The Study On the Effectiveness of Information Retrieval in the Vector Space Model and the Neural Network Inductive Learning Model

Seong Hee Kim*

〈Abstract〉

This study is intended to compare the effectiveness of the neural network inductive learning model with a vector space model in information retrieval. As a result, searches responding to incomplete queries in the neural network inductive learning model produced a higher precision and recall as compared with searches responding to complete queries in the vector space model. The results show that the hybrid methodology of integrating an inductive learning technique with the neural network model can help solve information retrieval problems that are the results of inconsistent indexing and incomplete queries--problems that have plagued information retrieval effectiveness.

* Department of Library and Information Science, Chungang University

1. Introduction

The major purpose of an information retrieval(IR) system is to respond to a request for information about a particular topic by retrieving a set of documents that are related to that topic. Information retrieval systems process files of records and requests for information: they identify and retrieve records from the files in response to information queries(Van Rijsbergen, 1975). Information retrieval systems are composed of three principal parts (Cooper, 1971). The first component extracts content representations from documents and creates content indicators for the system to use to identify the documents. The second function concerns the formulation of queries, when the user chooses the words that will make up query and presents that query to the system. Finally, once the contents of the documents in the collection have been identified and the query has been formulated, the system must compare the query to the document indicators. The various methods used for this comparison are termed retrieval techniques or retrieval strategies.

Currently, most of information retrieval systems are based on the boolean algebra. The boolean retrieval process produces subsets of the document file based on a match or no-match selection process between query terms and index terms. Query terms can be matched against index terms in various combinations using AND, OR, NOT operators (Bookstein, 1978).

Typically, a user interacts with the system by formulating a query, examining the retrieval results, and then reformulating the query until satisfied with the results. A query may be formulated as follows:

(REAL AND TIME) OR CONTROL

Where REAL, TIME and CONTROL are terms to be matched in the database records and AND and OR are boolean operators. The typical response to such a query would be in the form:

(REAL AND TIME) OR CONTROL=X HITS

meaning that X records in the database meet the query requirements. The user may broaden or narrow the search requirements in response to the results.

Boolean retrieval systems assume that a query and a document can be exactly represented in the form of a set of controlled vocabulary or in natural language (Belkin et al., 1982). Thus, they simply match query descriptors against document descriptors. In the end, boolean retrieval systems generate a set of documents containing descriptors that match the query term exactly; all other documents are rejected.

There are two difficult problems with such a boolean retrieval system(Mori et al., 1990; Mozer, 1984). The first problem

centers on the users of the system, who often have a difficult time accurately specifying the information they are seeking. They may not know the specific terms used by the system to represent the documents they seek, or they may inadvertently present terms that have ambiguous meanings for the system. They may fail to include relevant descriptors in the query, or they may include irrelevant descriptors. The end result is that the set of descriptors chosen for the query is often inaccurate or incomplete.

Belkin(1982) maintains that users recognize their need for more informative queries but find it difficult to specify precisely what information is missing. Thus, a user presents information to the system, hoping that it will be sufficient for satisfactory retrieval. The user's need may be satisfied completely, partially, or not at all.

Clearly, the user of an information retrieval system has a limited ability to choose the semantically correct indicator. Kuhltau's study(1991) of the information search process from the user's perspective indicates that in the early stages of the search process, the user's state of knowledge is vague, uncertain and accompanied with confusion, disruption and frustration. When a user cannot devise correct query terms, retrieval effectiveness is very low (Bates, 1986; Blazek&Bial 1988; Borgman, 1986; Larson, 1992).

The second major problem with information retrieval systems lies within those systems. The assignment of indicators to documents is itself often inconsistent, because documents are added to the collection over long periods of time by many individuals. In addition, relevant descriptors are sometimes omitted from a search. Furthermore, since a boolean-based retrieval system depends on character by character matching of words, word ambiguities such as homographs can result in the retrieval of non-relevant items (Radecki, 1988).

Current retrieval processes are literal minded; they simply match query descriptors against document descriptors. When either query or document descriptors are faulty, traditional retrieval systems will perform poorly.

To improve the boolean-based retrieval systems, the use of partial matching and ranking techniques for information retrieval has been explored with some success (Coyle, 1985; Doszkocs, 1991; Larson, 1992). Some studies have concentrated on improving system effectiveness through devising new models of document retrieval and search strategies based on these new models (Salton and McGill, 1983; Van Rijsbergen, 1979) which include the probabilistic model (Bookstein & Swanson, 1974), the extended boolean model (Salton, Fox, & Wu, 1983; Salton & McGill, 1985), and vector space

representations (Salton, Wong, & Yang, 1975: Wong, Ziarko, & Wong, 1985).

Each of these models is based on a form of key-word retrieval that operates at a symbolic, text-matching level and ignores any semantic and contextual information in the retrieval process (Watters, 1989). Thus, these models share the literal-mindedness of their forerunners. Because of such inherent limitations, it is questionable whether extensions of the traditional approaches to information retrieval will be able to provide the mechanisms needed for intelligent information retrieval systems.

