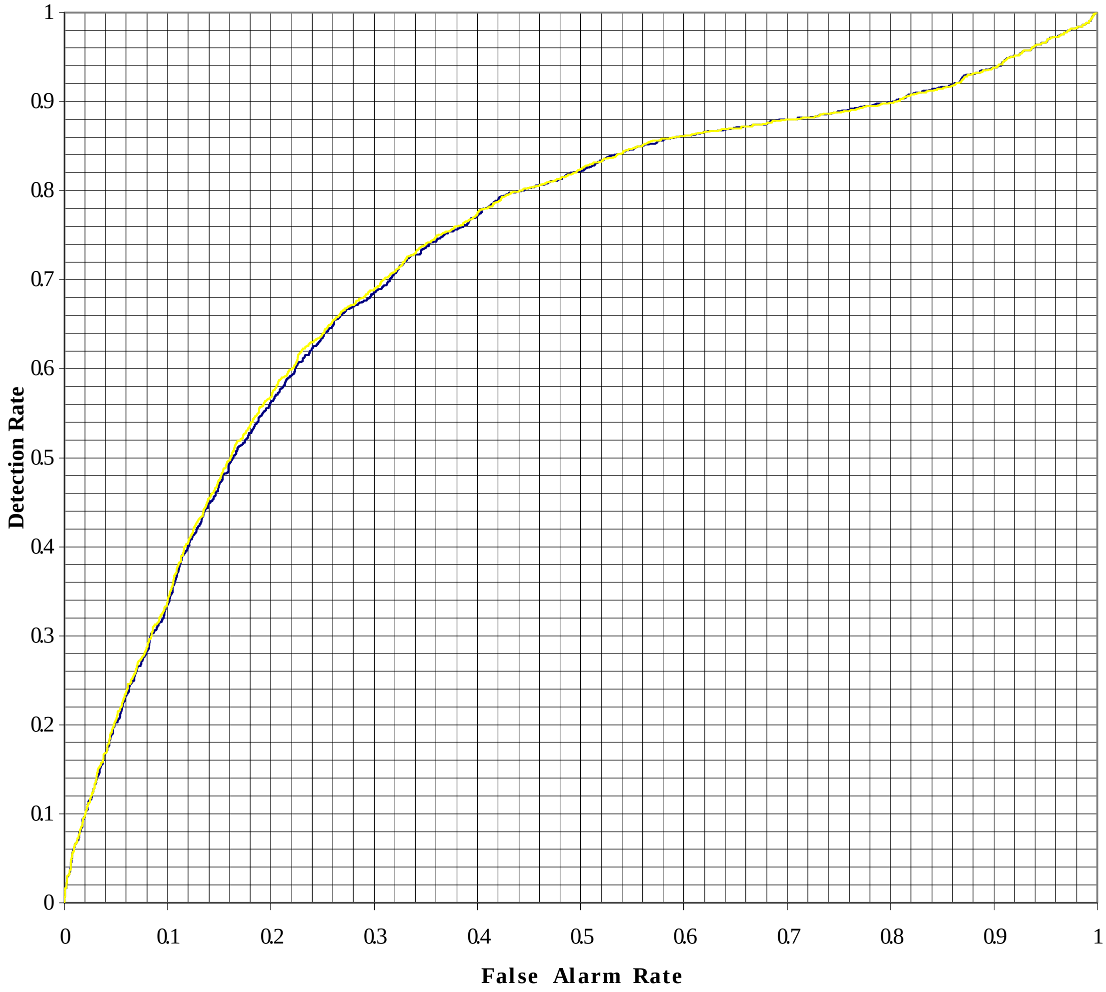

Figure 11.1 The two ROC curves (testing data set) of predicted discriminant scores by discrminant analysis (dark curve) and predicted probabilities by logistic regression (light curve). They almost completely overlap with each other.

When the detection rate is below about 75%, it is about one to two times higher than the false alarm rate. For example, when the detection rate is about 60%, the false alarm rate is about 22%, the detection rate is about 2.73 times of the false alarm rate; when the detection rate is about 70%, the false alarm rate is about 31%, the detection rate is about 2.26 times of the false alarm rate.

However, when the detection rate is higher than 75%, although the detection rate is not much higher than the false alarm rate as before, it is still better than chance alone. For example, when the detection rate is about 80%, the false alarm rate is about 47%; when the detection rate is about 90%, the false alarm rate is about 80%, still lower than detection rate.

We see that the two statistical analysis techniques have good predictive power in the low false alarm rate and moderate detection rate region, that is, from about 16% to 31% of false alarm rate and 50% to 70% detection rate.

11.3 Non-Parametric Training and Testing

The above two statistical techniques are parametric techniques. In other words, the model has a few parameters (the Betas). Also, the underlying statistical measures assumes that variables have multivariate normal distribution.

Besides parametric techniques, we can apply a non-parametric analysis which is totally empirical and highly nonlinear.

In Chapter 9, we grouped the positive cases and the negative cases of the training data into 100 square bins by defining 10 ranges of normalized dissimilarity and 10 ranges of ratio of precisions. Each bin contains the number of cases in a particular range of normalized dissimilarity and ratio of precisions.

If we divide the number of positive cases in a bin by the number of negative cases in the bin, we build the *Table 11.1* If we arrange the bins in rank order, with the cell have the highest rate of positive cases to negative case ratio ranked as “1”, we get *Table 11.2* (next page).

		Normalized Dissimilarity									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Ratio of Precisions	0.1	0.000	0.000	0.000	0.504	0.090	0.113	0.280	0.087	0.132	0.221
	0.2	0.000	0.000	0.315	0.000	0.000	0.107	0.031	0.238	0.188	0.793
	0.3	0.000	0.000	0.105	0.360	0.320	0.303	0.182	0.266	0.244	1.222
	0.4	0.000	0.000	0.063	0.291	0.233	0.187	0.343	0.372	0.744	1.453
	0.5	0.000	0.280	0.000	0.170	0.240	0.423	0.460	0.639	1.471	3.393
	0.6	0.000	0.373	0.111	0.123	0.236	0.760	0.667	1.735	0.802	1.512
	0.7	0.000	0.000	0.089	0.161	0.488	0.602	1.220	1.704	1.557	3.664
	0.8	0.000	0.450	0.455	0.686	1.169	1.812	2.262	2.465	2.927	2.240
	0.9	0.000	0.510	0.904	1.492	3.077	3.720	4.217	5.880	6.298	7.132
	1.0	0.734	2.771	4.644	6.184	7.892	8.277	8.860	11.024	14.092	16.348

Table 11.1: Table of comparative performance of data fusion in different ranges of normalized dissimilarity and ratio of precisions. Each cell contains the ratio of positive cases to negative cases in a particular range of normalized dissimilarity and precision ratio.

The highest rank in *Table 11.2* is 1 and the lowest rank is 83. If we use these data to plot an ROC curve, there are 83 points in the curve, point 1 represents the detection rate and false alarm rate we get if we use rank 1 as the cutoff point, point 2 represents the detection rate and false alarm rate we will get if we use rank 2 as cutoff point, etc.

		Normalized Dissimilarity									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Ratio of Precisions	0.1	83	83	83	44	78	74	58	80	72	66
	0.2	83	83	55	83	83	76	82	63	67	35
	0.3	83	83	77	52	54	56	69	60	61	30
	0.4	83	83	81	57	65	68	53	51	37	29
	0.5	83	59	83	70	62	49	46	41	28	15
	0.6	83	50	75	73	64	36	40	23	34	26
	0.7	83	83	79	71	45	42	31	24	25	14
	0.8	83	48	47	39	32	22	20	19	17	21
	0.9	83	43	33	27	16	13	12	10	8	7
	1.0	38	18	11	9	6	5	4	3	2	1

Table 11.2: Each cell contains the rank order of the comparative performance of a particular range of normalized dissimilarity and precision ratio. The cell with the rank one represents the range of normalized dissimilarity and precision ratio which will give the highest ratio of positive cases to negative cases.

The detection rate at point 1 will be the number of positive cases in the rank 1 bin divided by the total number of positive cases in the data set; the false alarm rate of point 1 will be the total number of negative cases in the rank 1 bin, divided by the total number of negative cases in the data set. The detection rate at point 2 is the number positive cases in the rank 1 bin and rank 2 bin combined, divided by the total number of positive cases in the data set; the false alarm rate of point 2 is the number of negative cases in the rank 1 bin and rank 2 bin combined, divided by the total number of negative cases in the data set. The detection rate of point ***n*** will be the number of positive cases in the rank 1 bin, rank 2 bin, …, rank ***n*** bin combined, divided by the total number of positive cases in the data set. The false alarm rate of point ***n*** is the total number of negative cases in the rank 1 bin, rank 2 bin, …, rank ***n*** bin combined, divided by the total number of cases in the data set.

