NEIGHBORHOOD-BASED INFORMATION SYSTEMS

A Thesis
Submitted to the Faculty of Graduate Studies and Research
In Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Mathematical Sciences
Department of Computer Science
School of Mathematical Sciences
Faculty of Arts and Science
Lakehead University

By
Xuechun Chen
Thunder Bay, Ontario
July 1997

0-612-33352-3

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENT

I would like to express heartfelt thanks and appreciation to my thesis supervisor, Dr. Y. Y. Yao, for his guidance, creative ideas, valuable comments, and especially for his persistent encouragement.

I also would like to thank my external examiner, Prof. W. W. Koczkodaj, and my internal examiner Prof. L. D. Black, for their valuable contributions.

I would like to thank all the faculty of the Department of Computer Science. Dr. M. W. Benson, Dr. X. Li and Dr. Y. Y. Yao provided academic instructions. Dr. L. K. Roy provided valuable help, and Lana Rizzuto, secretary, greatly helped me in many ways.

I would also like to thank my wife, Charlene Fang, my daughter, Angela Chen, my parents and parents-in-law for their understanding and encouragement.

ABSTRACT

The concept of neighborhood systems originated from studies in topology. In this study, semantical interpretations of neighborhood systems are given by introducing certain assumptions on neighborhoods. Based upon the notion of neighborhood systems, a structure called $\cap$−closure is proposed, which is a convenient and powerful tool to model the nearness among elements in a universe. A binary relation is defined on the $\cap$−closure, based on which it is possible to rank all elements in a universe according to their nearness to a given element.

The notion of information systems is enriched by adding neighborhood systems on attribute values. A Pawlak information system,

$$PT = \{O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}\},$$

is generalized to a neighborhood-based information system,

$$NT = \{O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}, \{NS(V_a) \mid a \in AT\}\}.$$

In a Pawlak information system, the retrieval method is based upon an exact matching mechanism by using the equality relation =. In this study, a retrieval model is proposed in the framework of neighborhood-based information systems. Queries are relaxed by using the notion of neighborhood systems. The retrieved results are ranked by adopting the proposed ranking structure. This model provides more useful results than that of Pawlak information systems.

Chapter 1

INTRODUCTION

The concept of information systems, proposed by Pawlak (1981) and Marek (1973, 1975, 1976, 1977, 1981), has received much attention in the area of information sciences. It provides a framework for the study of information retrieval and data analysis. There are many applications in such areas as: machine learning, data analysis, pattern classification, and rough set approximation (Pawlak, 1991).

In a Pawlak information system, each object in a universe is characterized by a set of attributes. The information about an object is represented by the values of its attributes. By defining the trivial equality relation = on attributes, all objects of a universe in a given information system are partitioned into classes (subsets). Each class may be associated with a unique description, and conversely, with a description one may associate the corresponding class (Pawlak, 1981).

In many situations, the trivial equality relation on attributes is not sufficient to describe vague and imprecise information in some applications. To solve this problem, many other types of relations on attributes have been proposed (Orlowska, 1986; Slowinski and Vanderpooten, 1995; Wasilewska, 1989; Yao and Wong, 1995).

Wasilewska (1989) extended the equality binary relation = on attributes to an equivalence relation. Two values selected from an attribute are considered to be equivalent if they satisfy the conditions of reflexivity, symmetry and transitivity. These two values are not necessarily equal to each other. The equivalence class induced on an attribute by this relation is interpreted as a condition of knowledge representation. Orlowska (1986) proposed the idea of classifying a universe into various classes by seeking similar and conditional relationships between attributes, instead of equivalence relations. With respect to some attributes, a subset of objects is considered to be similar and is associated with a certain condition. Slowinski and Vanderpooten (1995) further presented a method that extends the indiscernibility relation to a similarity relation. The only required property is reflexivity. Through a similarity relation, elements that are sufficiently close or similar are grouped into the same class. Yao and Wong (1995) used arbitrary binary relations on attribute values in information systems. The properties of relations on attributes determine the corresponding properties of induced relations of objects. In this way, the universe can be also grouped into different classes of subsets by using an arbitrary binary relation.

The studies on information systems mentioned above are focused on the classification of elements through binary relations. All elements in an information system are classified into individual subsets of a universe, whose elements are considered to be indiscernible, or similar. Each class of subset represents one piece of knowledge about the information system.

Lin (1988, 1996) adopted the notion of neighborhood systems from topology, which generalized the concept of indiscernibility into that of neighborhood. In this framework, an element of a universe is associated with a nonempty family of subsets of the universe. This family of subsets is called a neighborhood system of the element, and each subset is called a neighborhood. Neighborhood systems directly classify the universe into distinctive classes. The elements in one neighborhood of an element can be interpreted to be somewhat indiscernible or at least not noticeably distinguishable. Those elements of another neighborhood of the same neighborhood system can be interpreted to be somewhat indiscernible from an alternative point of view. With respect to an equivalence relation, the equivalence class containing a given element may be interpreted as a neighborhood of that element. For an arbitrary binary relation, the successor of an element may be interpreted as a neighborhood. To some extent, one may consider a neighborhood system as defined by n binary relations, each of which defines a particular neighborhood.

The main objective of this thesis is to propose a framework of neighborhood based information systems by combining Pawlak information systems and neighborhood systems. Semantic information is added into neighborhood systems. It provides a more realistic method to represent relationships among attribute values. Operations on neighborhood systems are introduced and studied, based on the notion of power algebra (Brink, 1993). A nearness relation is defined, which captures the information provided by a neighborhood system. Within the proposed framework, a new model is introduced for approximate retrieval. This extends the traditional retrieval methods using exact matching (Salton and McGill, 1983). The retrieval results can be arranged into a hierarchy more flexible than that of traditional methods.

This thesis is arranged as follows: In Chapter 2, basic concepts and notions are introduced and studied. In Chapter 3, the process of constructing a neighborhood based information system is provided. In Chapter 4, a retrieval model is introduced. Examples are used to illustrate the basic idea of the model. Chapter 5 gives a summary of this thesis.

Chapter 2

NEIGHBORHOOD SYSTEMS

In this chapter, basic concepts of neighborhood systems will be discussed. The notion of neighborhood systems is formulated within the framework of power set algebra (Brink, 1993). It provides a convenient and powerful tool for representing nearness between elements. Operations on neighborhood systems are defined and their properties are studied. An ordering relation is introduced, which can be used to rank elements with respect to their nearness to an element.

2.1 Set Algebra

In this study, a reference set, or universe U, is assumed to be finite and nonempty. The family of all subsets of U is called the power set of U. The cardinality of the power set is $2^{|U|}$, where $|U|$ is the cardinality of U. The power set of U is often denoted by 2^U.

Set algebra is the system $(2^U, \cap, \cup, \neg)$, where $\cap$, $\cup$, $\neg$ are set intersection, union, and complement, respectively. Set-theoretic operations satisfy the following axioms (Whitesitt, 1962), for $A, B, C \in 2^U$:

$$Idempotence: \quad A \cap A = A, \quad A \cup A = A;$$

$$
\begin{aligned}
\textit{Commutativity}: &\quad A\cap B = B\cap A,\quad A\cup B = B\cup A;\\
\textit{Associativity}: &\quad (A\cap B)\cap C = A\cap(B\cap C),\\
&\quad (A\cup B)\cup C = A\cup(B\cup C);\\
\textit{Distributivity}: &\quad A\cap(B\cup C) = (A\cap B)\cup(A\cap C),\\
&\quad A\cup(B\cap C) = (A\cup B)\cap(A\cup C);\\
\textit{Absorption}: &\quad A\cap(A\cup B) = A,\\
&\quad A\cup(A\cap B) = A;\\
\textit{De Morgan's Law}: &\quad \neg(A\cap B) = (\neg A)\cup(\neg B),\\
&\quad \neg(A\cup B) = (\neg A)\cap(\neg B).
\end{aligned}
$$

If a set A consists entirely of elements that are members of another set B, then A is called a subset of B, written $A \subseteq B$. If B contains at least one element not in A, A is called a proper subset of B, written $A \subset B$.

2.2 Power Set Algebra

From an arbitrary algebra, one may define another algebra called power algebra (Brink, 1993). Let $\circ : U \times U \longrightarrow U$ be a binary operation defined on a set U. One can define the power operation $\circ^+$ on subsets of U as follows, for $X, Y \subseteq U$:

$$X \circ^+ Y \;=\; \{x \circ y \mid x \in X, y \in Y\}. \tag{2.1}$$

In general, one may lift an n-nary operation $f : U^n \longrightarrow U$ on elements of U into its corresponding power operation f^+ (Brink, 1993):

$$f^+(X_0, \ldots, X_{n-1}) = \{f(x_0, \ldots, x_{n-1}) \mid x_i \in X_i,\;\; i = 0, \ldots, n-1\}.$$

for any $X_0, \ldots, X_{n-1} \subseteq U$. This provides a universal-algebraic construction approach. For any algebra $(U, f_1, \ldots, f_k)$ with base set U and operations $f_1, \ldots, f_k$, its power algebra is given by $(2^U, f_1^+, \ldots, f_k^+)$.