Because of the limitations of the models discussed above, many researchers have suggested applying neural networks the problem of information retrieval. Evidence provided by Belew(1986), Mozer(1984), and Wilkinson and Hingston (1992) demonstrated that neural networks could be applicable to information retrieval. However, the existing work done with neural network models is inadequate. Most previous studies didi not fully demonstrate how

neural networks improve retrieval effectiveness. Recently, Cortez, Park, and Kim(1995) investigated the usefulness of a neural network and an inductive learning model for intelligent information retrieval. In this study, the neural network inductive learning model was a hybrid model which uses a neural network augmented by an inductive algorithm. According to the study, the results of searches in response to complete queries in a neural network inductive learning model were not significantly different to that of searches in response to incomplete queries. The result suggest that a future retrieval system should have a flexibility that can respond more accurately to individual preference, changing information needs, and different retrieval situations even though the user queries not complete. However, they did not compare the results of neural network information retrieval with other existing information retrieval systems.

The retrieval effectiveness in the neural network inductive learning model will be compared to the effectiveness of the vector space model to investigate the predicted superiority of the neural network inductive learning model. A vector space model is a document representation and retrieval model. Its basic premise is that documents and queries are vectors in an n-dimensional space, where each dimension corresponds to an index term. The vector space model has been studied by information retrieval researchers. On the other hand, the neural network inductive information retrieval system which I have designed is a hybrid model which uses a neural network augmented by an inductive learning algorithm. Inductive learning and neural

network models have demonstrated the kind of flexibility and computational power that characterize cognition, particularly those aspects of cognition that users perform effortlessly and naturally (Croft and Harper, 1979; Oddy, 1977). Thus, such an approach to information retrieval promises to be particularly useful for queries that demand flexible inferencing and reasoning from incomplete or imprecise information.

2. Previous Works

As indicated in the previous section, the aim of this investigation was to determine whether or not the use of an inductive learning method and the neural network results in improved information retrieval performance. It is therefore essential to have an understanding of the current state of knowledge of the use of neural networks for information retrieval. A survey of the literature uncovered very little information on neural networks from studies of information retrieval conducted over thirty years. In this section, I discuss some of the research that has been done on enhancing retrieval system performance. The proposed IR model considered in this paper builds on and incorporates many aspects of that research.

Commercially available IR systems are based on boolean searching. The boolean retrieval process produces subsets of document files based on a match or no match selection between query terms and index terms. Boolean retrieval systems assume that a query and a document can be exactly represented in the form of a set of keywords or in natural language (Belkin et al., 1982). The disadvantage of this type of technique are well known and well documented, and a variety of aids such as thesauri are required to achieve reasonable performance. In simple cases, exact match searching 1) misses many relevant texts whose representations match the query only partially, 2) does not rank retrieved texts, 3) cannot take into account the relative importance of concepts either within the query or within the text, 4) requires complicated query logic formulation, and 5) depends on the two compared representations being drawn from the same vocabulary. Bookstein (1985) mentioned several other undesirable characteristics of boolean retrieval.

Research in exact match retrieval techniques has to, some extent, dealt with all of the problems mentioned above. The major efforts in the logic of exact match retrieval have been in making it less exact, in taking into account relative importance, and in achieving sensible ranking rules. For example, Salton and his colleagues have developed an extended vector space model, and Croft (1986) has proposed a method for making queries within a probabilistic search

model. Their experimental and theoretical investigations have indeed proven that many innovations can be made to the existing operational retrieval systems based on conventional boolean design principles.

Other efforts to improve the retrieval performance are related to partial matching techniques. The following section concentrates on the major modeling approaches that have been used for information retrieval: the vector space model, the probabilistic model, and the fuzzy set model. Many researchers (Bookstein, 1985; Robertson, 1977; Salton, 1979, and Van Rijsbergen, 1979) discuss these models of information retrieval in detail.

The evaluation of the these IR models has been a major topic of research for a number of years (Van Rijsbergen, 1979). The first important result is that all available evidence points to the superiority of partial match techniques over exact match techniques (Salton et al., 1983). Although there are some problems with making direct comparisons between the sets of retrieved documents, it appears that the difference in effectiveness is significant. Results that indicate the superiority of one technique over another in this context can be interpreted in terms of the estimates used for the weights. The use of good estimates is the major factor in obtaining the best performance from these techniques.

The use of term dependencies to modify document rankings can also improve performance, but only if the dependencies are accurately identified by the user or natural language processing techniques (Croft, 1986). Although the techniques described so far can achieve reasonable levels of performance and can be implemented efficiently in operational systems, there is still a lot of room for improvement in terms of performance. Each of these models is based on keyword retrieval which operates at a symbolic, text-matching level and ignores any semantic and contextual information in the retrieval process. To obtain much higher levels of performance, it is apparently necessary to consider knowledge-intensive techniques such as artificial intelligence.

A number of these studies in improving IR effectiveness indicate that a system that relies on a single search strategy which operates on a single document representation is not adequate for many environments. That is, these studies imply that flexibility is a key factor in improving the effectiveness of information retrieval systems.