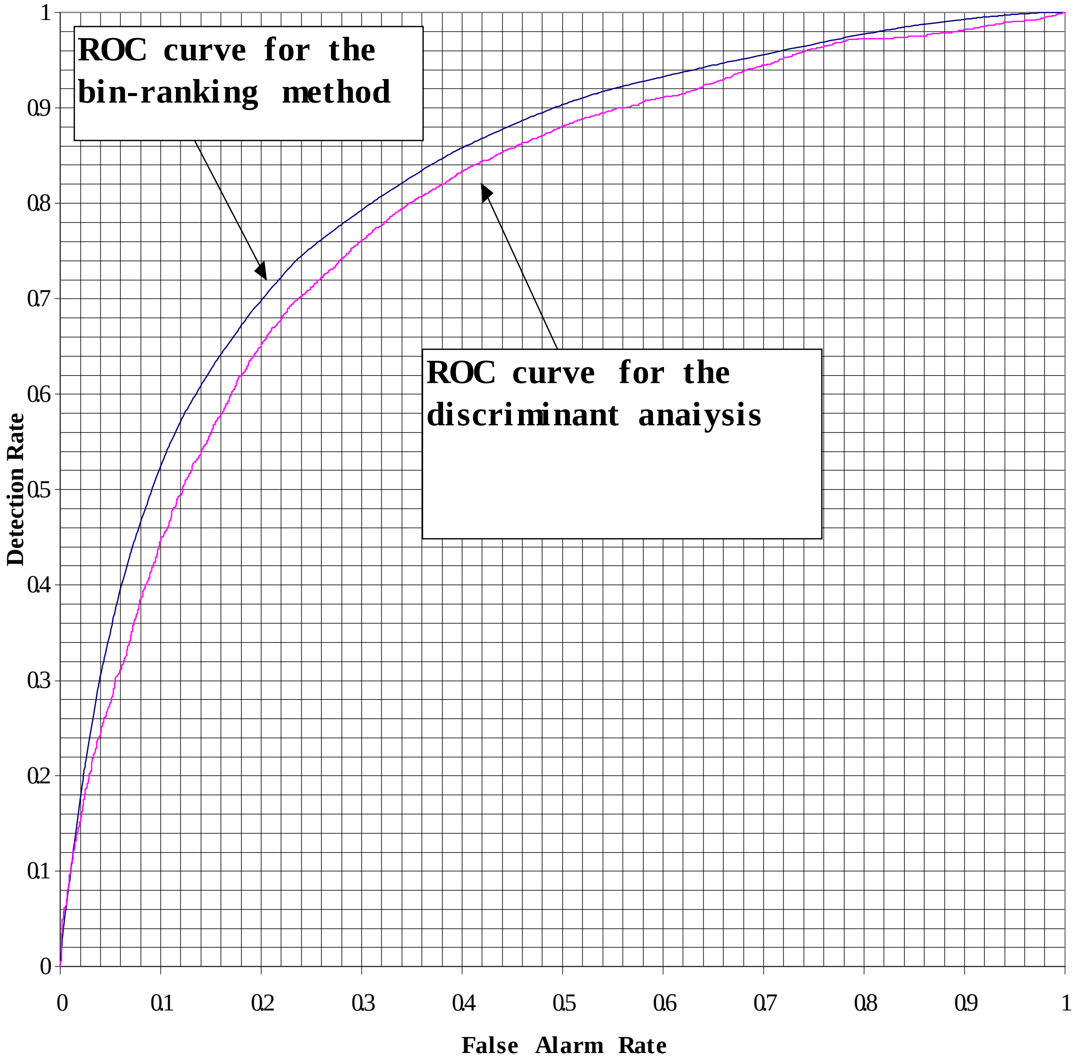

Figure 11.2: Comparison of performance between the non-parametric bin-ranking method and the parametric discriminant analysis for the training data set.

Figure 11.2 (previous page) is the ROC curve of this non-parametric bin-ranking method and the ROC curve of one of the two parametric methods, i.e., linear discrminant analysis.

From *Figure 11.2*, we can see that the ROC curve of this non-parametric method is more powerful than the ROC curves of the parametric method. With the same false alarm rate, the detection rate of the non-parametric method is always higher than that of the discriminant analysis.

For example, when the false alarm rate of the non-parametric method is about 30%, the detection rate is about 79%; while for ROC of the discriminant analysis is about 12%, the detection rate is about 57%, while for the discriminant analysis, when the false alarm rate is about 12%, the detection rate is about 50%.

Since the ROC curve of the bin ranking method is apparently better the parametric method, perhaps rank the bins of the testing data set according the rank order determined by the training set, we will get a better prediction.

We group the cases of the testing data set into the same 100 square bins. Next we divide the number of positive cases in a bin by the number of negative cases in the corresponding bin. We rank the cells according to the rank order of the training data, and then plot an ROC curve accordingly (*Figure 12.3*, next page).

From *Figure 11.3*, we see that the ROC curve of arranging the bins of the testing data set according to the order determined by the training data set is not always concave, as was the ROC curve of the training data set. For example, check the point when the detection rate is about 60% and the false alarm rate is about 21%. These turning points indicate that the bins corresponding at the consecutive points are out of rank order.

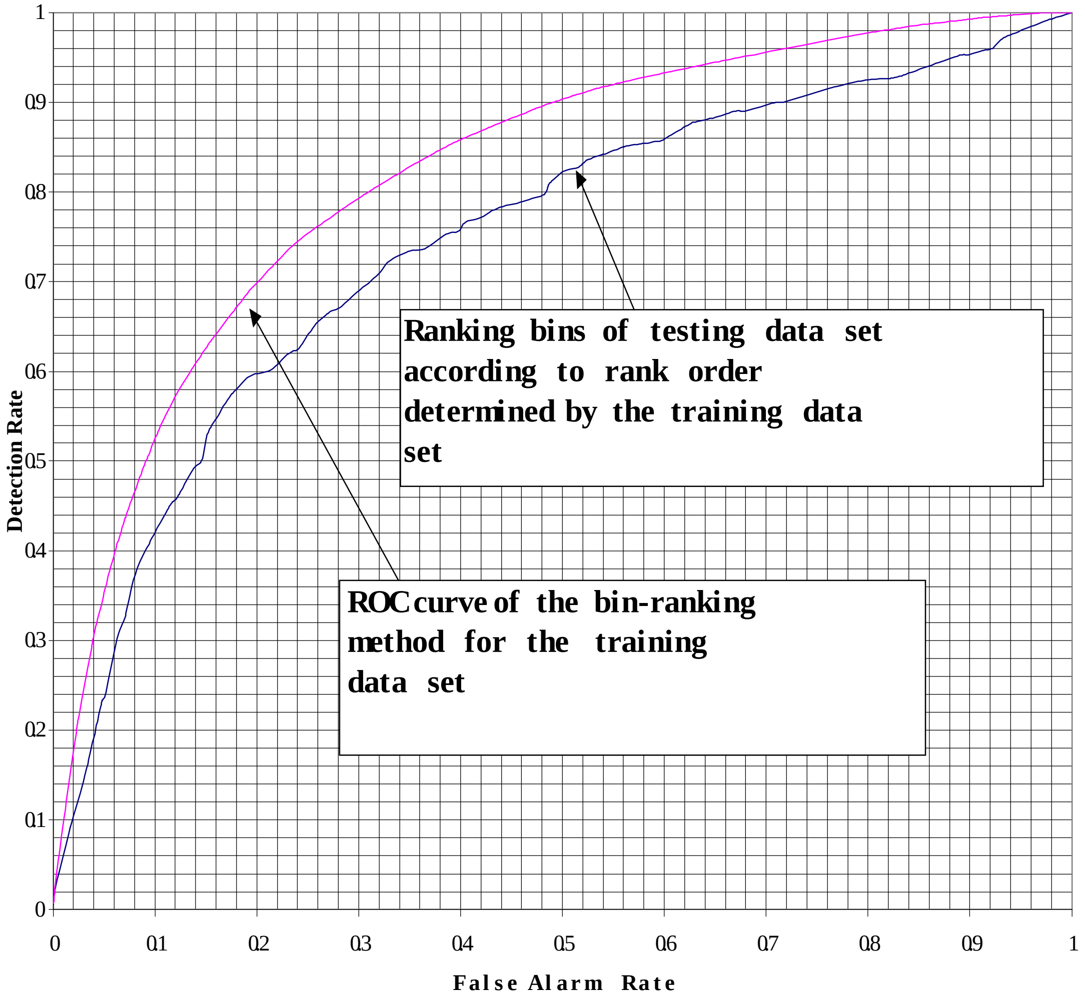

Figure 11.3: The ROC curves of the bin-ranking method: training and testing.

Let’s use the point when the detection rate is about 60% and the false alarm rate is about 21% as example. It corresponds to the bin with rank order 23, the exact detection rate is 60.17% and the exact false alarm rate is 21.40%. The ratio of positive cases to negative cases of this rank is 0.1477. The ratio of positive cases to negative cases of the rank 22 bin is 0.2518, while for the rank 24 bin, it is 0.5192 (*Table 11.3.*)

Predicted rank	**Ratio of positive cases to negative cases in the bin**
22 (A)	0.2518
23 (B)	0.1477
24 (C)	0.5192
Predicted rank order: A > B > C Ideal rank order: C > A >B	

Table 11.3: predicted rank order vs. ideal rank order for three bins in the testing data set.

From *Figure 11.2* and *Figure 11.3*, it seems that although the bin ranking method is more powerful than the two parametric methods in classifying the training cases into positive and negative cases, its predictive power drops a lot when applied to the testing data set. Of course it is unlikely that applying parameters estimated from training data to testing data will give the same performance. However, the difference between the training stage and testing stage seems to be quite large.