Example 1 Let $(R, *, /, +, -)$ be the algebra of real numbers R with four arithmetic binary operations. From a binary operation $*$ on R, one can obtain its power operation on closed intervals of R as follows (Alefeld and Heraberger, 1983; Moore, 1966):

$$[a_1, a_2] *^+ [b_1, b_2] = \{a * b \mid a_1 \le a \le a_2, b_1 \le b \le b_2\}. \tag{2.2}$$

This produces an interval number algebra $(I(R), *^+, /^+, +^+, -^+)$, where $I(R)$ is the set of all intervals of R. The power operations are closed, i.e., the resulting sets are also intervals. ■

One may define the power set algebra by lifting the set algebra $(2^U, \cap, \cup, \neg)$ into the system:

$$(2^{2^U}, \sqcap, \sqcup, \xi), \tag{2.3}$$

where $\sqcap, \sqcup$ and ξ correspond to power set intersection, union and complement, respectively. According to equation (2.1), it is easy to obtain the following definitions: for $\mathcal{A}, \mathcal{B} \in 2^{2^U}$:

$$\begin{aligned} \mathcal{A} \sqcap \mathcal{B} &= \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\}, \\ \mathcal{A} \sqcup \mathcal{B} &= \{A \cup B \mid A \in \mathcal{A}, B \in \mathcal{B}\}, \\ \xi \mathcal{A} &= \{\neg A \mid A \in \mathcal{A}\}. \end{aligned} \tag{2.4}$$

$\mathcal{A}$ and $\mathcal{B}$ are families of subsets of U.

Example 2 Let $U = \{a, b, c, d\}$ and $\mathcal{A} = \{\{a\}, \{b, c\}\}$, $\mathcal{B} = \{\{b\}, \{a, b\}\}$ be two subsets of 2^U. From equation (2.4), it follows:

$$\begin{aligned}
\mathcal{A} \sqcap \mathcal{B} &= \{\{a\} \cap \{b\}, \{a\} \cap \{a, b\}, \{b, c\} \cap \{b\}, \{b, c\} \cap \{a, b\}\} \\
&= \{\{\emptyset\}, \{a\}, \{b\}\}, \\
\mathcal{A} \sqcup \mathcal{B} &= \{\{a\} \cup \{b\}, \{a\} \cup \{a, b\}, \{b, c\} \cup \{b\}, \{b, c\} \cup \{a, b\}\} \\
&= \{\{a, b\}, \{b, c\}, \{a, b, c\}\}, \\
\xi \mathcal{A} &= \{\neg\{a\}, \neg\{b, c\}\}, \\
&= \{\{b, c, d\}, \{a, d\}\}.
\end{aligned}$$

■

Theorem 2.1 *Suppose $\circ$ is a binary operation on 2^U and $\circ^+$ is the corresponding power operation on 2^{2^U}. Then*

$$\begin{aligned}
&(a). \quad \textit{if } \circ \textit{ is commutative,} \quad \circ^+ \textit{ is commutative,} \\
&(b). \quad \textit{if } \circ \textit{ is associative,} \quad \circ^+ \textit{ is associative.}
\end{aligned}$$

Proof. Suppose $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are three subsets of 2^U.
(a) Assume if $D \in \mathcal{A} \circ^+ \mathcal{B}$. There must exist $A \in \mathcal{A}, B \in \mathcal{B}$ satisfying $D = A \circ B$. By the commutativity of $\circ$, $B \circ A = D$. Hence $D \in \mathcal{B} \circ^+ \mathcal{A}$. Similarly, if $D \in \mathcal{B} \circ^+ \mathcal{A}$, then $D \in \mathcal{A} \circ^+ \mathcal{B}$. Therefore, $\circ^+$ is commutative.
(b) Assume $D \in \mathcal{A} \circ^+ (\mathcal{B} \circ^+ \mathcal{C})$. There must exist $A \in \mathcal{A}, B \in \mathcal{B}, C \in \mathcal{C}$ satisfying $D = A \circ (B \circ C)$. By the associativity of $\circ$, $(A \circ B) \circ C = D$, so $D \in (\mathcal{A} \circ^+ \mathcal{B}) \circ^+ \mathcal{C}$. Similarly, if $D \in (\mathcal{A} \circ^+ \mathcal{B}) \circ^+ \mathcal{C}$, then it follows $D \in \mathcal{A} \circ^+ (\mathcal{B} \circ^+ \mathcal{C})$. Therefore, $\circ^+$ is associative. ■

For the algebra $(2^U, \cap, \cup)$, its corresponding power algebra is $(2^{2^U}, \sqcap, \sqcup)$. Operations $\cap$ and $\cup$ are distributive. However, $\sqcap$ and $\sqcup$ are not necessarily distributive. This may be illustrated by an example. Consider a universe $U = \{a, b, c\}$, and three subsets of 2^U:

$$\mathcal{A} = \{\{a\}, \{b, c\}\}, \quad \mathcal{B} = \{\{b\}, \{c\}\}, \quad \mathcal{C} = \{\{c\}, \{a, b\}\}. \tag{2.5}$$

One obtains the following results:

$$\begin{aligned}
\mathcal{B} \sqcup \mathcal{C} &= \{\{c\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}, \\
\mathcal{A} \sqcap (\mathcal{B} \sqcup \mathcal{C}) &= \{\{\emptyset\}, \{a\}, \{b\}, \{c\}, \{b, c\}\}, \\
\mathcal{A} \sqcap \mathcal{B} &= \{\{\emptyset\}, \{b\}, \{c\}\}, \\
\mathcal{A} \sqcap \mathcal{C} &= \{\{\emptyset\}, \{a\}, \{b\}, \{c\}\}, \\
(\mathcal{A} \sqcap \mathcal{B}) \sqcup (\mathcal{A} \sqcap \mathcal{C}) &= \{\{\emptyset\}, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}.
\end{aligned}$$

Note that

$$\mathcal{A} \sqcap (\mathcal{B} \sqcup \mathcal{C}) \neq (\mathcal{A} \sqcap \mathcal{B}) \sqcup (\mathcal{A} \sqcap \mathcal{C}).$$

Theorem 2.2 *With respect to power operations $\sqcap$ and $\sqcup$, the following properties hold: for $\mathcal{A}, \mathcal{B}, \mathcal{C} \in 2^{2^U}$,*

$$(a).\ \mathcal{A} \sqcap (\mathcal{B} \sqcup \mathcal{C}) \subseteq (\mathcal{A} \sqcap \mathcal{B}) \sqcup (\mathcal{A} \sqcap \mathcal{C}), \tag{2.6}$$

$$(b).\ \mathcal{A} \sqcup (\mathcal{B} \sqcap \mathcal{C}) \subseteq (\mathcal{A} \sqcup \mathcal{B}) \sqcap (\mathcal{A} \sqcup \mathcal{C}). \tag{2.7}$$

Proof. (a) Assume $D \in \mathcal{A} \sqcap (\mathcal{B} \sqcup \mathcal{C})$. There must exist $A \in \mathcal{A}, B \in \mathcal{B}, C \in \mathcal{C}$ satisfying $D = A \cap (B \cup C)$. By the distributivity of $\cap$, $D = (A \cap B) \cup (A \cap C)$. Hence, $D \in (\mathcal{A} \sqcap \mathcal{B}) \sqcup (\mathcal{A} \sqcap \mathcal{C})$. Therefore, equation (2.6) holds.

(b) Assume $D \in \mathcal{A} \sqcup (\mathcal{B} \sqcap \mathcal{C})$. There must exist $A \in \mathcal{A}, B \in \mathcal{B}, C \in \mathcal{C}$ satisfying $D = A \cup (B \cap C)$. By the distributivity of $\cap$, $D = (A \cup B) \cap (A \cup C)$. Hence $D \in (\mathcal{A} \sqcup \mathcal{B}) \sqcap (\mathcal{A} \sqcup \mathcal{C})$. Therefore, equation (2.7) holds. ■

In general, the power operation $\circ^+$ may retain some of the properties of $\circ$, but not all of them (Brink, 1993). Following Theorem 2.1, if $\circ$ is commutative and associative so is $\circ^+$. If an unary operation $\circ$ is an involution, i.e., $\circ(\circ(x)) = \circ(x), x \in U$, so is $\circ^+$. The distributivity is not retained in the power algebra, but there exist the relationships as shown in equations (2.6) and (2.7). Idempotence does not hold either. The De Morgan's law may not be retained.