Since 1980 many researchers have suggested applying neural networks to the area of information retrieval to improve information retrieval effectiveness. Current evidence provided by Mozer (1984), Belew (1986), and Wilkinson and Hingston (1992) demonstrated that neural networks could be

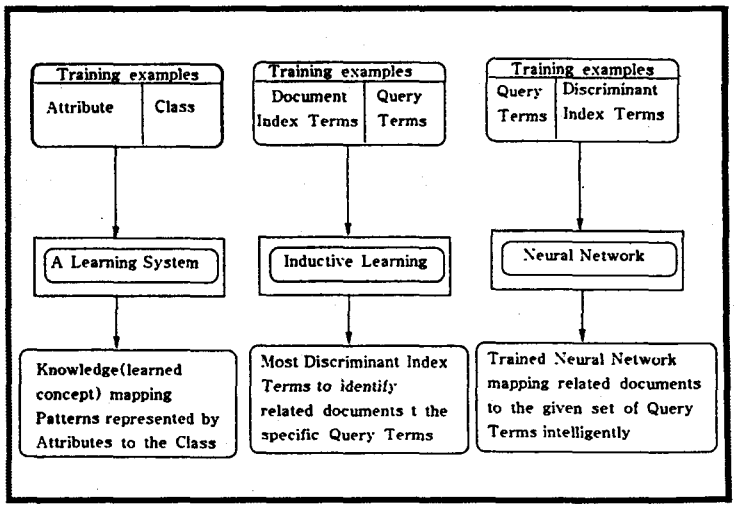
applicable to information retrieval. However, most previous studies did not fully demonstrate how neural networks improve retrieval effectiveness. Moreover, when neural networks were applied to information retrieval, the researchers did not control a number of factors that could determine retrieval effectiveness. Cherkassky and Vassilas (1989) showed how the quality of recall is affected by the number of hidden nodes for the network. The results agreed with earlier back-propagation studies (Burr, 1986) in that:

- too few hidden nodes cannot capture essential features of presented data, and
- too many nodes also have negative effects, i.e. the network attempts unnecessary classification by overreacting in response to small changes in the input data.

The implications of these studies are such that structured approaches need to be developed to determine the optimal number of nodes. Since the variability of the number of nodes remains unsolved, the optimal neural network architecture will need to be developed in future.

3. Inductive Learning and Neural Networks

The proposed information retrieval system in this study is based on a specific learning system; that is, a neural network and an inductive learning algorithm (See Figure 1). A learning system is a computer program that makes decisions based on the accumulated experience contained in successfully solved cases. Certain learning systems are concerned with learning from examples. Widely used learning algorithms in this category include inductive learning and neural networks. For this type of learning, bundles of training examples (observations of successfully solved cases) are given to the learning system. Each training example is represented by a set of attribute values describing the observation



⟨Fig1⟩ A Learning System-Suggested Methodology

and its decision class. Then, the system extracts the knowledge (learned concept for the decision).

A detailed description of the inductive learning and neural network for intelligent information retrieval can be found in the Cortez et al.' study. To provide a theoretical background on the suggested methodology, I briefly describe inductive learning and neural network in this section.

3.1. Inductive Learning

Inductive learning is a process of acquiring knowledge by drawing inductive inferences from training examples (Michalski, 1983). Such a process involves operations of generalizing, specializing, transforming, correcting and refining knowledge representations. The input to an inductive learning algorithm consists of three parts: (1) a set of training examples, (2) generalization rules and other transformation rules, and (3) the criteria for a successful inference (Park et al., 1990). Each training example has two components: a database consisting of a set of attributes, each with an assigned value; and the classification decision made by a domain expert according to the given data case. The output generated by this inductive learning algorithm is a set of decision rules consisting of inductive concept definition for each of the classes. The basis of the induction task is a set of positive and negative training examples. In the case of data collected for this study, the positive examples are documents related to the given query term, and the negative examples are all irrelevant documents.

Learning programs falling into this category include AQ-Star (Michalski, 1983), PLS(Rendell, 1986), and ID# (Quinlan, 1986). A detailed description of the learning programs can be found in Shaw et al. (1990). The inductive learning program used in this study is Quinan's (1988, 1992) C4.5, a descendant of ID3.

The induction method is based on the process of dividing a group of training examples by the value of a selected attribute, in the hope that the examples in a subgroup would belong to the same class. This program generates a classifier in the form of a decision tree. The decision tree structure includes:

- *a leaf*, indicating a class or
- *a decision node*, specifying a test to be performed on a single attribute values, with one branch and subtree for each possible outcome of the set.

C4.5 uses a method of selecting the most discriminant attribute and its threshold value to form the node of a tree using the information gain ratio criterion (Quinlan, 1988). Suppose a set of examples S contains

p positive examples and n negative examples. Then the entropy of the set (the degree of instability of the information the set has) is defined by H(S) where

If an attribute A is selected to partition

$$H(S) = -p/(p+n)\log_2 p/p+n) - n/(p+n)\log_2 n/(p+n).$$

the set into two subsets depending upon the A's value range, each subset contains positive examples and P_k positive examples and n_k negative examples where k=1,2. Then the entrophy or the instability of the information of the subset can be defined as $H(A_k)$

$$H(A) = \Sigma_k [(p_k+n_k)/(p+n)H(A_k)]$$

Thus G(A), the information gained by branching the set on attribute A is

G(A) = H(S) - H(A).

At each iteration of branching, the inductive learning algorithm examines all candidate attributes and selects the one that maximize the information gain G(A).

Inductive learning in this study is proposed to improve the retrieval effectiveness of the information retrieval system. Rather than relying on the given indexing system, the proposed system relies on index terms chosen to discriminate the most related documents as to the given query. In representing documents, the system ignores the dictated potentially inconsistent indexing scheme. Instead, it utilizes an inductive learning algorithm that isolates the most discriminant indexing terms which are related to given query terms. The selected indexing terms, are then ranked by their potential importance in discriminanation so that the degree of relatedness can be controlled by adjusting the term inclusion boundary.

By doing so, the proposed system can maintain its retrieval effectiveness even in the presence of the inconsistent and/or incomplete indexing. Semantic and syntactic ambiguity can also be mitigated by using these ranked discriminant index terms because ambiguity of terms or inconsistency of indexing are absorbed and imbedded in the rank which clarifies and quantifies the relatedness.