For example, from *Figure 11.3*, at point where the false alarm rate is about 20%, the detection rate of the training data set is about 70% but the detecting rate of the testing data set is about 60% , decreases 10%; when the false alarm rate is about 40%, the detection rate of the training data set is about 86%, the detection rate of the training data set is about 76%, also decrease by 10%.

The ROC curves for discriminant analysis do not change so much from the training data set to the testing data set (*Figure 11.4.)*

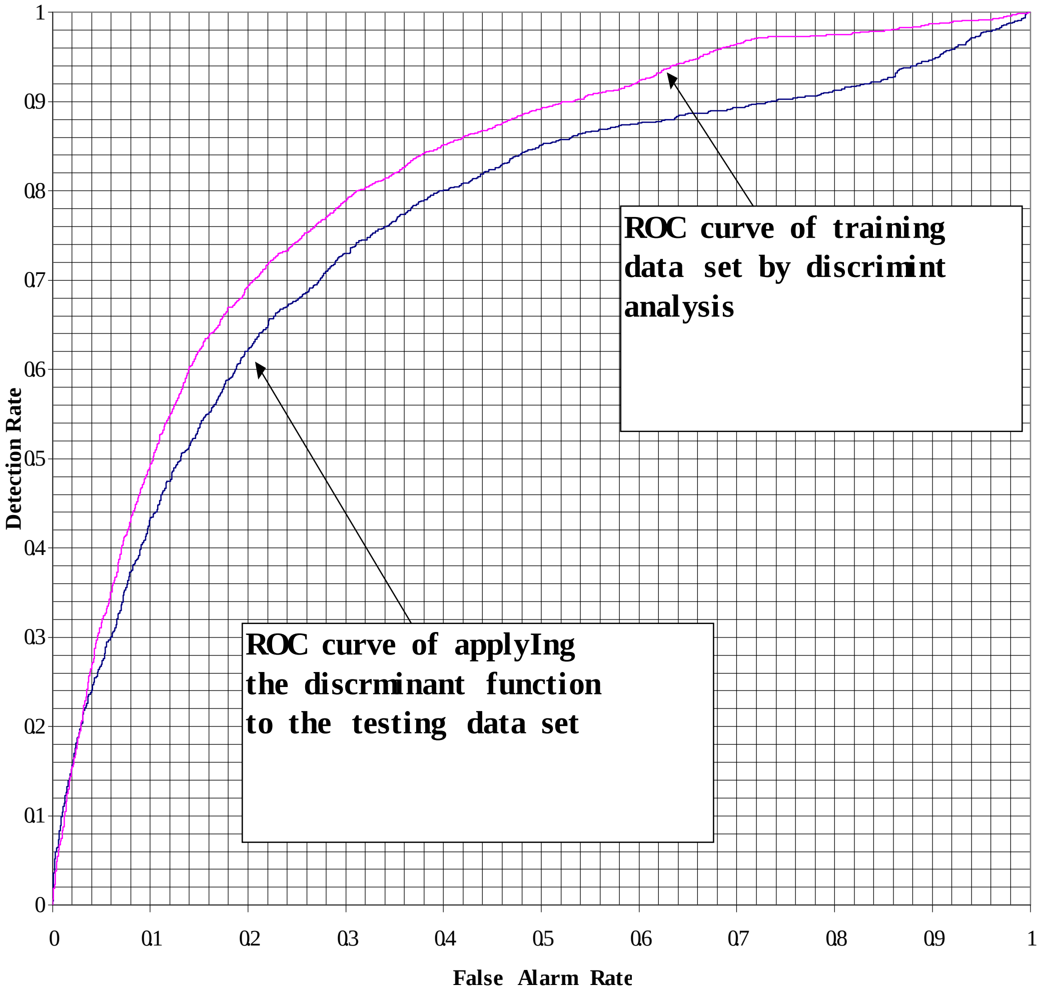

Figure 11.4: *The performance of discriminant analysis on the training data set (lighter curve) and the performance of the same discriminant function applied to the testing data set (darker curve) .*

From *Figure 11.4*, we see that generally the vertical distance between the two curves is not as great as for the non-parametric method (*Figure 12.3*). The biggest difference is in the range of false alarm rate about 68% to 78%. The ROC curve for the training data suddenly rises at this point while the ROC curve for the testing data suddenly flattens, indicating that it may be a region of over-training. In other words, because of the pattern of the distribution of positive cases and negative cases in the plane of ratio of precisions and normalized dissimilarity, the detection rate suddenly has a more rapid rise than the false alarm rate in this region for this data set, however this is only a specific characteristic of the training data, not a general characteristic for all data set.

The rank order of the same region in another data set probably will be different from this training data set. Therefore, applying the trained rank order (which has a sudden improvement) of that region to other data set probably will not get the same rapid rise in detection rate, but a rapid rise in false alarm rate because the actual rank other is different. That is why we say this may be a region of over-training. Detailed explanation follows.

As we have mentioned in Chapter 7 Section 7.4, when we move the line corresponding to the constant value of the discriminant function upward and perpendicular to itself across the plane of normalized dissimilarity and ratio of precision, we will have more and more cases under the line. The increase in detection rate and false alarm rate on the ROC curve corresponds to moving the line in such a direction and labeling all the cases under the line as positive cases. If we label all the cases under the line as positive cases, when we move the line upward, the detection rate will become higher and higher, as will the false alarm rate. Generally, for each strip of the plane added, the ratio between the increase in detection rate to the increase in false alarm rate will become smaller and smaller making the ROC curve concave in shape. That is because, in the ideal case, with an optimal discriminant function, the ratio of the number of positive cases to the number of the negative cases in each added strip will be smaller and smaller.

However, in our training data set, after the detection rate reaches 92% (i.e., 92% of the positive cases are under the line, let's call this region " ***x*** "), the ratio of positive cases to negative cases in the next region is higher than the region just below it, and this ratio keeps more or less constant until the detection rate is equal to about 97% (making this portion of the ROC curve to have a larger slope than the previous portion), then the ratio returns back to becoming smaller and smaller until the detection rate is equal to about 98% (making this portion of the ROC curve to be horizontal line), then the ratio keeps constant (making this portion of the ROC curve to be a diagonal straight line).

In other words, because the discriminant function is optimized over all the regions in the training data set, although in general the ratio of positive cases to negative cases is descending from one region to the next region, there may exist some regions which have a higher ratio than their preceding regions .

Applying the same discriminant function to the testing set, region “ ***x*** ” no longer has a higher ratio of the number of positive cases to the number of negative cases than the region below it. The ROC curve flattens instead of rising up. Now it is the region above “ ***x*** ” that has a higher ratio compared to region “***x***”, making this portion of the ROC curve to have a larger slope. n other words, the ROC curve of the testing data set flattens when the ROC curve of the training data set rises up; and the former rises up when the latter flattens. Therefore, putting the ROC curves for the two data set together, we see a larger gap near region “***x***”.

The direction of the moving of the discriminant line is determined by the slope of the equation of the discriminant function, in other words, it is determined by the ratio of the coefficients of the parameters. This explains why, for the training data set, the ROC curves for the discriminant analysis and logistic regression are very close to each other while they are quite different from the ROC curve for the multiple regression (see Chapter 10 Section 10.4).

11.4 Optimal ROC Curve of the Testing Data Set

We can investigate why there is a comparatively large drop of detection rate when using the non-parametric method, by comparing the ROC curves of the testing data set with the corresponding optimal ROC curve. From *Figure 11.5 (*next page*)*, we can see the predicted ROC curve based on arranging the 100 bins of the testing data set according to the rank order of the bins in the training data set is very close to its corresponding optimal ROC curve which is based on the actual rank order of the bins of the testing data set.