2.3 Interval Set Algebra

The power set algebra is a general extension of set algebra. One may consider various special cases. The notion of interval sets is such an example. Interval set algebra is a counterpart of interval-number algebra for the purpose of representing qualitative information (Yao, 1993).

Definition 2.1 *Let U be a finite and nonempty set. An interval set $\mathcal{A}$ is a family of subsets, i.e., $\mathcal{A} \in 2^{2^U}$, such that each element in $\mathcal{A}$ falls between two subsets of U. An interval set can be expressed as:*

$$\mathcal{A} = [A_1, A_2] = \{A \in 2^U \mid A_1 \subseteq A \subseteq A_2\}, \tag{2.8}$$

where A_1 is called the lower bound, and A_2 the upper bound, of interval set.

Let $I(2^U)$ denote the set of all closed interval sets. Let $\mathcal{A}, \mathcal{B}$ be two interval sets in $I(2^U)$, their intersection, union and complement, are defined by:

$$\begin{aligned} \mathcal{A} \sqcap \mathcal{B} &= \{A \cap B \mid A \in \mathcal{A}, B \in \mathcal{B}\}, \\ \mathcal{A} \sqcup \mathcal{B} &= \{A \cup B \mid A \in \mathcal{A}, B \in \mathcal{B}\}. \\ \xi \mathcal{A} &= \{\neg A \mid A \in \mathcal{A}\}. \end{aligned} \tag{2.9}$$

In fact, the intersection, union and complement of interval sets are closed on $I(2^U)$, namely, $\mathcal{A} \sqcap \mathcal{B}, \mathcal{A} \sqcup \mathcal{B}$ and $\xi \mathcal{A}$ are also interval sets. They are explicitly computed by the following formulas:

$$\begin{aligned} \mathcal{A} \sqcap \mathcal{B} &= [A_1 \cap B_1, A_2 \cap B_2], \\ \mathcal{A} \sqcup \mathcal{B} &= [A_1 \cup B_1, A_2 \cup B_2], \\ \xi \mathcal{A} &= [\neg A_2, \neg A_1]. \end{aligned} \tag{2.10}$$

The operations of interval set algebra retain the properties of commutativity, associativity. Interval set algebra is a subalgebra of power set algebra $(2^{2^U}, \sqcap, \sqcup, \xi)$, it has more properties than arbitrary power sets. For example, the operations of interval set algebra satisfy idempotence, distributivity, absorption and De Morgan's law.

2.4 Basic Concepts of Neighborhood Systems

The concept of neighborhood systems originated from studies of topological space and its generalization called Frechet (V)space (Sierpinski and Krieger, 1956). Let U be a finite and nonempty set. For an element $x \in U$, a neighborhood of x, denoted by $n(x)$, is a subset $n(x) \subseteq U$. A neighborhood of x may or may not contain x itself. A nonempty family of neighborhoods of x, denoted by $NS(x)$, is called a neighborhood system of x. A neighborhood system of U, denoted by $NS(U)$, is the collection of $NS(x)$ for all $x \in U$. $NS(U)$ determines a Frechet space, or briefly (V)space, written $(U, NS(U))$. There is no additional requirement on neighborhood systems. Different neighborhood systems define different Frechet (V)spaces (Sierpinski and Krieger, 1956). A neighborhood system may also be interpreted as an operator that transforms an element in U to one in 2^{2^U}, i.e. $NS(x) \subseteq 2^U$.

Example 3 Consider a universe $U = \{a, b, c\}$. One can define neighborhoods of a, such as $n_1(a) = \{a\}$ and $n_2(a) = \{a, b\}$. One may also define some neighborhood systems of a, such as

$$NS_1(a) = \{\{a\}\}, \quad NS_2(a) = \{\{a,b\},\{a,c\}\}, \quad NS_3(a) = \{\{a,b\},\{a,c\},\{a,b,c\}\}.$$

The power set of U is

$$2^U = \{\{\emptyset\},\{a\},\{b\},\{c\},\{a,b\},\{a,c\},\{b,c\},\{a,b,c\}\}.$$

Neighborhood systems are subsets of 2^U, i.e.,

$$NS_1(a), NS_2(a), NS_3(a) \subseteq 2^U.$$

Two examples of Frechet (V)spaces are:

$$\begin{aligned} (U, NS_1(U)): \quad & NS_1(a) = \{\{a\},\{a,c\}\}, \\ & NS_1(b) = \{\{b\},\{a,b\}\}, \\ & NS_1(c) = \{\{c\},\{b,c\}\}; \\ (U, NS_2(U)): \quad & NS_2(a) = \{\{a,b\},\{a,c\},\{a,b,c\}\}, \\ & NS_2(b) = \{\{b\},\{a,b\},\{a,b,c\}\}, \\ & NS_2(c) = \{\{a,c\},\{b,c\},\{a,b,c\}\}. \end{aligned}$$

■

The notion of neighborhood systems may be formulated based on the power algebra $(2^{2^U}, \sqcap, \sqcup)$. By lifting set-theoretic operations $\cap$ and $\cup$ on 2^U to power operator $\sqcap$ and $\sqcup$ on 2^{2^U}, one may define operations on neighborhood systems.

Definition 2.2 *For an element $x \in U$, suppose*

$$NS_1(x) = \{n_{1,i}(x) \subseteq U \mid i \in I\}, \quad NS_2(x) = \{n_{2,j}(x) \subseteq U \mid j \in J\}, \tag{2.11}$$

are two neighborhood systems. Their intersection and union are defined by:

$$\begin{aligned} NS_1(x) \sqcap NS_2(x) &= \{n_{1,i}(x) \cap n_{2,j}(x) \mid n_{1,i}(x) \in NS_1(x), n_{2,j}(x) \in NS_2(x)\}, \\ NS_1(x) \sqcup NS_2(x) &= \{n_{1,i}(x) \cup n_{2,j}(x) \mid n_{1,i}(x) \in NS_1(x), n_{2,j}(x) \in NS_2(x)\}. \end{aligned}$$

There is a special case in which a neighborhood system contains only one neighborhood, such as $NS_1(x) = \{n_{1,1}(x)\}$ and $NS_2(x) = \{n_{2,1}(x)\}$. The intersection and union of $NS_1(x)$ and $NS_2(x)$ can be simplified into:

$$\begin{aligned} NS_1(x) \sqcap NS_2(x) &= \{n_{1,1}(x) \cap n_{2,1}(x)\}, \\ NS_1(x) \sqcup NS_2(x) &= \{n_{1,1}(x) \cup n_{2,1}(x)\}. \end{aligned} \tag{2.12}$$

The results are again neighborhood systems with one neighborhood only. The properties of neighborhood system algebra $(2^{2^U}, \sqcap, \sqcup)$ are those of power set algebra. Both commutativity and associativity hold for $\sqcap$ and $\sqcup$. Idempotence and distributivity do not hold in general.

2.5 Modeling Nearness Using Neighborhood Systems

So far, the notion of neighborhood systems is formulated from a purely mathematical point of view. In order to apply such a notion, it is necessary to examine its semantic interpretations. Neighborhood models nearness between two elements in a space. It originated from the abstraction of geometric notion, in which two points are "close to", "analogous to", or "approximate to" each other. Lin (1988, 1996) discussed applications of neighborhood systems in database management systems. By interpreting neighborhood systems as a method to represent closeness, or nearness, one may talk about approximate retrieval in information systems. When retrieving elements with certain characteristics, those elements that are close enough are examined, if the exact ones do not exist.

The general definition of neighborhood systems does not impose any constraints on neighborhoods. A neighborhood of an element x can either contain or does not contain x. In order to capture the physical interpretation of nearness, certain assumption should be made.

Assumption 2.1 *In a neighborhood system $NS(x) = \{n_i(x) \in 2^U \mid i \in I\}$, the element x is contained in each of its neighborhoods, namely, $x \in n_i(x)$, for all $i \in I$.*

This assumption gives an intuitive interpretation of neighborhoods. It states that an element x must be in all of its neighborhoods. A neighborhood system having this property is called a reflexive neighborhood system.

Definition 2.3 *For a neighborhood system $NS(x)$, the $\cap$-closure $NS^*(x)$ is the minimum subset of 2^U, which contains $NS(x)$ and entire set U, and is closed under set intersection $\cap$.*

The $\cap$-closure $NS^*(x)$ of a neighborhood system $NS(x)$ satisfies the following conditions:

$$\begin{aligned}
&(a1) \quad NS(x) \subseteq NS^*(x),\\
&(a2) \quad U \in NS^*(x),\\
&(a3) \quad n_j(x) = \left[\bigcap_{i \in J \subseteq I} n_i(x)\right] \in NS^*(x).
\end{aligned}$$

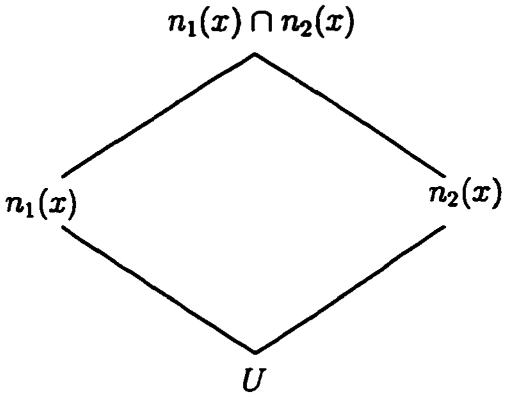


Figure 2.1: Diagram of $\cap$-closure for $NS(x) = \{n_1(x), n_2(x)\}$

The notion of $\cap$-closure will be used to design a method for evaluating the nearness of elements in a universe.