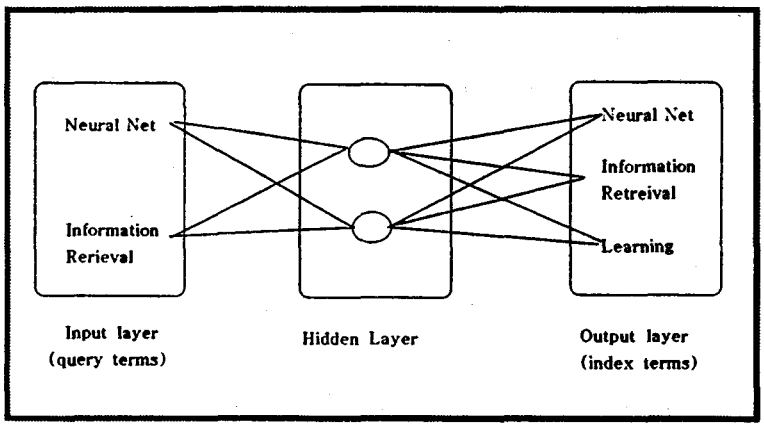
3.2. Neural Network Model

A neural network is an information processing system consisting of a number of very simple and highly interconnected processors called neurodes. The neurodes are connected with each other by many weighted links, over which signals can pass. There are several layers of neurodes in the network. Generally, the connection between neurodes occurs between layer although some networks allow the connection within one

layer. The control parameters in building the network include: (1) the network topology, (2) the learning or training algorithm, and (3) learning and momentum ratios.

A special class of neural network called multi-layer feed-forward neural network, has proved its utility and power in the development of complex classification systems (Pao, 1989; Wilson & Sharda, 1992) and is used in this study. A feed-forward network with appropriately linked weights can be used to model the causal relationship between a set of query terms and a set of related documents. The correlation is then fed into the inductive learning component of the retrieval system. The architecture of this system, as the name implies, consists of multiple layers of neurodes as shown in Fig.2. These layers are:(1) an input layer that introduces information from the environment to the network, (2) an output layer that holds the responses of the network to a given pattern, and (3) middle or hidden layers that are any layers between the input and output layers. Each unit within the middle and the output layers can have a threshold, usually referred to as biases, associated with it. Neural networks with hidden layers have the ability to develop internal representations. The middle layer neurodes are often characterized as features detectors that combine raw observations into higher order features, thus permitting the network to make reasonable generalizations (Salchenberger et al., 1992). Since there is no rigorous way of deriving a right number of hidden layers and neurodes, in later phases of this research I conducted an experimental study to determine the suitable network topology.

The outputs of nodes in one layer are transmitted to nodes in another layer through links that amplify or attenuate such outputs through weight factors. Except for the input layer nodes, the net input to each node is the sum of the weighted outputs of the nodes in the prior layer. For example, the net input to a node in layer j is