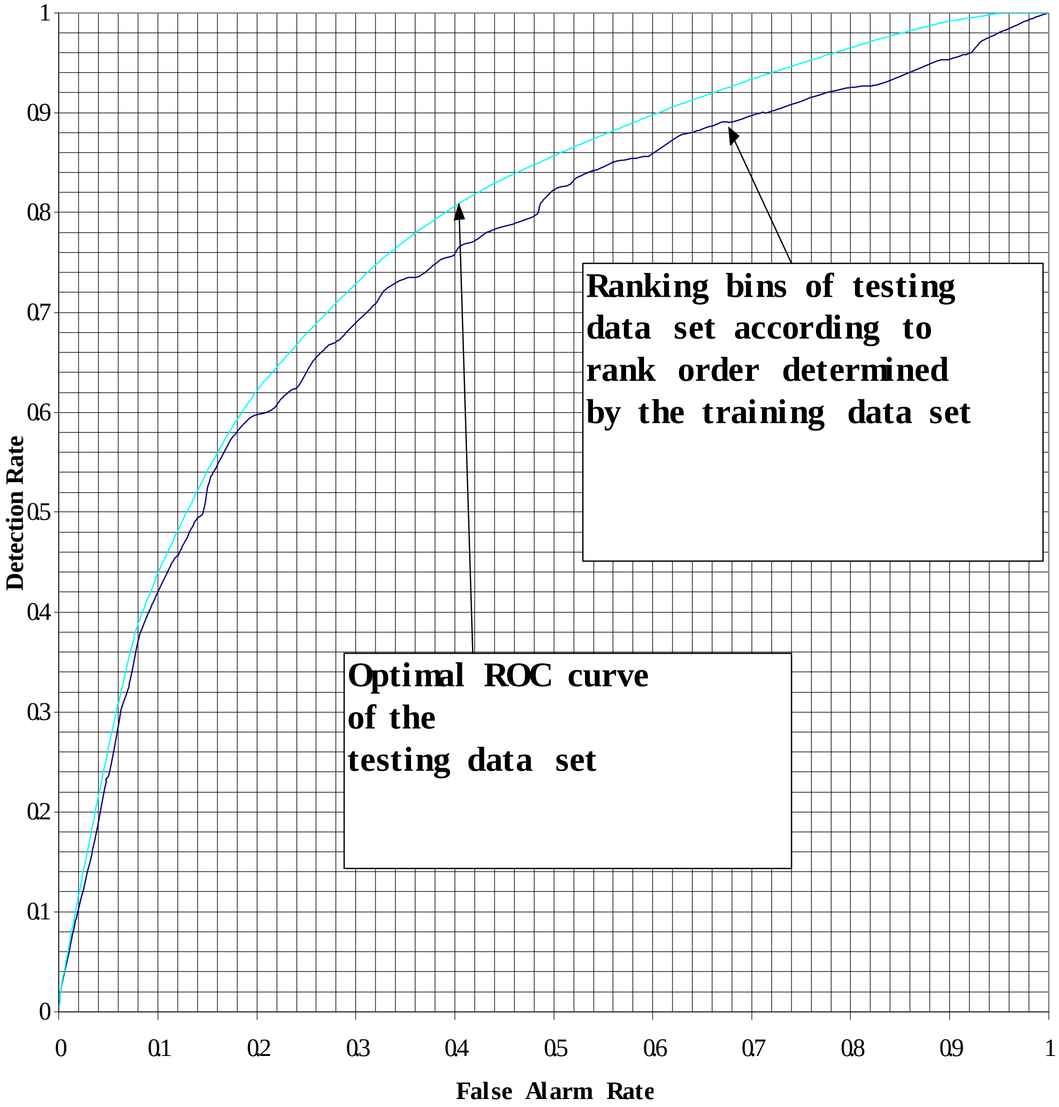

Figure 11.5 *Comparing the ROC curves of the training data set to its corresponding optimal ROC curve .*

In other words, the predicted ROC curve of the non-parametric method is very close to its corresponding optimal curve. The drop in detection rate from training data set to testing data set is only because the corresponding optimal curve for the testing data set is lower than that of the training data set (*Figure 11.6*).

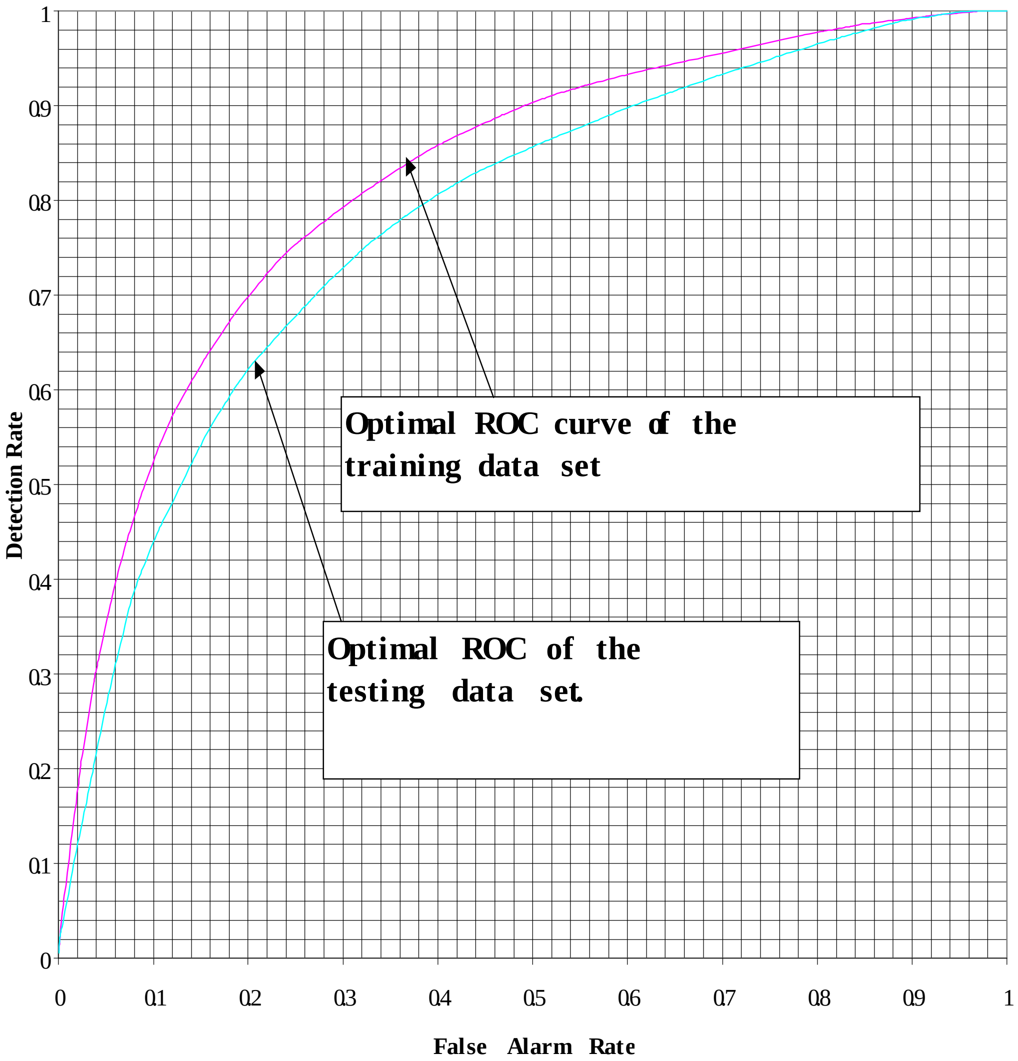

Figure 11.6: *Comparison of the ROC curves for testing and training data set arranging bins according to their actual rank in the data set.*

11.5 The Performance of Non-Parametric Method.

Although compared to the parametric method, e.g., discriminant analysis, the predictive power of the non-parametric bins ranking methods seems to drop more from training data set to testing data set, it may still have a better performance.

Since the ROC curves of the discriminant analysis and logistic regression almost completely overlap each other (see *figure 11.1*), the discussion here also applies to the logistic regression analysis.

Figure 11.7 (next page) is the ROC curves of the two predictive methods: parametric discriminant analysis and the non-parametric bin ranking method. When the false rate is below about 60%, the two ROC curves are very close to each other, with less than 1% difference in detection rate, indicating that the predictive power of the parametric method and the non-parametric method are more or less the same until false alarm rate is about 60%. When the false alarm rate is above 60%, the ROC curve of the non-parametric method is always better than the parametric method by about 2%.

In the above we have compared the performance of different predictive methods using ROC curves. In the following, we are going to discuss how to rationally predict the sign of effectiveness of data fusion rational based on the ROC curve.