Example 4 Consider a neighborhood system with two neighborhoods, $NS(x) = \{n_1(x), n_2(x)\}$. Assume $n_1(x) \neq n_2(x)$. The $\cap$-closure is given by:

$$NS^*(x) = \{n_1(x) \cap n_2(x), n_1(x), n_2(x), U\}, \tag{2.13}$$

which is shown in Figure 2.1, where the edges represent the relation $\subset$. For example, there is an edge connecting $n_1(x) \cap n_2(x)$ and $n_1(x)$. Edges that can be obtained from the transitivity of $\subset$ are not shown. ■

Example 5 Consider a neighborhood system with three neighborhoods, $NS(x) = \{n_1(x), n_2(x), n_3(x)\}$. Assume $n_i(x) \neq n_j(x)$, where $i, j \in \{1, 2, 3\}$ and $i \neq j$. The $\cap$-closure is given by:

$$\begin{aligned} NS^*(x) &= \{n_1(x) \cap n_2(x) \cap n_3(x), \\ &\quad n_1(x) \cap n_2(x), n_1(x) \cap n_3(x), n_2(x) \cap n_3(x), \\ &\quad n_1(x), n_2(x), n_3(x), U\}, \end{aligned}$$

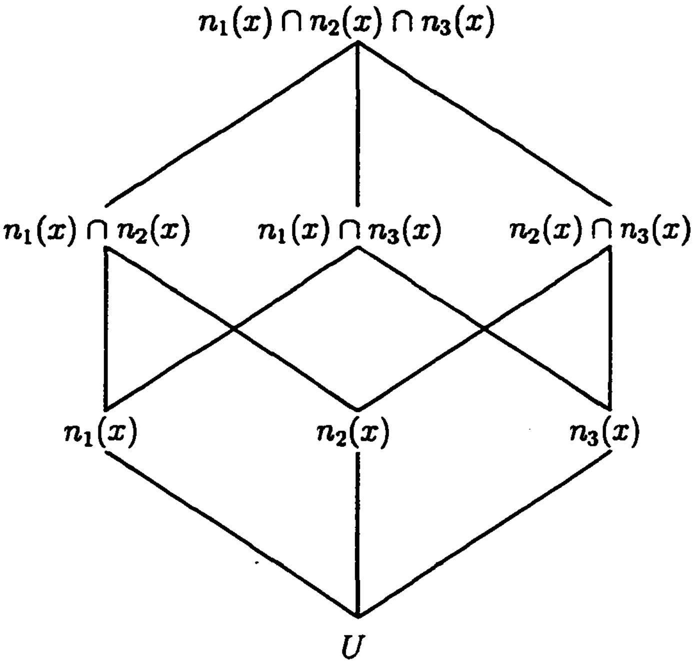


Figure 2.2: Diagram of ∩-closure for $NS(x) = \{n_1(x), n_2(x), n_3(x)\}$

which is demonstrated in Figure 2.2. ■

For a neighborhood system including more neighborhoods, the corresponding ∩-closure can be similarly created by extending the procedure shown above.

2.6 Rankings Induced by ∩-Closure

Let L be a set and $\preceq$ a binary relation on L. The relation $\preceq$ is called partial ordering if $\preceq$ is reflexive, antisymmetric and transitive. If $\preceq$ is defined to be a partial order on L, the pair $(L, \preceq)$ is called a partially ordered set or poset. Let $(L, \preceq)$ be a poset and let $A \subseteq L$. An element x in L is called an upper bound of A if for all $a \in A$, $a \preceq x$. Similarly, an element $x \in L$ is called a lower bound of A if for all $a \in A$, $x \preceq a$. An element $x \in L$ is called the least upper bound (LUB) of A, if x is an upper bound of A, and for any upper bound y of A, $x \preceq y$. Likewise, x is called the greatest lower bound (GLB) of A if x is a lower bound of A, and for any lower bound y of A, $y \preceq x$. A lattice is a partially ordered set $(L, \preceq)$ in which every pair of elements $a, b \in L$ has a GLB and a LUB. GLB and LUB of $\{a, b\}$ are denoted by $a \otimes b$ and $a \oplus b$, i.e.,

$$\begin{aligned} GLB(a,b) &= a \otimes b, \\ LUB(a,b) &= a \oplus b. \end{aligned} \tag{2.14}$$

The operator $\otimes$ is called "meet" and $\oplus$ is called "join". Using these symbols, a lattice is denoted by $(L, \otimes, \oplus)$. If there only exists the meet operation in the poset, then it is called a meet semilattice, denoted by $(L, \otimes)$. Similarly, If there only exists the join operation in the poset, then it is called a join semilattice, denoted by $(L, \oplus)$.

The $\cap$-closure $NS^*(x)$ is meet semilattice on which the defined partial ordering relation is $\subseteq$, and the closed meet operation is $\cap$, which is denoted by $(NS^*(x), \cap)$.

Lemma 2.1 *Given the $\cap$-closure $NS^*(x)$ of a neighborhood system $NS(x)$. Suppose $n_1(x), n_2(x) \in NS^*(x)$. If $n_1(x) \neq n_2(x)$, then*

$$\begin{aligned} n_1(x) \cap \left(n_2(x) - \bigcup_{n(x) \subset n_2(x)} n(x)\right) &= \emptyset, \\ n_2(x) \cap \left(n_1(x) - \bigcup_{n(x) \subset n_1(x)} n(x)\right) &= \emptyset. \end{aligned}$$

Proof. Let $\delta(x) = n_1(x) \cap \left(n_2(x) - \bigcup_{n(x) \subset n_2(x)} n(x)\right)$. One has

$$\begin{aligned} \delta(x) &= n_1(x) \cap n_2(x) - n_1(x) \cap \left(\bigcup_{n(x) \subset n_2(x)} n(x)\right) \\ &= n_1(x) \cap n_2(x) - \bigcup_{n(x) \subset n_2(x)} n_1(x) \cap n(x). \end{aligned}$$

Let $n_\lambda(x) = n_1(x) \cap n_2(x)$. One has $n_\lambda(x) \in NS^*(x)$. From $n_1(x) \neq n_2(x)$, it follows $n_\lambda(x) \subset n_1(x)$ and $n_\lambda(x) \subset n_2(x)$. Hence,

$$\begin{aligned}
\delta(x) &= n_\lambda(x) - (n_1(x) \cap n_\lambda(x)) \cup \left(\bigcup_{\substack{n(x) \subset n_2(x) \\ n(x) \neq n_\lambda(x)}} n_1(x) \cap n(x) \right) \\
&= n_\lambda(x) - n_\lambda(x) \cup \left(\bigcup_{\substack{n(x) \subsetneq n_2(x) \\ n(x) \neq n_\lambda(x)}} n_1(x) \cap n(x) \right) \\
&= \emptyset.
\end{aligned}$$

Similarly, $n_2(x) \cap \left(n_1(x) - \bigcup_{n(x) \subset n_1(x)} n(x)\right) = \emptyset$, can be proved. ■

Definition 2.4 *Given the $\cap$-closure $NS^*(x)$ of a neighborhood system $NS(x)$, a binary relation $\prec$ is defined as follows:*

$$\begin{aligned}
a \prec b \iff & \textit{there exists a pair of neighborhoods } n_1(x), n_2(x) \in NS^*(x) \\
& \textit{such that } n_1(x) \subset n_2(x), \; a \in n_1(x), \textit{ and } \; b \in n_2(x) - \bigcup_{n(x) \subset n_2(x)} n(x).
\end{aligned}$$

For two elements $a, b \in U$ with $a \prec b$, a is said to be nearer to x than b.