⟨Fig2⟩ An example of Neural Network

$$net_j = \sum_i w_{ij} O_i$$

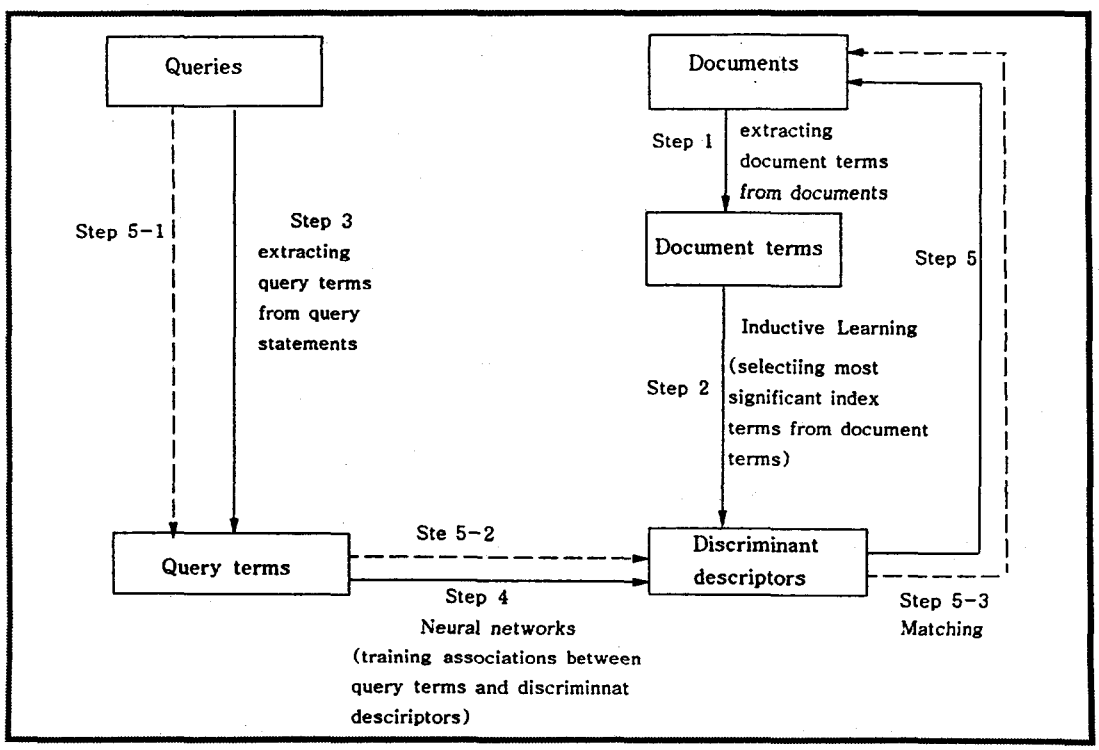
The output of node j is

$$O_j = f(net_j)$$

where f is the activation function. Activation functions map a neurode's input to its output. It is generally a threshold level of a neurode's activity at which the neurode will output a signal. Usually, a neural network model starts with a random set of weights and a training algorithm is used to adjust these weights. In this study, the backpropagation learning algorithm (Rumelhart et al., 1985; Rumelhart & McClelland, 1986) is used to perform the training requirements. It has been widely used in the development of many neural network applications today.

4. Design of the Proposed IR System

The proposed information retrieval system is based on a hybrid model consisting of an inductive learning and neural network system. The logical flowchart for the proposed information retrieval system is shown in Fig. 3. To develop the proposed system, the queries and target document database are selected initially. At first, I select all words, other than so called noise



(Fig3) Proposed Information Retreival System

words or position-holders, such as prepositions and articles from the document database to form a set of document index terms. Likewise, I select a set of query terms. Next, for each query term, I identify a relevant document subset. Using the inductive learning algorithm described above, I extract the most discriminating indexing terms that distinguish the relevant documents out of the given document database. The inducting learning technique is used as a preprocessor to create a structured set of discriminant indexing terms which when grouped by their relative importance, enable a stepwise increase or decrease of document representation.

After the set of discriminant indexing terms are identified, a process of training the neural network is initiated. The input layer of the network consists of all query terms, whereby the output layer consists of all discriminant indexing terms. When trained, the network can respond effectively to a given query in enumerating relevant documents.

4.1. Illustration of the proposed system using case examples

In this section, I explain step by step how the proposed system works. Suppose one of the queries for training reads "I want to know what neural networks are and what kinds of neural networks exist". From this query, the term "neural networks" is identified.

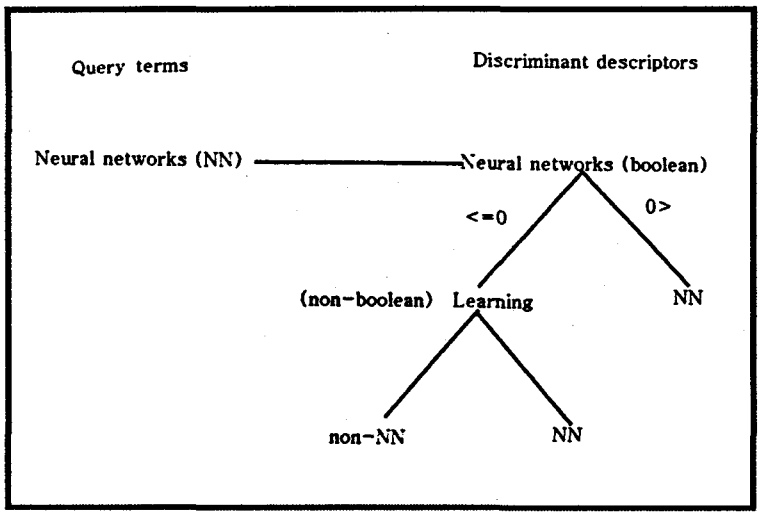
Correspondingly, the relevant set of documents is selected from the given document database. The relevant document titles are as follows:

(1) Understanding neural networks
(2) Back-propagation neural network
(3) Artificial neural network
(4) Unsupervised learning
(5) Learning internal presentations by error propagation.

From this set of documents, the following document index terms are extracted:

(1) understanding
(2) neural networks
(3) back-propagation
(4) artificial
(5) unsupervised
(6) learning
(7) internal representations
(8) error-propagation.

Next, the inductive learning algorithm is applied to extract discriminant index terms (most significant) among those described above. The resulting induction is shown in Fig.4. Through this process the terms "neural networks" and "learning" are identified as



⟨Fig4⟩ An example of induction tree

the most discriminating. Implied is that the proposed system can produce more relevant documents even if the document terms are different from the query terms, (contrasted with current IR systems which cannot find those documents whose titles contain the term "learning", but do not contain the term "neural networks" (i.e. documents #4 and #5). The following example of an "IfThenElse" statement found in the decision tree:

```
If index term includes "neural networks"
   then it is a relevant document for the
   query contaning the term "neural networks"
Else if in dex term includes "learning"
   then it is a relevant document for the
   query containing the term "learning"
Else it is an irrelevant document for the
   query containing the term "neural network"
```

Then, a simple neural network is built to map the queries and discriminant terms. The training starts by feeding a pair of query and corresponding related documents, as represented by their respective discriminant index terms (refer to Fig.2). After the network is trained, the proposed system can provide the most effective way of matching subsets of documents to a given query regardless of its completeness.

5. Implementation

I have implemented the intelligent information retrieval system on a Macintosh Quadra and a Unix-based DEC station 5000/200 workstation. ADI(American Documentation Institute) document and query collections provided by the SMART system were selected to implement the proposed system. The collections are composed of 82 documents and 35 queries on the subject of information science. The document and query collections are explained in the next section.

Generally, the sample cases are divided into two groups: a training and a test group.