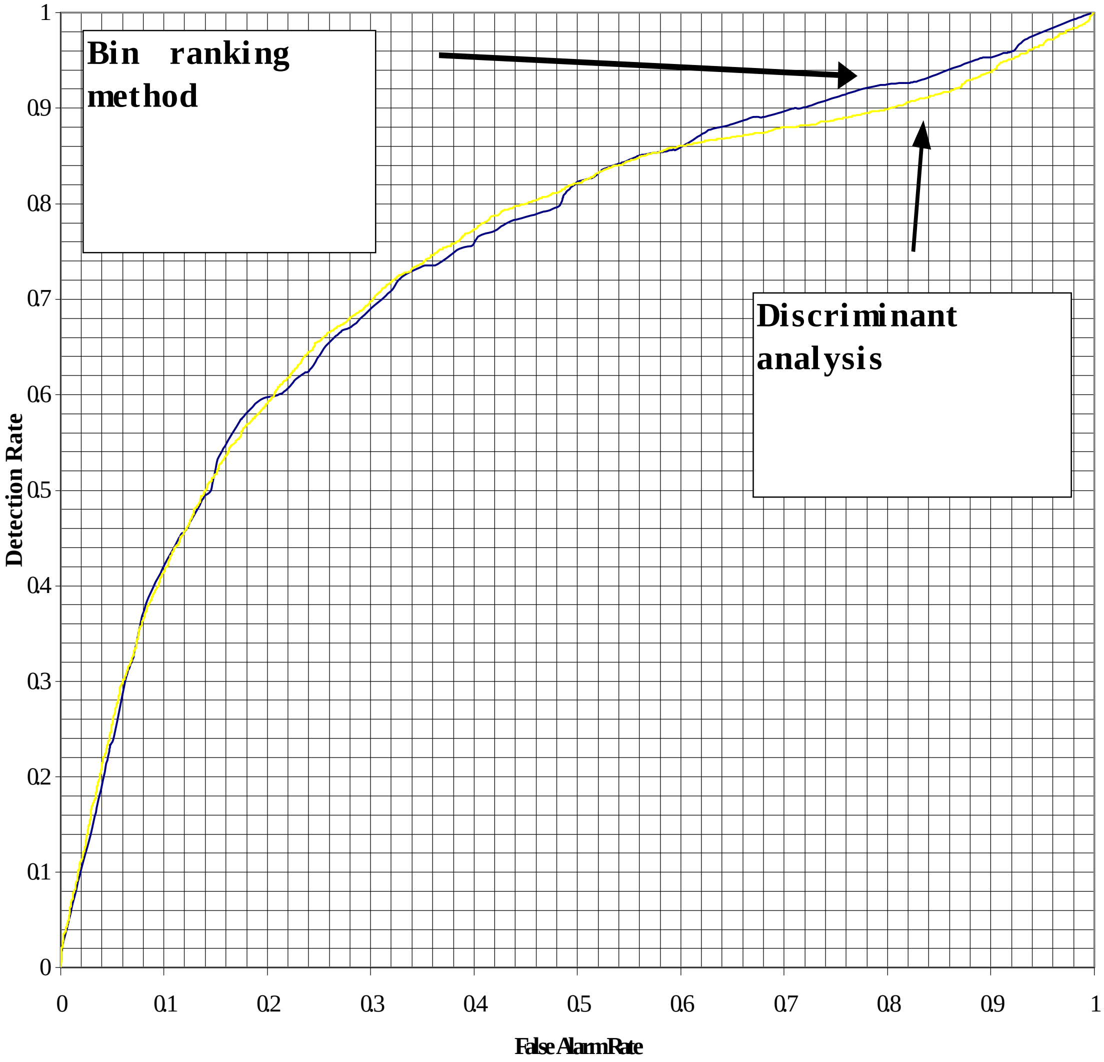

Figure 11.7: The prediction performance of the two methods in the testing data set: parametric linear discriminant analysis and non-parametric bin ranking method.

11.6 ROC and Effectiveness Prediction

We represent the problem of detecting positive effectiveness of data fusion in terms of signal detection. In particular, we use the ROC curve of signal detection theory. In the following, we will use the ROC of discriminant analysis as example, to discuss how to use the ROC curves to predict effectiveness of data fusion in IR.

Although we will use only discriminant analysis as example, the discussion also can also be applied to other predictive methods which will assign some kind of descriminant scores to the cases, e.g., probability based on logistic regression analysis, or inverse rank order based on the bin-ranking method.

We have two events, i.e.,, $e_{positive}$ and $e_{negative}$. The former means the effectiveness of data fusion is greater than zero while the latter means the effectiveness of data fusion is less than zero. We also have two responses, i.e.,, $r_{positive}$ and $r_{negative}$: predict as positive and predict as negative respectively. There are four event-response combinations: $(e_{positive}, r_{positive})$, $(e_{positive}, r_{negative})$, $(e_{negative}, r_{positive})$, $(e_{negative}, r_{negative})$, the conditional probabilities of these events satisfy the constrains:

$$Pr(e_{negative}) + Pr(e_{positive}) = 1$$

$$Pr(r_{positive} \mid e_{positive}) + Pr(r_{negative} \mid e_{positive}) = 1$$

$$Pr(r_{positive} \mid e_{negative}) + Pr(r_{negative} \mid e_{negative}) = 1$$

We can represent the above equations in *Table 11.4.*

	$r_{positive}$	$r_{negative}$	Sum of Pr
$e_{positive}$	$Pr(r_{positive} \mid e_{positive})$	$Pr(r_{negative} \mid e_{positive})$	*1*
$e_{negative}$	$Pr(r_{positive} \mid e_{negative})$	$Pr(r_{negative} \mid e_{negative})$	*1*

Table 11.4: The probabilities of the four event-response combinations.

In addition, we can define a ratio, ***L(d)***, where ***d*** is the discriminant score of a case calculated by the discriminant function estimated from the training data set..

$$\boxed{L(d) = \frac{Pr(d \mid e_{positive})}{Pr(d \mid e_{negative})}}$$

L(d) summarizes the changing ratio of the corresponding ordinates of the conditional probability distributions of the discriminant scores for positive cases and negative cases.

Our purpose is to predict the sign of the effectiveness of data fusion based on the discriminant score. The rational goals of decision theory is to maximize expected value. To achieve this goal, different values (and cost, i.e., negative value) are assigned to different outcomes.

The payoff matrix is shown in *Table 11.5*, where v_1 and v_4 are positive, v_2 and v_3 are negative:

	$\mathbf{r}_{\mathbf{positive}}$	$\mathbf{r}_{\mathbf{negative}}$
$\mathbf{e}_{\mathbf{positive}}$	v_1	V_3
$\mathbf{e}_{\mathbf{negative}}$	v_2	V_4

Table 11.5: Payoff Matrix of Events and Responses

Thus, the expected value function, ***E(V)***, of a decision rule with the conditional probabilities shown in *Table 11.4* becomes:

$$
\begin{aligned}
E(V) = \quad & Pr(r_{positive} \mid e_{positive})\, Pr(e_{positive}) \times v_1 + Pr(r_{positive} \mid e_{negative}) \times v_2 \\
& + Pr(r_{negative} \mid e_{positive})\, Pr(e_{positive}) \times v_3 \\
& + Pr(r_{negative} \mid e_{negative})\, Pr(e_{negative}) \times v_4
\end{aligned}
$$

Given discriminant score ***d***, we can compute the expected value ***E(V)*** of each assignment of $r_{positive}$ and $r_{negative}$:

$$
\begin{aligned}
E(V \mid d, r_{positive}) &= Pr(e_{positive} \mid d) \times v_1 + Pr(e_{negative} \mid d) \times v_2 \\
E(V \mid d, r_{negative}) &= Pr(e_{negative} \mid d) \times v_4 + Pr(e_{positive} \mid d) \times v_3
\end{aligned}
$$

To maximize expected value, when we observe ***d***, we should predict the effectiveness of data fusion as positive if and only if

$$\boxed{E(V \mid d, r_{positive}) > E(V \mid d, r_{negative})}$$

that is, when

$$\boxed{\begin{array}{l} Pr(e_{positive} \mid d) \times v_1 + Pr(e_{negative} \mid d) \times v_2 > \\ \qquad\qquad Pr(e_{negative} \mid d) \times v_4 + Pr(e_{positive} \mid d) \times v_3 \end{array}}$$

If $Pr(e_{negative} \mid d)$ is not equal to 0 and v_1 is not equal to v_3, then the above inequality can be re-arranged as:

$$\boxed{\begin{array}{ll} & Pr(e_{positive} \mid d) \times (v_1 - v_3) > Pr(e_{negative} \mid d) \times (v_4 - v_2) \\ \Rightarrow & \dfrac{Pr(e_{positive} \mid d)}{Pr(e_{negatives} \mid d)} > \dfrac{v_4 - v_2}{v_1 - v_3} \end{array}}$$

Using Bayes' Rule, we substitute $\dfrac{Pr(d \mid e_{positive})Pr(e_{positive})}{Pr(d)}$ for $Pr(e_{positive} \mid d)$ and $\dfrac{Pr(d \mid e_{negative})Pr(e_{negative})}{Pr(d)}$ for $Pr(e_{negative} \mid d)$.

This yields:

$$\boxed{\begin{aligned} &\frac{\boldsymbol{Pr(d \mid e_{positive})}}{\boldsymbol{Pr(d \mid e_{negative})}} > \frac{\boldsymbol{Pr(e_{negative})}}{\boldsymbol{Pr(e_{positive})}} \times \frac{\boldsymbol{v_4 - v_2}}{\boldsymbol{v_1 - v_3}} \\ \Rightarrow \quad &\boldsymbol{L(d)} > \frac{\boldsymbol{Pr(e_{negative})}}{\boldsymbol{Pf(e_{positive})}} \times \frac{\boldsymbol{v_4 - v_2}}{\boldsymbol{v_1 - v_3}} \end{aligned}}$$

where ***L(d)*** is the ratio we defined above.

The above decision rule divides the ***d***-axis into two regions which may consist of several disconnected parts.. In one region we should predict the effectiveness of data fusion as positive, and in the other region we should predict the effectiveness of data fusion as negative, with the boundary point(s) point ***c*** satisfying the equation:

$$\boxed{\boldsymbol{L(c)} = \frac{\boldsymbol{Pr(e_{negative})}}{\boldsymbol{Pr(e_{positive})}} \times \frac{\boldsymbol{v_4 - v_2}}{\boldsymbol{v_1 - v_3}}}$$

In parametric analysis we look for the best connected regions approximately this rule. In the non-parametric analysis, we permit disconnected regions. The disconnection occurs precisly when bin number 14 is added (see *Talbe 11.2*) and it is not adjacent to the bins already added to the positive region.

Chapter 12:

Further Discussions about The Experiment

12.1 The Power of Simple Symmetrical Data Fusion

In our experiments, we used a very simple rule to combine the ranked output lists of IR schemes, ranking the sum of normalized relevancy scores. We did not apply any fine tuning mechanism or machine learning method to find the best rule for each pair of schemes.