Theorem 2.3 *The binary relation $\prec$ defined on $NS^*(x)$ is asymmetric and transitive.*

Proof. **(i)** $\prec$ is asymmetric: Suppose $a \prec b$. There must exist two neighborhoods $n_1(x), n_2(x) \in NS^*(x)$ such that $n_1(x) \subset n_2(x)$, $a \in n_1(x)$ and $b \in n_2(x) - \bigcup_{n(x) \subset n_2(x)} n(x)$. Assume that $b \prec a$ also holds. Thus there also exist two subsets $n_3(x), n_4(x) \in NS^*(x)$ such that $n_3(x) \subset n_4(x)$, $b \in n_3(x)$ and $a \in n_4(x) -$

$\bigcup_{n(x)\subset n_4(x)} n(x)$. Hence,

$$a \in n_1(x) \cap \left(n_4(x) - \bigcup_{n(x)\subset n_4(x)} n(x) \right),$$

$$b \in n_3(x) \cap \left(n_2(x) - \bigcup_{n(x)\subset n_2(x)} n(x) \right).$$

To make a and b belong to nonempty subsets, it is necessary that $n_1(x) = n_4(x)$, $n_2(x) = n_3(x)$ hold by following Lemma 2.1. Thus, from $n_1(x) \subset n_2(x)$, one has $n_4(x) \subset n_3(x)$, contradiction appears. Similarly, from $n_3(x) \subset n_4(x)$, one has $n_2(x) \subset n_1(x)$, contradiction appears again. Therefore, if $a \prec b$, then $\neg(b \prec a)$. This means $\prec$ is asymmetric.

(ii) $\prec$ is transitive: Suppose $a \prec b$ and $b \prec c$, there must exist subsets $n_1(x), n_2(x), n_3(x), n_4(x) \in NS^*(x)$ such that

$$n_1(x) \subset n_2(x), \quad a \in n_1(x), \quad b \in n_2(x) - \bigcup_{n(x)\subset n_2(x)} n(x),$$

$$n_3(x) \subset n_4(x), \quad b \in n_3(x), \quad c \in n_4(x) - \bigcup_{n(x)\subset n_4(x)} n(x).$$

Hence, one has

$$b \in n_3(x) \cap \left(n_2(x) - \bigcup_{n(x)\subset n_2(x)} n(x) \right).$$

To make b belong to nonempty subset, is is necessary that $n_2(x) = n_3(x)$ hold by following Lemma 2.1. Thus one has

$$n_1(x) \subset n_4(x), \quad a \in n_1(x), \quad c \in n_4(x) - \bigcup_{n(x)\subset n_4(x)} n(x).$$

By following the Definition 2.4, one has $a \prec c$. This means $\prec$ is transitive. ■

From $\prec$, two more relations are introduced. The first is the reverse of $\prec$, denoted by $\succ$. By $a \succ b$, it means that a is farther away from x than b. The second relation is $\sim$, which describes the nearness of two elements a and b with respect to a given element x. The relation $\sim$ can be defined by

$$\begin{aligned} a \sim b &\iff \neg(a \prec b) \text{ and } \neg(b \prec a) \\ &\iff \neg(a \succ b) \text{ and } \neg(b \succ a). \end{aligned}$$

It suggests that a and b are incomparable. With order relation $\prec$, one may organize or rank elements in the universe U according to their nearness to x.

Example 6 Suppose $R \subseteq U \times U$ is a binary relation on the universe. The successor neighborhood of an element $x \in U$ is defined by (Yao, 1996):

$$n_R(x) = \{y \mid xRy\}. \tag{2.15}$$

In this case, the neighborhood system of each element has only one neighborhood, namely, $\mathrm{NS}_R(x) = \{n_R(x)\}$. The $\cap$-closure of the neighborhood system is $\{n_R(x)\}$ if $n_R(x) = U$, otherwise it is $\{n_R(x), U\}$. In the former case, every element of U is as near to x as any other element. In the latter case, one may say that element in $n_R(x)$ is closer to x than elements in $U - n_R(x)$. One can obtain a ranked list by using the order relation $\prec$ as follows:

$$n_R(x) \prec U - n_R(x).$$

∎

Example 7 For a neighborhood system containing two neighborhoods, the $\cap$-closure is shown in Figure 2.1. It is easy to obtain the following ranked list indicating their nearness to x:

$$n_1(x) \cap n_2(x) \prec \begin{array}{c} n_1(x) - n_1(x) \cap n_2(x) \\ n_2(x) - n_1(x) \cap n_2(x) \end{array} \prec U - n_1(x) \cup n_2(x).$$

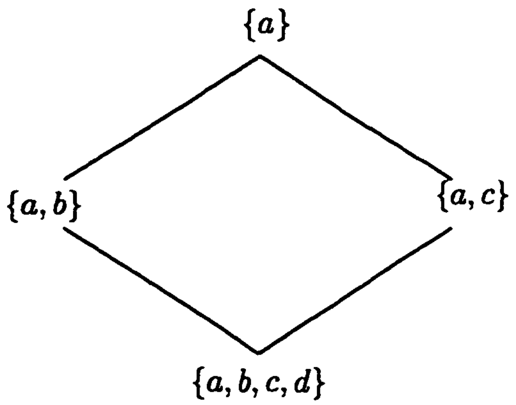


Figure 2.3: $\cap$-closure for $NS(a) = \{\{a,b\},\{a,c\}\}$

Consider a universe $U = \{a,b,c,d\}$. Suppose a neighborhood system of a is given by $NS(a) = \{\{a,b\},\{a,c\}\}$. The $\cap$-closure is $NS^*(a) = \{\{a\},\{a,b\},\{a,c\},\{a,b,c,d\}\}$, as shown in Figure 2.3. One may obtain the following ranked list:

$$\{a\} \prec \begin{matrix}\{b\}\\ \{c\}\end{matrix} \prec \{d\},$$

which is depicted in Figure 2.4. ■

Example 8 Consider a neighborhood system containing three neighborhoods. The $\cap$-closure structure is given by Figure 2.2. The following equations show the details

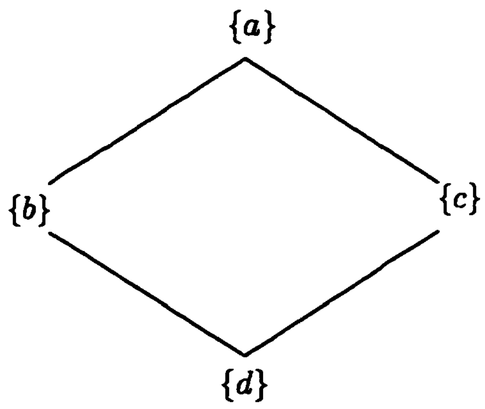


Figure 2.4: Ranked List for $NS(a) = \{\{a,b\},\{a,c\}\}$

of how the ranked list are formulated:

$$n_1(x)\cap n_2(x)\cap n_3(x) \;\prec\; \left\{\begin{array}{l} n_1(x)\cap n_2(x)-n_3(x) \prec \left\{\begin{array}{l} n_1(x)-n_2(x)\cup n_3(x)\\ n_2(x)-n_1(x)\cup n_3(x)\end{array}\right. \\ n_1(x)\cap n_3(x)-n_2(x) \prec \left\{\begin{array}{l} n_1(x)-n_2(x)\cup n_3(x)\\ n_3(x)-n_1(x)\cup n_2(x)\end{array}\right. \\ n_2(x)\cap n_3(x)-n_1(x) \prec \left\{\begin{array}{l} n_2(x)-n_1(x)\cup n_3(x)\\ n_3(x)-n_1(x)\cup n_2(x)\end{array}\right. \end{array}\right.$$

$$\prec\; U - n_1(x)\cup n_2(x)\cup n_3(x).$$

Given a universe $U = \{a,b,c,d,e,f,g,h\}$. Suppose a neighborhood system of a is $NS(a) = \{\{a,b,d,e\},\{a,b,c,g\},\{a,c,d,f\}\}$. The $\cap$-closure is:

$$\begin{aligned} NS^*(a) \;=\; & \{\{a\},\{a,b\},\{a,d\},\{a,c\},\{a,b,d,e\},\{a,b,c,g\},\\ & \{a,d,c,f\},\{a,b,c,d,e,f,g,h\}\}, \end{aligned}$$

which is shown as in Figure 2.5. One has:

$$\{a\} \prec \left\{ \begin{array}{l} \{b\} \prec \left\{ \begin{array}{l} \{e\} \\ \{g\} \end{array} \right. \\ \{c\} \prec \left\{ \begin{array}{l} \{f\} \\ \{e\} \end{array} \right. \\ \{d\} \prec \left\{ \begin{array}{l} \{g\} \\ \{f\} \end{array} \right. \end{array} \right. \prec \{h\},$$

which is depicted as in Figure 2.6. There is no doubt that a is the nearest element. After $\{a\}$ is selected, $\{b\}, \{c\}, \{d\}$ may be selected. Elements $\{e\}$ or $\{g\}$ may be selected after $\{b\}$. By this procedure, one define a structure on elements of U, as shown in Figure 2.6. ■

In general, one may obtain the ranked list of any neighborhood system containing finite neighborhoods. The $\cap$-closure $NS^*(x)$ is first constructed, a ranked list is then generated from the $\cap$-closure. The ranked list provides users with a series of choices. The immediate concern is that how end users could make a decision from a variety of possible selections. Generally speaking, there are no intrinsic rules to govern the user's choice. Each user may select a result at his own discretion. In the structure of a ranked list, the selection starts from the top element. One selection is accomplished by taking a path from top to bottom along the edges in the structure. At any point where there exist multiple edges, either of them may be taken. This point will be further discussed in Chapter 4.