Generally, the sample cases are divided into two groups: a training and a test group. The proposed system is built upon using the training group, and its performance is examined on the test group. In this study, the training group consisted of all 35 queries and 82 document titles, and the test group consisted of 18 queries out of 35 with 82 document titles.

First, the inductive learning method, C4.5, was applied to select the most significant indexing terms of the 82 document titles. The inductive learning algorithm produced 66 discriminant descriptors for 35 queries. That is to say, from the 35 queries, the system yielded 66 discriminant descriptors judged relevant to each query term. The discriminant descriptors for each query are given in Cortez's study (1995).

Then, a fully connected neural network of 118 input nodes, one hidden layer with 88 nodes, and 66 output nodes was used to train the query and discriminant descriptors. 118 input nodes and 66 output nodes were selected because the number of query terms and discriminant descriptors was 118 and 66 respectively. In order to select the hidden layer size, some exploratory experiments were performed. After several trial runs, the number having the lowest error rate and the lowest iteration times was chosen as the number of nodes in the hidden layer. As a result, the model with 88 nodes in the hidden layer 1 size resulted in the lowest number of iterations. A learning rate of 0.5 and momentum of 0.9 were chosen to control the learning process. The selection of the training parameters was also made on the basis of the lowest error rate. The stopping criteria were set such that the maximum error for each pattern in the training set did not exceed 0.01, and the maximum total number of iterations did not exceed 2000.

6. Research Design

The objective of this study is to compare the retrieval effectiveness of a neural network inductive learning model with a vector space model in terms of precision and recall. That is, this study investigates whether a neural network inductive learning system can perform the task of information retrieval more effectively than a vector space model.

This investigation deals with an evaluation of the effects of one independent variable on two dependent variables. In this study, the independent variable are the retrieval models (the neural network inductive learning model and a vector space model). The dependent variables are the precision and recall. A nonparametric test was chosen to analyze the experimental results: the Wilcoxon Ranked Test. The

one-tail test was selected to measure the effects of the dependent variable because this study is intended to compare not only the difference between variables but also the magnitude.

To evaluate the independent variable, two hypotheses are stated:

Searches responding to incomplete queries within a neural network inductive learning model result in higher precision than searches responding to complete queries in a vector space model.

$H_o : \mu_n \leq \mu_v$ $H_1 : \mu_n > \mu_v$

Searches responding to incomplete queries within a neural network inductive learning model result in higher recall than searches responding to complete queries in a vector space model.

$H_o : \mu_n \leq \mu_v$ $H_1 : \mu_n > \mu_v$

The American Documentation Institute (ADI) document and query collections were selected for use in this study so that the effectiveness of the neural network inductive learning model could be compared with previous studies which also made use of the ADI data collection. The ADI collection has been used for several experimental studies in the Library and Information Science area. It comprises 82 document and 35 queries on the subject of Library and Information Science. The documents are automatically indexed from abstracts and titles by the System for the Mechanical Analysis and Retrieval of Text (SMART) system. The queries were devised by Harvard computer science students. Each document and query term is automatically weighted in terms of term frequency by the SMART system. The stemming operation for all phase descriptors and single terms was also applied by SMART.

In order to fairly evaluate the relative effectiveness of the neural network inductive learning model and a vector space model, the document and query vectors used in the comparative experiments should differ only with respect to the retrieval method. However, for this study the queries and document terms needed to be modified, 1) in order to investigate the superiority of th neural network inductive learning model over the vector space model, and 2) due to the processing time required to urn the inductive learning and the neural network.

The specific modifications are as follows:

1) Document representation: In a vector space model, the document terms are based on the abstracts and titles. However, in a neural network inductive learning model, the document index terms were based only on titles, because of the processing time required to run the inductive learning and the neural

retrieval performance between a vector space model and the neural network inductive learning model, queries were also modified. For the neural network inductive learning model, incomplete queries were used. By contrast, complete queries were applied to the vector space model.

A simple comparison of the two models would call for identical queries to be used with both systems. However, the hypothesized advantage of the neural network inductive learning model lies in its ability to respond effectively to incomplete queries, which vector space models cannot do. This study has created incomplete queries through the random removal of three terms from each item in the original set of complete query terms.

The document collections and queries were divided into two parts -- training data and test data. All 35 queries were used to train the neural network inductive learning model, and then 18 queries out of the 35 were randomly selected to test the retrieval performance.

7. Data Analysis

The purpose of this study was to compare the retrieval results in response to incomplete queries in the neural network inductive learning model with the results in response to complete queries in a vector space model. The measures applied were precision and recall.

The hypotheses stated that searching in response to incomplete queries in the neural network inductive learning model will achieve higher precision and recall than would searching in response to complete queries in a vector space model. The experimental results were analyzed by conducting the Wilcoxon Ranked Test, using SAS (Statistical Analysis System) computer package. For all tests, a 5% probability was selected for measuring significance using a one-tail test.

The descriptive statistics for precision ratios are shown in Table 1. As a result,

〈Table 1〉 Precision ratios of incomplete queries in the NNILM and complete queries in the VSM

	Complete queries in VSM	Incomplete queries in NNILM
Mean	0.12	0.57
Median	0.13	0.53
Std Dev	0.10	0.28

the hypothesis that the retrieval in response to incomplete queries in the neural network inductive learning model yields a higher precision ratio than the retrieval in response to a complete query in a vector space model.