Although our fusion rule is very simple, it is also a powerful rule. As we have shown in Chapter 7, this naïve fusion approach performs better than the uninformed decision makers (who picks either scheme with equal probability) in 11,785 cases out of 14,395 cases (over 80%), and even beats an oracle (who knows which scheme will perform better and always picks the better scheme) in some 3,623 cases out of 12,794 cases (more than 28%)

In fact, for the fifty topics in the training data set, if we compare the precision at the 100th document of all single schemes and all fused schemes, we will see how powerful the simple symmetrical data fusion rule is. In thirty-five topics, the highest precision belongs to a fused scheme. For fourteen topics, they are tied. In only one topic (topic 45) does the highest precision belongs to single scheme, and in this topic, the precision of the highest scheme is 0.78 (scheme INQ203), higher than the fused scheme by just 0.01 (scheme INQ203 fused with scheme ORAdL1, precision 0.77), about 1.3% (0.01/0.77).

Moreover, the power of the fusion reported in the above paragraph is a conservative estimate and the actual performance is likely to be higher. The reason is as follow. Our raw data are from TREC 4 and TREC 5. In TREC 4 and TREC 5, only the top 100 documents of all submitted lists have been judged as relevant or non-relevant. If we use the sum of normalized scores to re-rank the 2,000 documents retrieved by two IR schemes, it is possible that some of the documents within the top 100 documents of the fused list are non-judged documents. In other words, the actual precision at the 100th document of the fused list may be higher than what we have estimated in our experiments using the relevant documents list provided by NIST.

The above factor will affect the performance of our predictive methods in a complex way. If we did not underestimate the precision of the fused lists, the number of positive cases will be larger than we think and the number of negative cases will be smaller than we think. The effect on the ROC curve is, however, unknown because the same factor will affect the effectiveness of data fusion on both the training data set and the testing data set. The estimation of the parameters of our predictive methods may change. The rank of the 100 bins may be different. Whether or not we will get better ROC curves for the training data set and testing data set depends on:

1. which regions will gain relatively more positive cases, and
2. which regions will lose relatively more negative cases.

We can use the bin ranking method as an example. Let there be only three bins. Originally, when we use the underestimated precision of the fused list, the three bins are:

	Number of positive documents	Number of negative documents	Ratio of positive to negative documents	Accumulated Detection rate	Accumulated False alarm rate
Bin 1	5	1	5.00	0.5	0.1
Bin 2	3	4	0.75	0.8	0.5
Bin 3	2	5	0.40	1.0	1.0

Table 12.1 Original three bins for a fused list with underestimated precision.

Assuming now we have the true precision of the fused list, the three bins become:

	Number of positive documents	Number of negative documents	Ratio of positive to negative documents	Accumulated Detection rate	Accumulated False alarm rate
Bin 1	6	0	Infinity	0.4	0
Bin 2	6	1	6.00	0.8	0.2
Bin 3	3	4	0.75	1.0	1.0

Table 12.2 The number of positive documents and negative documents in each bin changes and give a better ROC curve.

Which will give us a better ROC curve compared to the ROC curve for the original three bins. However, the three bins becomes:

	Number of positive documents	Number of negative documents	Ratio of positive to negative documents	Accumulated Detection rate	Accumulated False alarm rate
Bin 1	6	1	6.0	0.40	0.20
Bin 2	5	1	5.0	0.73	0.40
Bin 3	4	3	1.33	1.00	1.00

Table 12.3 The number of positive documents and negative documents in each bin changes and give a worse ROC curve.

Which will give us a worse ROC curve compared to the ROC curve for the original three bins.

Even if we have better ROC curves for the training data set as well as for the testing data set, the predictive power of the ROC curve for the training data set is not necessary better than before. The predictive power may be the worse, or the same. Whether an ROC curve of a predictive method will have the same predictive power as before, depends only on whether the gain of positive cases and loss of negative cases are in the same proportion for all regions from the training data set to the testing data set, not on how well each individual ROC curve performs. If the proportion is not the same, the predictive power will change, to better or to worse.

12.2 Magnitude and Sign of the Effectiveness of Data Fusion

The purpose of this research is to investigate the conditions of effective data fusion. There are very few similar studies in the area of information retrieval. As a first step, we did not try to predict the magnitude of effectiveness of data fusion, but the sign of effectiveness of data fusion, i.e., whether the fusion will be positive or negative. Accordingly, in our experiments, any pair of the IR schemes which has the highest discriminant score (or the highest probability to be a positive case, or belongs to the bin of the first rank) is likely to be a positive case. However, it does not necessarily have the highest magnitude of relative improvement, let alone best absolute performance. In addition, since the dependent variable is relative improvement of fusion, not absolute performance of fusion, higher magnitude of the effectiveness of data fusion does not lead to better performance.

For example, consider the following case (*table 12.4*).

	Precision at the 100th document	**Effectiveness of data fuson**
Scheme ***A***	0.01	(0.03 – 0.02) / 0.02 = 50%
Scheme ***B***	0.02	
Scheme ***A*** fused with Scheme ***B***	0.03	
Scheme ***C***	0.4	(0.6 – 0.5)/0.5 = 20%
Scheme ***D***	0.5	
Scheme C fused with Scheme ***D***	0.6	

Table 12.4: Example of high effectiveness of data fusion but low absolute fusion performance.

Although the effectiveness of scheme ***A*** fused with scheme ***B*** is higher than the effectiveness of scheme ***C*** fused with scheme ***D*** by 50% – 20% = 30%, its absolute performance as well as the amount of improvement are much worse compared with the latter, i.e., for absolute performance, it is 3% compared with 60%; for amount of improvement (compared to the best scheme of the two), it is 1% compared with 10%.

12.3 Limitations of The Research

According to the two conditions for effective data fusion, and our explorations, we suggest that the ratio of precisions and the normalized dissimilarity are positively related with the effectiveness of data fusion. Generally speaking, the results of our experiments support our suggestion. However, this suggestion does not implies that the effect of ratio of precisions and the normalized dissimilarity on the effectiveness of data fusion are the same for different types of topics.

We do not have a good scheme to classify the topics into different categories in such a way that for the topics in the same category the effect will be the same. We normalized the effects by averaging. In other words, when we estimated the values of the parameters in our analyses, we gave equal weight to all the topics in the training data set. The result is one set of values (coefficients for parametric equations or rank order of bins) for all topics. This normalization will make the estimation far from optimal for individual topics.

For example, the estimated coefficients of the multiple regression equation normalized over all topics are very different from the estimated coefficients for individual topics. We randomly pick up two topics from the TREC 4 routing task for comparison. The two topics are topic 28 and topic 182. (*Table 12.5*).

	Coefficient for the normalized dissimilarity	**Coefficient for the ratio of Precision**	**R^2**
For all topics	0.164	0.408	0.204
For topic 28 only	0.409	0.479	0.345
For topic 182 only	0.601	0.619	0.376

Table 12.5: Different estimations for the coefficients of multiple regression analysis for overall case and two individual topics.

From *Table 12.5*, we see that the relative weights assigned to the two parameters are very different from the overall case to the individual cases. As we have discussed in Chapter 11, if we use a line of constant value for the multiple regression equation as a discriminant line on the two dimensional plane of normalized dissimilarity and ratio of precisions, label all the cases below the line as positive and all the cases above the line as negative, we will have a specific detection rate associated with a specific false alarm rate. If we move the line perpendicular to itself across the plane, we will get a series of detection rates associated with a series of false alarm rate. The detection rates and false alarm rates we will get should be the same as using the predicted values of the effectiveness of data fusion to plot an ROC curve. Therefore, different relative weights of the two parameters means the shape of the line will be different, and, in the process of moving the line, the regions below and above the line will be different. The discriminant line that is optimal for all cases of course is unlikely to be optimal for specific groups of cases.

The $\boldsymbol{R^2}$ of the multiple regression equations for the two individual cases (0.345 and 0.376) is much larger compared to the $\boldsymbol{R^2}$ for all case (0.204). $\boldsymbol{R^2}$ is the square of the correlation coefficient between the observed value of the effectiveness of data fusion and the predicted value of the effectiveness of data fusion. It means the points representing the cases in the plane of normalized dissimilarity and ratio of precisions are more closer to the regression line. Larger $\boldsymbol{R^2}$ means we can predict the effectiveness of data fusion from the normalized dissimilarity and ratio of precision better. If $\boldsymbol{R^2} = 1$, we know exactly where to set the threshold and we can get a 100% detection rate with 0% false alarm rate. Since the $\boldsymbol{R^2}$ of the multiple regression equations for the two individual cases is much larger, the predicting power of the equations is much better.

There are many possible classification schemes for topics. One of the possible classification schemes to categorize the topic is to categorize according to the difficulty of the topics. The difficulty of a topic may be determined by the number of relevant documents in the whole collection. For TREC 4 and TREC 5 routing tasks, there are some very easy and some very difficult topics.

For the topics of TREC 4 and TREC 5 routing task, some have less than five relevant documents, some of them have more than hundreds (*Table 12.6*). Difficulty is just one of the possible classification criteria. Since our primary concern in the research is the general conditions for effective data fusion, we did not differentiate topics. It is possible that for those topics which have only a small number of relevant documents in the collection, the effect of normalized dissimilarity and ratio of precisions on the effectiveness of data fusion, is different from that of those topics which have hundreds or more relevant documents in the collection.