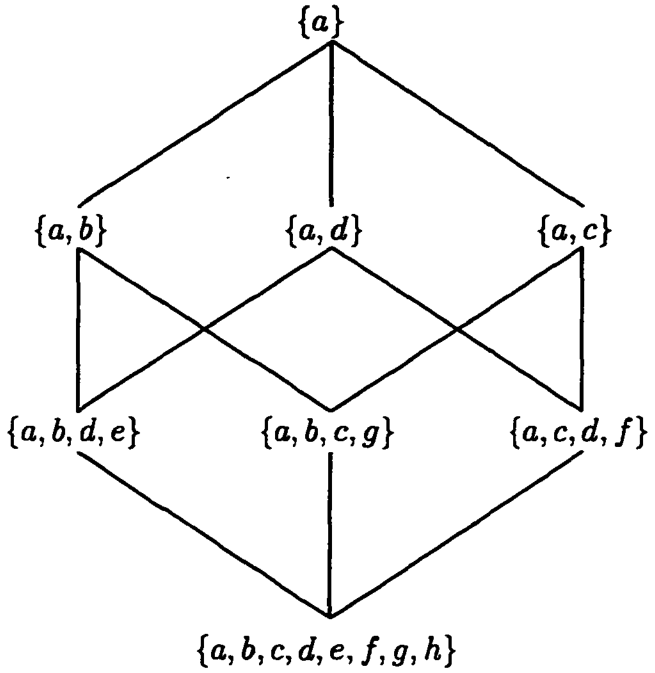


Figure 2.5: $\cap$-closure for $NS(x) = \{\{a, b, d, e\}, \{a, b, c, g\}, \{a, c, d, f\}\}$

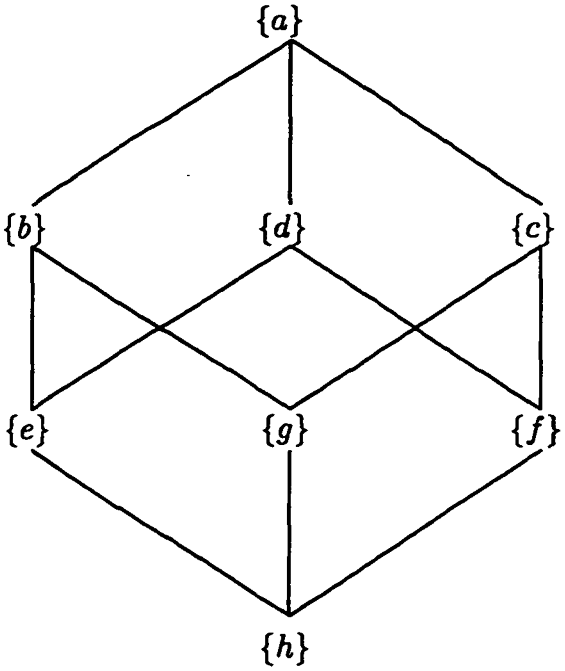


Figure 2.6: Ranked List for $NS(x) = \{\{a, b, d, e\}, \{a, b, c, g\}, \{a, c, d, f\}\}$

Chapter 3

INFORMATION SYSTEMS

A generalized information system is presented in this chapter by combining Pawlak information systems (Pawlak, 1981; Marek and Pawlak, 1973, 1976, 1981; Marek and Rode-Babezenko, 1975; Marek and Traczyk, 1977) and neighborhood systems (Lin, 1988, 1996). An equality binary relation is used in Pawlak information systems. By using an arbitrary binary relation that induces a single neighborhood, relation-based information systems are reviewed (Yao and Wong, 1995). The adoption of neighborhood systems in information systems presents a more general framework, called neighborhood-based information systems.

3.1 Pawlak Information Systems

A basic component of a Pawlak information system is a finite and nonempty set of objects O called the universe. A universe can be anything that is of interest, e.g., people, automobiles, or any entities. The objects in the universe O are characterized by a finite and nonempty set of attributes AT. Every element $a \in AT$ is associated with a finite and nonempty set of values V_a called the domain of attribute a.

Following Pawlak (1981), Marek and Pawlak (1973, 1976, 1981), Marek and Rode-Babezenko (1975), Marek and Traczyk (1977), Lipski (1979, 1981), Orlowska and

Pawlak (1984), Vakarelov (1991), and Yao and Noroozi (1994), a Pawlak information system is defined as:

$$PT = (O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}), \tag{3.16}$$

where

O is a finite and nonempty set of objects, i.e., universe,

AT is a finite and nonempty set of attributes,

V_a is a finite and nonempty set of values for each $a \in AT$,

$f_a : O \times AT \longrightarrow V_a$ is an information function for each a ∈ AT.

An information function f_a associates each object and attribute with a value in the domain of that attribute.

Example 9 An information system can be conveniently represented by a table. Table 3.1 is an example of an information system. In this table, a group of people $O = \{o_1, o_2, \ldots, o_{12}\}$ are described by three attributes $AT = \{$AGE, GENDER, OPINION$\}$. The domain of AGE is a closed interval $[20, 30]$, the domain of GENDER is $V_{\text{GENDER}} = \{male, female\}$, and the domain of OPINION is:

$$\begin{aligned} V_{\text{OPINION}} &= \{\textit{positive}, \textit{negative}, \textit{medium}, \\ &\quad \textit{slightly positive}, \textit{slightly negative}, \\ &\quad \textit{considerably positive}, \textit{considerably negative}, \\ &\quad \textit{highly positive}, \textit{highly negative} \\ &\quad \textit{extremely positive}, \textit{extremely negative}\}. \end{aligned}$$

For object o_1, information functions produce the following description:

$$f_{\text{AGE}}(o_1) = 25, \quad f_{\text{GENDER}}(o_1) = male, \quad f_{\text{OPINION}}(o_1) = medium.$$

	AGE	GENDER	OPINION
o_1	25	*male*	*medium*
o_2	27	*male*	*negative*
o_3	22	*female*	*positive*
o_4	30	*male*	*extremely negative*
o_5	23	*female*	*highly positive*
o_6	20	*female*	*extremely positive*
o_7	27	*male*	*negative*
o_8	24	*female*	*slightly positive*
o_9	21	*female*	*highly positive*
o_{10}	26	*male*	*slightly negative*
o_{11}	23	*male*	*positive*
o_{12}	30	*male*	*highly negative*

Table 3.1: Example of An Information System

This sample information system may be used to represent the opinions of people on a specific issue. ■

The notion of information system provides a conventional tool for the representation of objects in terms of their attribute values. A database system may be considered as an information system. Therefore, the results developed in this study can be readily extended to database management systems.

With respect to an attribute $a \in A$, if a relation E_a is defined by: for o, o',

$$oE_ao' \Longleftrightarrow f_a(o) = f_a(o'). \tag{3.17}$$

the relation E_a is an equivalence relation. Reflexivity, symmetry and transitivity of E_a follow trivially from those of the equality relation =. With the relation E_a, two objects are considered to be indiscernible with respect to attribute a, if and only if they have the same value on attribute a.

The equivalence relation E_a partitions the universal set O into disjoint subsets called equivalence classes. Elements in the same equivalence class are said to be indiscernible. The family of all equivalence classes is called a quotient set, denoted by O/E_a. The equivalence class containing object o is denoted by $[o]_{E_a}$.

Example 10 In the information system given by Table 3.1, the partition of the universe with respect to attribute OPINION, is $\{\{o_1\}, \{o_2, o_7\}, \{o_3, o_{11}\}, \{o_4\}, \{o_5, o_9\}, \{o_6\}, \{o_8\}, \{o_{10}\}, \{o_{12}\}\}$. For instance, objects of o_2 and o_7 are equivalent, i.e., $o_2 E_a o_7$, because $f_{\text{OPINION}}(o_2) = f_{\text{OPINION}}(o_7) = \textit{negative}$. Similarly, with respect to attribute GENDER, the partition is $\{\{o_1, o_2, o_4, o_7, o_{10}, o_{11}, o_{12}\}, \{o_3, o_5, o_6, o_8, o_9\}\}$. ■

The definition of equivalence relation from single attribute can be extended into an arbitrary subset $A \subseteq AT$ as follows:

$$oE_A o' \iff (\forall a \in A) f_a(o) = f_a(o'). \tag{3.18}$$

The relation E_A is an equivalence relation (Orlowska, 1985). The indiscernibility of o and o' is determined in terms of all attributes in the subset A. This means that two objects are indiscernible with respect to a subset $A \subseteq AT$ if and only if they have the same values for all elements in A.