The mean and median of precision ratios

for incomplete queries in the neural network inductive learning model were 57% and 53%, respectively. The mean and median of precision ratios for complete queries in the vector space model were 12%and 13%, respectively. The Wilcoxon Ranked Test showed these results to be significant at P<0.05. Data from the Wilcoxon Ranked Test for precision ratios are shown in Table 2. On the other hand, searching in response to incomplete queries in the neural network inductive learning model retrieved between 25% and 100% of the relevant documents for any given question. With the vector space model, the searching retrieved between 0% and 100% of the relevant documents for any given query.

Table 3 indicates that searching in response to incomplete queries in the neural network of recall ratios in the vector space model were 49% and 50% of the relevant documents, while the mean and median of recall ratios in the neural network inductive learning model were 92% and 100%. Thus, the neural network inductive learning model produced higher recall ratios than did the vector space model. As shown in Table 4, the result was significant at the 0.05 level of testing.

〈Table 2〉 Wilcoxon results for precision ratios

Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Vector	18	179.0	333.0	31.42	9.94
Neural	18	487.0	333.0	31.42	27.06

Average Scores were used for Ties

Wilcoxon 2-sample Test (Normal Approximation)

Prob>|z| = 0.0001

〈Table 3〉 Recall ratios of incomplete queries and complete queries

	Complete queries in VSM	Incomplete queries in NNILM
Mean	0.49	0.92
Median	0.5	1
Std Dev	0.35	0.20

8. Discussions and Future Research

This study investigated the effectiveness of the neural network inductive learning model for information retrieval. It presented an original design for the application of inductive learning to information retrieval systems, in which an inductive algorithm was merged with a neuralnetwork to create a new information retrieval model. The performance results of this neural network inductive learning model were measured by comparing searches in response to complete queries in vector space model with

searches in response to incomplete queries in the neural network inductive learning model.. To measure the results, precision and recall were used. ADI (American Documentation Institute) documents and queries were selected. The collections consisted of 82 documents and 35 queries on the subject of Library and Information Science. As a result, this study demonstrated that a neural network inductive learning model produced higher precision and higher recall than the vector space model, even though incomplete query searching was performed in the neural network inductive learning model. These results may be attributable to the property of the neural network and the inductive learning model. The vector space model and neural network inductive learning model in this section are examined to try to explain these results.

〈Table 4〉 Wilcoxon results for recall ratios

Group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Vector	18	216.0	333.0	29.51	12.0
Neural	18	450.0	333.0	29.51	25.0

Average Scores were used for Ties

Wilcoxon 2-sample Test (Normal Approximation)

Prob>|z| = 0.0001

In a vector space model, the documents retrieved are ranked in terms of the similarity between query and documents. Consequently, after the system has retrieved all the documents that match the query terms exactly, it will go on to retrieve documents whose descriptors match the query term in part. A vector space model depends upon the similarity between a query and each document in the collection. Even though a weighted search based on term frequency in a vector space model was used, there is no mechanism for a vector space system to use to represent topical interest--except for the words that describe the subject. Thus, matching was based only on terms present in the query and in the document; the vector space model cannot match to a holistic concept. Thus, since a vector space model is based on keyword retrieval--which operates at a symbolic, text-matching level and ignores any semantic and contextual information in the retrieval process--the retrieval results can produce non-relevant documents.

On the other hand, a neural network associates items it is taught--that is, it groups similar items together in its structure--it can retrieve stored information from incomplete input cues (More et al., 1990; Mozer, 1984). The prototype system created for this study can start retrieving relevant documents by using whatever the query terms were submitted to the system

initially, even if those queries are not complete. During the course of retrieval, an active query term can activate other query terms. Thus, the retrieved set of documents will be ones that match either the query terms that were initially active or the induced query terms that were internally activated. This flexibility can allow the system to help compensate for incomplete queries. Futhermore, inductive learning technique has the ability to distinguish the most significant indexing terms along with varying degrees of their semantic significance. Thus, the meaning attached to a word by one document need not to be equivalent to the meaning stored in other relevant documents. Thus, the proposed system may help to overcome inconsistency in the indexing of a collection, because inductive learning models can infer the significant relationship of items through their semantic relation. The results encourage creation of a new information retrieval system that has a flexibility that can respond more accurately to individual preferences, changing information needs, and different retrieval situations even though the user queries are not complete. Furthermore, this study offers effective new techniques to the field of information retrieval techniques that could be engineered into improved information retrieval software.

Since the proposed system is an initial prototype, there are a number of phases yet to follow before complete system performance can be tested. The first phase is based on training examples. This implies that well chosen training examples should be used, because different training examples can drastically alter the retrieval of documents. For example, when the prototype is applied to different subjects and different size of training examples, the retrieval results could be changed. Thus, for future research design, a method for choosing the training examples for appropriate retrieval results needs to be considered. The second phase is related the the optimal design of the neural network topology. In this study, the selection of topology in the neural network was based on several trial runs. According to Cherkassy and Vassilar (1989), a neural network based on back-propagation learning is strongly affected by the number of layers and the choice of learning parameters. That conclusion was confirmed by the present study. The implication is that better approaches need to be developed to determine the optimal number of layers and learning parameters.

〈References〉

김 성희, 박상찬 (1995). 지능형 정보검색을 위한 신경망설계, *제 2차 한국정보관리학회 전국논문학술대회 논문집.*

남 태우, 김 성희 (1996). Integration of Information Retrieval System and Relational Database Management System, *중앙대학교 인문과학 연구, 제24집*, 153-176.