Number of relevant documents for the topic in the collection	**TREC 4 routing task topic frequencies**	**TREC 5 routing task topic frequencies**
0 - 5	4	6
6 - 10	0	5
11 – 20	3	5
21 – 50	5	7
51 – 100	8	7
201 – 200	18	9
201 – 500	12	2
501 - 1000	0	4

Table 12.6: Number of relevant documents for the topics of TREC 4 and TREC 5 routing task.

In addition to topics, there are other dimensions of the problem space which we did not differentiate but averaged or simply picked up one convenient point in the dimension, to limit the complexity of our analysis. For example, we used the relevance score of the 1,000th document produced by each IR scheme as the minimum relevance score of that scheme to calculate the normalized relevancy for fusion. As we have demonstrated in Chapter 1 Section 1.3.3, using different cutoff points will affect the rank order of the documents in the fused list, therefore the statistics and the results of our experiments may be different if we use the top 500, or 100, documents for fusion.

Another example is, we used the precision at the 100th document as the efficacy measure. The decision directly affects the value of ratio of precisions and the effectiveness of data fusion. However, in IR the precision usually is not constant for different cutoff points. The precision at the 100th is very different from the precision at 50th. In other words, the coefficients and the significance of the functions estimated in our experiments may vary if we use the precision at the 50th documents as our efficacy measure.

Other simplistic decisions include (1) the fusion rule, (2) using 1,000 documents to calculate the number of out-of-order pairs instead of 500 or 100 or any other cutoff points (if we use a much smaller number, e.g., 10 or 50, to calculate the schemes dissimilarity and use all the information provided by the 1,000 documents to determine the order of pairs, we will have much fewer non-deterministic pairs to which we assign “0.5” as their out-of-order score in our experiments); (3) using 25 bins or 400 bins instead of 100 bins in the non-parametric bins ranking method; etc.

Chapter 13:

Conclusion

13.1 A Brief Summary of Our Experiments

Since data fusion can improve information retrieval performance without developing new retrieval principles or algorithms, and sometimes even without using another IR system, it is a potentially powerful technique. However, data fusion does not always improve performance. Sometimes the performance of data fusion is worse than the individual schemes even though the fusion rule which been proven effective in many other cases.

This research investigates the conditions for effective data fusion in IR. We tried to identify good predictive variables for predicting effectiveness of data fusion for two IR schemes. In this research, we focus on predicting the direction of the effectiveness of data fusion, not the magnitude of the effectiveness. we have proposed two conditions for effective data fusion in IR. They are:

1. <u>Condition of Efficacy:</u> Fusion of two IR schemes with comparable

performance tends to be effective;

2. Condition of Dissimilarity: Fusion of two dissimilar IR schemes tends to be effective.

To measure dissimilarity between two IR schemes, we developed a scale, i.e., pairs out of order, and a set of algorithms, to measure and calculate the normalized distance between the ranked output lists of two IR schemes (see Chapter 4).

We used the relative improvement of the fused list, compared to the best scheme, to measure the effectiveness of data fusion.

We used the precision at the 100th document to measure efficacy of an IR scheme. In the process of our investigation, it became clear that the ratio of the precision of the two IR schemes has higher predictive power than the precision of individual schemes. We switched from the individual IR scheme precision to the ratio of precisions as our predictive variable in the later stage of our research.

Using all the IR schemes participating in the TREC 4 routing tasks (see Chapter 5 Section 5.2) as training data, we found that the two independent variables, (1) normalized dissimilarity and (2) ratio of precisions, are fairly good predictor variables to predict the sign of effectiveness of data fusion. For example, using logistic regression analysis based on the two independent variables, we can estimate the probability of a case to have positive value of effectiveness of data fusion. If we classify the cases with probability higher than 50% to be positive cases, we can correctly classify more than 78% of all cases.

Applying multiple regression analysis to the training data, we successfully reject the null hypothesis that there is no linear relationship between the effectiveness of data fusion and the two independent variables (normalized dissimilarity and ratio of precisions) at the significance level of 0.0000. The regression equation shows that the effectiveness of data fusion increases with increases in normalized dissimilarity and ratio of precisions. This is consistent with the Alan Smeaton’s founding (Smeaton, 1998). Smeaton found that, data fusion of conceptually dissimilar information retrieval strategies tend to have positive effectiveness.

On average, an increase in 10% of ratio of precisions will produce about 4.08% increase in effectiveness of data fusion, and an increase in 10% of normalized dissimilarity will produce 1.6% increase in effectiveness of data fusion.

We used ROC curve (Receiver Operating Characteristic curve, see Chapter 7 for details) to investigate the relationship between the detection rate and the false alarm rate for different statistical methods: discriminant analysis, logistic regression analysis, multiple regression analysis, and a non-parametric bin-ranking method.

We first estimate the parameters of our models by fitting the equations to training data set and then test their predictive power using the testing data set, which were the output lists of the IR schemes participating in the TREC 5 routing tasks.

The predictive power of all the three methods we have used are more or less the same in the sense that they have very similar ROC curves. Comparatively speaking, the predictive power of the non-parametric bin ranking method has the highest detection rate for the region of false alarm rate greater than approximately 60%. In the region where false alarm rate is less than approximately 60%, the ROC curves of all methods almost overlap with each other. For example, when the detection rate is approximately 60%, the false alarm rate of all the methods are approximately 20%.

If we pick the point on the ROC curve where ***detection rate + false alarm rate = 1*** as a convenient point for comparison, the performance of the three methods are the same. The results are summarized in the following table (*Table 13.1*.)

Predicting Method		Detection Rate	False Alarm Rate
Logistic regression	Training	76%	24%
	Testing	69%	31%
Discriminant analysis	Training	76%	24%
	Testing	69%	31%
Non-parametric bin ranking	Training	75%	25%
	Testing	69.5%	30.5%

Table 13.1: Table of performance of different predicting methods

Based on the ROC curves, we developed a set of rules to rationally (maximizing value and minimizing cost) predict whether the combination of two IR schemes will be more effective than an oracle.

13.2 Significance of the Research and Directions for Further Study

It appears that there is no theory to tell ***a priori*** when one should use data fusion methods to combine the outputs of different IR schemes to improve performance. In this research, we have identified two independent variables, ratio of precisions and normalized dissimilarity, to predict the effectiveness of data fusion. We have demonstrated how to use these two variables to construct ROC curves to predict when a data fusion will be positive. It seems that predictive power of the two variables is not extraordinary, but much better than predicting by chance alone.

Since the two proposed conditions for effective data fusion seem valid (but may be not complete), now we have a basic set of criteria to tell ***a priori*** whether one should employ data fusion. We can expect two IR scheme to meet the condition of efficacy if we know that on average the previous performance of the two IR schemes are pretty close, while the dissimilarity can be calculated from the two ranked output lists without knowing the actual performance of each IR scheme.

In this research, we only discuss fusion of two IR schemes. However, in the process of implementing the idea of inter-IR-scheme disstance, we have constructed a reliable scale: normalized dissimilarity. This measure may be used to test whether the conditions for effective data fusion in IR can extend from fusion of two schemes to fusion of three or more.

We can use the normalized dissimilarity to represent the distance between a pair of IR schemes, and then use proximity clustering techniques (e.g., agglomerative hierarchical clustering techniques like simple linkage clustering and complete linkage clustering) to cluster IR schemes. Each cluster contains elements which are more dissimilar to other elements within the cluster than elements in other clusters. If the two conditions for effective data fusion can extend to fusion of more than two schemes, then in each cluster, the fusion of schemes of similar performance are more likely to have positive effectiveness of data fusion.

The two conditions for effective data fusion do not completely predict effectiveness of data fusion. It seems that, after establishing these two conditions researchers can continue to investigate the additional influences of the other dimensions (e.g., those we mentioned in the previous chapter, Chapter 12, about limitation of this research) on the effectiveness of data fusion.

This research is focused on predicting the sign of the effectiveness of data fusion. Another challenging direction for researchers is moving on from predicting the sign of effectiveness to magnitude of the effectiveness.

References

Bartell, B.T., Cottrell, G.W. & Belew R.K. (1994). Automatic combination of multiple ranked retrieval systems. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 173-181.

Bartell, B.T., Cottrell, G.W. & Belew R.K. (1992). Latent semantic indexing is an optimal case of multidimensional scaling. SIGIR'92 Proceedings, NY: ACM Press, pp. 173-181.

Belkin, N.J., Cool, C., Croft, W.B. & Callan, J.P. (1993). The effect of multiple query representations on information retrieval performance. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339-346.

Belkin, N.J., Kantor, P.B., Fox, E, and Shaw, J. (1995). Combining the evidence of multiple query representations for information retrieval. Information Processing and Management, vol. 31, No. 3, pp. 431-448.

Callan, J.P., Lu, Z., & Croft, W.B. (1995). Searching distributed collections with inference networks. SIGIR'95 Proceedings, Seattle: ACM., pp.21-28.

Carmines, E.G. & Zeller, R.A. (1979). Reliability and validity assessment. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-017. Beverly Hill and London: Sage Publications.

Cleverdon, C. (1967). The Cranfield tests on index language devices. Aslib Proceedings. Vol. 19, No. 6, June, 1967.

Croft, W.B. (1995). Effective text retrieval based on combining evidence from the corpus and users. IEEE expert: intelligent systems and their applications. Vol. 10. No.4. pp.59-63.