Example 11 In the information system given by Table 3.1, consider a subset $A = \{\text{GENDER, OPINION}\}$. The corresponding partition is $\{\{o_1\}, \{o_2, o_7\}, \{o_3\}, \{o_{11}\}, \{o_4\}, \{o_5, o_9\}, \{o_6\}, \{o_8\}, \{o_{10}\}, \{o_{12}\}\}$. For instance, objects of o_5 and o_9 are indiscernible, i.e., $o_5 E_A o_9$, because one has both $f_{\text{GENDER}}(o_5) = f_{\text{GENDER}}(o_9) = \textit{female}$ and $f_{\text{OPINION}}(o_5) = f_{\text{OPINION}}(o_9) = \textit{highly positive}$. ■

3.2 Relation-based Information Systems

It is straightforward to generalize the Pawlak information system by adopting an arbitrary binary relation on attribute values instead of using the trivial equality relation = (Orlowska 1985; Wasilewska 1989; Yao and Wong 1995). A relation-based information system is defined by introducing binary relations on the values of attributes as follows:

$$RT = (O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}, \{R_a \mid a \in AT\}), \tag{3.19}$$

where R_a is a binary relation on V_a for each $a \in AT$. Binary relations provide additional information on relationships between attribute values. This adds a new component to a Pawlak information system.

With respect to an attribute $a \in AT$, consider a binary relation $R_a \subseteq V_a \times V_a$. For any value $x \in V_a$, one may define the following subset of V_a:

$$R_a(x) = \{y \mid xR_a y\}. \tag{3.20}$$

The subset $R_a(x)$ may be interpreted as a successor neighborhood of the element x relation R_a (Yao, 1997). Thus, a relation R_a defines a special neighborhood system, in which each element of V_a has only one neighborhood.

With respect to attribute a, an object o is $\Re_a$-related to another object o' if their values on attribute a are R_a-related, namely:

$$o\Re_a o' \iff f_a(o)R_a f_a(o'). \tag{3.21}$$

For any object $o \in O$, one may define a subset of o using relation R_a:

$$R_a(o) = \{o' \mid o\Re_a o'\} = \{o' \mid f_a(o)R_a f_a(o')\}, \tag{3.22}$$

which may be interpreted to be a neighborhood of object o. If the relation R_a is chosen to be an equality relation on attribute a, then $R_a(x)$ in Equation (3.20) becomes a singleton subset, i.e., $R_a(x) = \{x\}$, which leads to a Pawlak information system.

The binary relation R_a on attribute $a \in A$ can be extended to a subset $A \subseteq AT$ as follows:

$$o\Re_A o' \iff (\forall a \in A) o\Re_a o',$$

which means that o and o' are $\Re_A$-related, if and only if they are R_a-related for all $a \in A$. For a subset $A \subseteq AT$, the class of the object o may be obtained based on those attributes $a \in A$, that is:

$$\begin{aligned} R_A(o) &= \{o' \mid o\Re_A o'\} \\ &= \bigcap_{a \in A} \{o' \mid o\Re_a o'\} \\ &= \bigcap_{a \in A} R_a(o), \end{aligned} \tag{3.23}$$

which means that each member in the class of the object o belongs to every class of attributes $a \in A$.

Example 12 In the information system given by Table 3.1, consider the attribute a = AGE. Assume there is a binary relation on attribute AGE such that two objects are R_{AGE}-related if their age difference is less than or equal to one. Thereby, for the object o_8, there is $f_{\text{AGE}}(o_8) = 24$, one can have $R_{\text{AGE}}(24) = \{24, 23, 22\}$, thus $R_{\text{AGE}}(o_8) = \{o_3, o_5, o_8, o_{11}\}$. Assume there is a binary relation R_{OPINION} on the attribute OPINION such that object o is R_{OPINION}-related to o' if the opinion of o' is mostly near to that of o. For $f_{\text{OPINION}}(o_8) = \{\textit{slightly positive}\}$, from $R_{\text{OPINION}}(\textit{slightly positive}) = \{\textit{slightly positive}, \textit{positive}, \textit{medium}\}$, one obtains $R_{\text{OPINION}}(o_8) = \{o_1, o_3, o_8, o_{11}\}$. With respect to subset A = {AGE, OPINION}, the subset class of o_8 is:

$$\begin{aligned} R_A(o_8) &= R_{\text{AGE}}(o_8) \cap R_{\text{OPINION}}(o_8) \\ &= \{o_3, o_8, o_{11}\}. \end{aligned}$$

■

3.3 Neighborhood-Based Information Systems

In the previous sections, two kinds of information system are introduced, Pawlak information system and relation-based information system. They have been used in many applications (Chu and Chen, 1994; Wasilewska, 1987; Yao and Wong,1995). This section will extend relation-based information systems to neighborhood-based information systems.

With respect to an attribute $a \in AT$, one may construct a neighborhood system for each $x \in V_a$:

$$NS_a(x) = \{n_{a,i}(x) \subseteq V_a \mid i \in I, x \in V_a\}. \tag{3.24}$$

For example, in the information system given by Table 3.1, a neighborhood system for AGE $= 24$ could be:

$$NS_{\text{AGE}}(24) = \{\{24, 23\}, \{24, 25\}\}. \tag{3.25}$$

For the attribute OPINION, a neighborhood system of *slightly positive* may be given by

$$\begin{aligned} NS_{\text{OPINION}}(\textit{slightly positive}) = \{&\{\textit{slightly positive}, \textit{medium}\}, \\ &\{\textit{slightly positive},\ \textit{positive}\}\}. \end{aligned} \tag{3.26}$$

As pointed out in Chapter 2, when modeling nearness of a neighborhood system, it is assumed that every neighborhood of an element contains that element. In other words, only reflexive neighborhood system will be considered.

By using neighborhood systems $NS(V_a)$ defined on attributes, a more generalized notion of information system is defined by Yao and Chen (1997):

$$NT = (O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}, \{NS(V_a) \mid a \in AT\}), \tag{3.27}$$

where,

$NS : V_a \longrightarrow 2^{2^{V_a}}$ is a neighborhood operator on attribute $a \in AT$.

In this study, NT is referred to as neighborhood-based information system, which is a natural generalization of relation-based information systems.

Given an arbitrary object $o \in O$, its value on an attribute $a \in AT$ is $f_a(o)$. From the value $f_a(o)$ and the neighborhood system

$$NS_a(f_a(o)) = \{n_{a,i}(f_a(o)) \subseteq V_a \mid i \in I\}, \tag{3.28}$$

one is able to define a neighborhood system of o as follows:

$$NS_a(o) = \{n_{a,i}(o) \subseteq O \mid i \in I\},$$

where

$$n_{a,i}(o) = \{o' \mid f_a(o') \in n_{a,i}(f_a(o))\}. \tag{3.29}$$

That means that the object o' is in a neighborhood of o if and only if its attribute value $f_a(o')$ is in the corresponding neighborhood of $f_a(o)$, namely:

$$o' \in n_{a,i}(o) \iff f_a(o') \in n_{a,i}(f_a(o)). \tag{3.30}$$

By using different attributes, various neighborhood systems can be obtained. With respect to a subset of attributes, one is required to combine different neighborhood systems through logical connectives. Let $NS_a(o), NS_b(o), NS_c(o)$ be three neighborhood systems of o with respect to attributes $a, b, c \in AT$, respectively. One may further formulate neighborhood systems of o in the following ways. For attributes a and b, there are two ways to define neighborhood systems of o:

$$\begin{aligned} NS_{a \wedge b}(o) &= NS_a(o) \sqcap NS_b(o), \\ NS_{a \vee b}(o) &= NS_a(o) \sqcup NS_b(o). \end{aligned} \tag{3.31}$$

For three attributes $a, b, c \in AT$, one may obtain combinations such as:

$$\begin{aligned} NS_{a \wedge b \wedge c}(o) &= NS_a(o) \sqcap NS_b(o) \sqcap NS_c(o), \\ NS_{a \vee b \vee c}(o) &= NS_a(o) \sqcup NS_b(o) \sqcup NS_c(o), \\ NS_{a \wedge b \vee c}(o) &= NS_a(o) \sqcap NS_b(o) \sqcup NS_c(o). \end{aligned} \tag{3.32}$$

For a subset $A \subseteq AT$ of attributes, one is able to obtain more combinations. The use of conjunction produces:

$$NS_A^{\wedge}(o) = \sqcap_{a \in A} NS_a(o). \tag{3.33}$$

Such a method is used in many applications of rough set theory (Pawlak, 1984, 1991). Alternatively, one may formulate the neighborhood systems of o as:

$$NS_A^{\vee}(o) = \sqcup_{a \in A} NS_a(o). \tag{3.34}$$

In general, one can use various combinations through operations of $\wedge$ and $\vee$.