Bates, M. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science, Vol.37, no.6, 357-376*

Belew, R. K. (1987). *Adaptive information retrieval: Machine learning in associative networks.* Doctoral dissertation, University of Michigan, Ann Arbor.

Belew, R. K. (1987). A connectionist approach to conceptual information retrieval. *Proceedings of the First International Conference on Artificial Intelligence and Law*, PP. 116-125. ACM

Belkin, N. J., Addy, R. N., & Brooks, H. M. (1982). Ask for information retrieval: Part I. Background and theory. *Journal of Documentation*, Vol. 38, no.3, 145-164

Blazek, R., & Bilal, D. (1988). Problems with OPAC: A case study of an academic library. RQ, Vol. 28, no.2. 169-178

Booksetin, A., & Swanson, D. R. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, Vol. 25, no.5, 312-318

Borgman, C. L. (1986). Why are online catalogs are hard to use?: Lessons learned from information retrieval studies? *Journal of the American Society for Information Science*, Vol. 37, no.6, 384-400.

Cherkassky, V., & Vassilas, N. (1989). Back-propagation networks for spelling correction. *International Journal of Neural Networks-Research & Applications*, Vol. 1, no.3, 166-174

Coyle, K. (1985). Record matching: A discussion. *Information Technology and Libraries*, vol.4, no.1, 57-59

Doszkocs, T. El, Regia, J., & Lin, X. (1990). Connectionist models and information retrieval. *Annual Review of Information Science and Technology*, Vol. 25, 209-260

Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W., & Dasgupta, P. (1982). A study of the overlap among documents representations. *Information Technology: Research and Development*, Vol. 1, no.4, 261-274

Kim, S. H. (1994). *Intelligent Information Retrieval Using An Inductive Learning Algorithm and A Back-propagation Neural Network*, Doctoral Dissertation, University of Wisconsin- Madison.

Cortez, E. M., Sang C. Park and Seonghee Kim (1995). The Hybrid Application of An Inductive Learning Method and A Neural Network For Intelligent *Information Retrieval, Information Processing & Management*, Vol. 31, No.6, 789-813.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, Vol. 42, no.5, 361-371

Kwok, K. L. (1989). A neural network for probabilistic information retrieval. *SIGIR Forum*, Vol. 23, no1-2, 21-30

Kwok, K. L. (1990). Application of neural network to information retrieval. *International Joint Conference on Neural Networks, II*, 623-626.

Larson, R. R. (1992). Evaluation of advanced retrieval techniques in an experimental online catalog. *Journal of the American Society for Information Science*, Vol. 43, no.1, 34-53

Michalski, R. S. (1983). A theory and methodology of inductive learning, *Artificial Intelligence*, Vol. 20, no.2, 111-161.

Mori, H., Cheng Long Chung, Yousuke Kinoe, & Yoshi Hayashi (1990). An adaptive document retrieval system using neural network. *International Journal of Human-Computer Interaction*, Vol. 2 no.3, 267-280

Mozer, M. (1984). Inductive information retrieval using parallel distributed computation. Technical Report. La Jolla, Calif.: ICS, UCSD.

Pao, M. L., & Worthen, D. B. (1989). Retrieval effectiveness by semantic and citation searching. *Journal of the American Society for Information Science*, Vol. 40, no. 4, 226-235

Park, S. C., Piramuthu, S., Raman, N., & Shaw, M. J. (1990). Integrating inductive learning and simulation in rule-based scheduling. In G. Gottlob, & W. Nejdl (Eds), *Lecture notes in artificial intelligence #462: Expert systems in engineering, subseries or lecture notes in computer science*, Berlin: Springer, 152-167

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, Vol. 1, no.1, 82-106

Quinlan, J. R. (1988). Decision trees and multi-valued attributes. *Maching Learning*, 11(2), 305-318

Quinlan, J. R. (1992). C4.5: *Programs for machine learning*. San Mateo, Calif: Morgan Kaufmann

Radecki, T. (1988). The trends in research on information retrieval-the potential for improvements in conventional Boolean retrieval systems. *Information Processing & Management*, Vol. 24, no.3, 219-227

Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*. Vol. 1, no.2, 177-226

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations*. Cambridge, Mass: Bradford.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. *Technical Reports*, La Jolla, Calif.: ICS, UCSD

Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Science*, 23, 899-916

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*, New York: McGraw-Hill

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM.*, Vol. 26., no.11, 1022-1036

Salton., G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613-620.

Shaw, M. J., Gentry, J. A., & Piramuthu, S. (1990). Inductive learning methods for knowledge-based decision support: A comparative analysis. *Computer Science in Economics and Management*, 3, 147-165.

Van Rijsbergen, C.J. (1979). *Information Retrieval Experiment(2nd edition)* London: Butterworths.

Watters, C. R. (1989). Logic framework for information retrieval. *Journal of the American Society for Information Science*, Vol. 40, no.5, 311-324

Wilkinson, R., & Hingston, P. (1992). Incorporating the vector space model in a neural network used for document retrieval. *Library HI Tech*, Vol. 10, no. 1-2, 69-75.

Wilkinson, R., & Hingston, P. (1991). Using the cosine measure in a neural network for document retrieval. *14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 202-210

Wilson, R., & Sharda, R. (1992). Neural networks, OR/MS Today, 36-42

Wong, S. K., Ziarko, Wl, & Wong, P. C. N. (1985). Generalized vector space model in information retrieval. *Proceedings of the Seventh International Conference on Information Storage and Retrieval*, 18-25.