Dreilinger, D. & Howe, A. (1997). Experiences with selecting search engines using metasearch. ACM Transactions on Information Systems. Vol. 15. No. 3, pp. 195-222.

Dunn, G. (1989). Design and analysis of reliability studies. London: Edward Arnold.

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Fox, E.A. & Shaw, J.A. (1993). Combination of multiple searches. Proceedings of the Second Text REtrieval Conference (TREC-2). National Institute of Standards and Technology Special Publication 500-215.

Fox, E.A. & Shaw, J.A. (1994). Combination of multiple searches. Proceedings of the Third Text REtrieval Conference (TREC-3). National Institute of Standards and Technology Special Publication 500-215.

Gomez, L.M., Lochbaum, C.C. & Landauer, T.K. (1990). All the right words: Finding what you want as a function of richness of indexing vocabulary. Journal of the American Society for Information Science. Vol. 41, No., 8, pp. 547-559.

Harman, D. (1993). Overview of the first Text REtrieval Conference. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 36-48.

Harman, D. (1996). Overview of the fourth Text REtrieval Conference. In D. Harman (ed.) Proceedings of the Fourth Text Retrieval Conference. Washington. DC: GPO.

Hosmer, D.W. & Lemeshow, S. (1989). Applied logistic regression. New York: John Wiley and Sons.

Hull, D. A., Pedersen, J.O. & Schutze, H. (1996). Method combination for document filtering. SIGIR' 96 Proceedings. Zurich: ACM., pp. 279-287.

Kantor, P.B. (1994a) Data fusion in information retrieval: Towards a theoretical foundation. A vector simulation model. APLab Technical Report. APLab/TR-93/3.

Kantor, P.B. (1994b) Information retrieval technique. Annual review of information science and technology. Vol 29, pp. 53-90

Kantor, P.B. (1995) Tutorial on data fusion in information retrieval. ACM SIGIR. Seattle Washington.

Kantor, P.B., Hull, D. & Ng, K.B. (1998) Advanced approaches to the statistical analysis of TREC information retrieval experiments. Technical report submitted to National Institute of Standards and Technology.

Kantor, P.B., Ng, K.B. & Hull, D. (1998) Comparison of system using pairs-out-of-order. Technical report submitted to submitted to National Institute of Standards and Technology.

Katzer, J, McGill, M.J., Tessier, J.A., Frakes, W. & Dasgupta, P. (1982). A study of the overlap among document representations. Information Technology: Research and Development. Vol. 1, No. 2, pp.261-274.

Keen, E. M. (1992) Presenting results of experimental retrieval comparisons. Information Processing & management. Vol. 28. No.4, pp.491-502. UK: Pergamon Press.

Kemeny, J.G. (1964). Random essays on mathematics, education and computers. Englewood Cliffs, N.J.: Prentice-Hall.

Klecka, William, R. (1980) Discriminant analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-019. Thousand Oaks, CA: Sage.

Larkey, L.S. & Croft, W.B. (1996). Combining classifiers in text categorization. Draft submitted to SIGIR' 96 Proceedings.

Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.180-188.

Lee, H.J. (1997). Analyses of multiple evidence combination. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.267-276.

Menard, Scott (1995) Applied logistic regression analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-106. Thousand Oaks, CA: Sage.

Ng, K.B. and Kantor, P.B. (1996). Two experiments on retrieval with corrupted data and clean queries in TREC 4 adhoc task environment: Data fusion and pattern scanning. In D. Harman (ed.) Proceedings of the Fourth Text Retrieval Conference. Washington. DC: GPO.

Ng, K.B., Loewenstern, D., Basu, C., Hirsh, H. & Kantor, P. (1997). Data fusion of machine learning methods for the TREC-5 routing task (and other works). In D. Harman (ed.) (in press) Proceedings of the Fifth Text Retrieval Conference. Washington. DC: GPO.

Norusis, Marija (1994). SPSS advanced Statistics 6.1. Chicago: SPSS Inc.

Saracevic, T, & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. Journal of the American Society for Information Science. Vol. 39, No. 3, pp. 197-216.

Smeation, A. (1998). Independence of contributing retrieval strategies in data fusion for effective information retrieval. Proceedings of the 20th BCS-IRSG Colloquium, Grenoble, France, Springer-Verlag Workshops in Computing, April 1998.

Thompson, P. (1990). A combination of expert opinion approach to probabilistic information retrieval, part 1: the conceptual model. Information Processing & Management, Vol. 26. No. 3., pp. 381 –382.

Thompson, P. (1990). A combination of expert opinion approach to probabilistic information retrieval, part 2: mathematical treatment of CEO model 3. Information Processing & Management, Vol. 26. No. 3., pp. 383 –394.

Turtle, H. & Croft, W.B. (1991). Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems. Vol. 9, No. 3, pp. 187-222.

Varshney, P.K. (1997). Scanning the issue: Special issue on data fusion. Proceedings of the IEEE. Vol. 85, No. 1, pp. 3-5.

Viswanathan, R. & Varshney, P.K. (1997). Distributed detection with multiple sensors: Part 1 -- fundamentals. Proceedings of the IEEE. Vol. 85, No. 1, pp. 54-63.

Vogt CC., Cottrell G., Belew R. and Bartell B. (1997) Using Relevance to train a Linear Mixture of Experts. In D. Harman (ed.) Proceedings of the Fifth Text Retrieval Conference. Washington. DC: GPO.

Voorhees, E.M. & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference. In D. Harman (ed.) Proceedings of the Fifth Text Retrieval Conference. Washington. DC: GPO.

Voorhess, E.M., Gupta, N.K., & Johnson-Larid, B. (1995). Learning collection fusion strategies. In Fox, Ingwersen, & Fidel (Eds.), SIGIR’95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 172-179.

Welch, T. (1984). A technique for high-performance data compression. IEEE Computer, 17(6): pp. 8-19.

Curriculum Vita

Kwong Bor Ng

Education:

1981-1985	The Chinese University of Hong Kong, Hong Kong. Major: Philosophy; Minor: Sociology. BA, June 1985.
1993-1994	Rutgers University, New Brunswick, NJ. School of Communication, Information and Library Studies.. Master of Library Studies Program. MLS, Oct., 1994.
1998-1998	Rutgers University, New Brunswick, NJ. School of Communication, Information and Library Studies.. Ph.D. program. Ph.D., October 1998.

Publications:

Ng, K.B. & Kantor, P.B. (1996). Two experiments on retrieval with corrupted data and clean queries in TREC 4 adhoc task environment: Data fusion and pattern scanning. In D. Harman (ed.) Proceedings of the Fourth Text Retrieval Conference. Washington, D.C.

Belkin, N.J., Cool, C., Koenemann, J., Ng, K.B. & Park, S. (1996). Using relevance feedback and ranking in interactive searching. In D. Harman (ed.) Proceedings of the Fourth Text Retrieval Conference. Washington, D.C.

Cool, C., Park, S., Belkin, N.J., Koeneman, J., & Ng, K.B. (1996). Information seeking behavior in new searching environments. Proceedings of the Second International Conference on Conceptions of Library and Informaion Science. Copenhagen, Denmark.

Ng, K.B., Loewenstern, D., Basu, C., Hirsh, H. & Kantor, P. B. (1997). Data fusion of machine learning methods for the TREC-5 routing task (and other works). In D. Harman (ed.) Proceedings of the Fifth Text Retrieval Conference. Washington, D.C.

Belkin, NJ., Cabezas, A., Cool, C., Kim, K., Ng, K.B., Park, S., Pressman, R., Rieh, S., Savage, P., & Xie, H. (1997). Rutgers interactive track at TREC 5. In Harmann, D., (ed.), Proceedings of the Fifth Text Retrieval Conference. Washington, D.C.

Ng, K.B., Park, S.,& Burnett, K. (1997). Control or management: A Comparison of the two approaches for establishing metadata schemes in the digital environment. Proceedings of the 60th Annual Meeting of the American Society for Information Science (ASIS) , November, 1997.

Kantor, P.B., Ng, K.B. & Hull, D.A. (1997). Comparison of systems using pairs-out-of-order. Technical report, submitted to Computer Systems Laboratory, National Institute of Standards and Technology. Gaithersbury, M.D.

Kantor, P.B., Lee, J.J., Ng, K.B. & Boros, E.. Application of LAD (Logical Analysis of Data) to the TREC6 routing task. Proceedings of the Sixth Text Retrieval Conference. Washington, D.C.

Hull, D.A., Kantor, P.B. & Ng, K.B. (1997). Advanced approaches to the statistical analysis of TREC information retrieval experiments. Technical report, submitted to Computer Systems Laboratory, National Institute of Standards and Technology. Gaithersbury, M.D.

Ng, K.B., Kantor, P.B. (1998). An Investigation of the Conditions for Effective Data Fusion in Information Retrieval: A Pilot Study To be presented in the 61th Annual Meeting of the American Society for Information Science (ASIS), October, 1998. To be published in Proceedings of the 61th Annual Meeting of the American Society for Information Science (ASIS) ,

Burnett, K., Ng, K.B. & Park, S. (1998). A Comparison of the two approaches for establishing metadata schemes. To be published in Journal of American Society for Information Science.