Example 13 Consider the attribute AGE and OPINION in the information system given by Table 3.1. By using neighborhood systems on attributes defined as in equations (3.25) one can easily construct neighborhood systems, for $o_8 \in O$:

$$\begin{aligned} NS_{\text{AGE}}(o_8) &= \{\{o_8, o_5, o_{11}\}, \{o_8, o_1\}\}, \\ NS_{\text{OPINION}}(o_8) &= \{\{o_8, o_1\}, \{o_3, o_{11}, o_8\}\}. \end{aligned}$$

By combining two attributes, one obtains:

$$\begin{aligned} NS_{\text{AGE}\wedge\text{OPINION}}(o_8) &= NS_{\text{AGE}}(o_8) \sqcap NS_{OPINION}(o_8) \\ &= \{\{o_8\}, \{o_8, o_1\}, \{o_8, o_{11}\}\}, \\ NS_{\text{AGE}\vee\text{OPINION}}(o_8) &= NS_{\text{AGE}}(o_8) \sqcup NS_{\text{OPINION}}(o_8) \\ &= \{\{o_1, o_8, o_5, o_{11}\}, \{o_5, o_3, o_8, o_{11}\}, \{o_1, o_8\}, \{o_1, o_3, o_8, o_{11}\}\}. \end{aligned}$$

■

Pawlak and relation-based information systems are special cases of neighborhood-based information systems. The former uses equivalence neighborhood, the latter uses only one neighborhood for each element. The use of many neighborhoods leads to graded approximation, rather than two-level approximations.

Chapter 4

A NEIGHBORHOOD-BASED RETRIEVAL MODEL

In this chapter, a retrieval model using neighborhood-based information systems will be presented. This model makes full use of the concept of neighborhood systems developed in Chapter 2 and the notion of neighborhood-based information systems proposed in Chapter 3. Retrieval in both Pawlak and relation-based information systems is first discussed. Retrieval in a neighborhood-based information system is based on the notion of neighborhood system operators and $\cap-$closure. These will be used to analyze the retrieved results in order to provide users with useful information.

4.1 An Exact Retrieval Model

The retrieval operation is the process of recovering information from an information system in response to requests from users. According to a query submitted by the user, a retrieval system selects objects that satisfy the query.

Definition 4.1 *An atomic query is an attribute and attribute value pair connected by the equality sign, i.e., "attribute_name = attribute_value".*

Definition 4.2 *A query language is defined recursively as follows:*

Rule 1 : *Every atomic query is a query;*

Rule 2 : *If q_1 and q_2 are two queries, $(q_1 \wedge q_2)$ and $(q_1 \vee q_2)$ are queries.*

For simplicity, conditions expressed by relations $\leq$ and $\geq$, and the negation operation ($\neg$) are not considered in this study. It is assumed that $\wedge$ has higher priority during the evaluation of a query.

Example 14 Consider the information system given by Table 3.1. Examples of atomic queries are:

$$q_1 : \text{AGE} = 24,$$

$$q_2 : \text{OPINION} = \textit{slightly positive},$$

and examples of queries are:

$$q_3 : (\text{AGE} = 24) \wedge (\text{OPINION} = \textit{slightly positive}),$$

$$q_4 : (\text{AGE} = 24) \vee (\text{OPINION} = \textit{slightly positive}).$$

■

A query q may be considered as a condition expressed in terms of attribute values. For an atomic query q:

$$q : a = v,$$

all objects that satisfy $f_a(o) = v$, will be fetched in the process of retrieval. The retrieved set can be computed by:

$$R(q) = \{o \mid f_a(o) = v\}. \tag{4.35}$$

That is, the universe O is divided into two classes. The subset $R(q) \subseteq O$ consists of all objects satisfying the condition expressed by q, while $O - R(q)$ consists of those objects that do not satisfy q.

Suppose there are two queries q_1 and q_2, whose retrieved subsets are $R(q_1)$ and $R(q_2)$, respectively. For query $q_1 \wedge q_2$, the following rule is used:

$$o \text{ satisfies } q_1 \wedge q_2 \Longleftrightarrow o \text{ satisfies } q_1 \;\text{ and }\; o \text{ satisfies } q_2.$$

The retrieved subset can be therefore computed by:

$$\begin{aligned} R(q_1 \wedge q_2) &= \{o \mid o \in R(q_1) \;\text{ and }\; o \in R(q_2)\} \\ &= R(q_1) \cap R(q_2). \end{aligned} \tag{4.36}$$

The query $q_1 \vee q_2$ is evaluated by using the rule:

$$o \text{ satisfies } q_1 \vee q_2 \Longleftrightarrow o \text{ satisfies } q_1 \;\text{ or }\; o \text{ satisfies } q_2.$$

Similarly, the retrieved set is computed by:

$$\begin{aligned} R(q_1 \vee q_2) &= \{o \mid o \in R(q_1) \;\text{ or }\; o \in R(q_2)\} \\ &= R(q_1) \cup R(q_2). \end{aligned} \tag{4.37}$$

Through applying the above evaluation rules recursively, one can obtain the retrieved set for any query. For example, the retrieved set for a query $q_1 \wedge (q_2 \vee q_3)$ can be obtained by

$$R(q_1 \wedge (q_2 \vee q_3)) = R(q_1) \cap (R(q_2) \cup R(q_3)).$$

Example 15 In the information system given by in Table 3.1, consider a query q_1 : AGE $= 23$. From equation (4.35), one obtains $R(q_1) = \{o_5, o_{11}\}$. Similarly, for query q_2 : OPINION $=$ *positive*, one obtains $R(q_2) = \{o_3, o_{11}\}$. According to equations (4.36) and (4.37), for queries $q_1 \wedge q_2$ and $q_1 \vee q_2$, one has

$$\begin{aligned} R(q_1 \wedge q_2) &= \{o_{10}\} \\ R(q_1 \vee q_2) &= \{o_3, o_5, o_{10}\}. \end{aligned}$$

■

One can see that retrieval in Pawlak information system is based on exact matching. This strategy is in fact used in database management systems.

4.2 A Relation-Based Retrieval Model

Retrieval operation in Pawlak information systems uses the trivial equality relation to select data satisfying a query. Such exact matching may provide limited answers, or even no information at all if the exact answer is not available (Chu and Chen, 1994). This is a main disadvantage of retrieval in Pawlak information systems. To remedy such a problem, relation-based retrieval methods may be used in the framework of relation-based information systems.

Suppose there is an atomic query $q: \quad a \; = \; v$ for an attribute $a \in AT$. Let $R_a \subseteq V_a \times V_a$ be a binary relation on V_a. The element v is associated with a subset $R_a(v) = \{v' \in V_a \mid vR_a v'\}$. Using all elements of $R_a(v)$, query q may be relaxed into q', denoted by $q \leadsto q'$, as follows:

$$q' : \bigvee_{v' \in R_a(v)} a = v'.$$

According to equation (4.37), the retrieved result of relaxed query q' can be computed by

$$R(q') = \bigcup_{v' \in R_a(v)} \{o \mid f_a(o) = v'\}. \tag{4.38}$$

Query relaxation using relation-based information systems is similar to that of neighborhood query answering (Chu and Chen, 1992, 1994), in which a query is relaxed and refined within one neighborhood through a type abstract hierarchy.

Given two queries q_1 and q_2, they are relaxed into q_1' and q_2', respectively, namely:

$$q_1 \leadsto q_1',$$

$$q_2 \leadsto q_2'.$$

Consequently, one has the retrieval subsets $R(q_1')$ and $R(q_2')$. Queries $q_1 \wedge q_2$ and $q_1 \vee q_2$ are relaxed into $q_1' \wedge q_2'$ and $q_1' \vee q_2'$:

$$\begin{aligned} q_1 \wedge q_2 &\leadsto q_1' \wedge q_2', \\ q_1 \vee q_2 &\leadsto q_1' \vee q_2', \end{aligned}$$

which produce the following two retrieved sets:

$$\begin{aligned} R(q_1' \wedge q_2') &= R(q_1') \cap R(q_2'), \\ R(q_1' \vee q_2') &= R(q_1') \cup R(q_2'). \end{aligned} \tag{4.39}$$

Example 16 Consider the information system given in Table 3.1 again. On attributes AGE and OPINION, suppose the same binary relations as in Example 12 are defined. For a query q_1 : AGE $= 24$, produces a class of subset $\{23, 24, 25\}$ for 24, thus one has $q_1 \leadsto q_1'$, where q_1' : $\bigvee_{v' \in \{23,24,25\}}$ AGE $= v'$. The retrieved subset is $R(q_1') = \{o_1, o_5, o_8, o_{11}\}$. For a query q_2 : OPINION $=$ *slightly positive*, similarly, the subset class for *slightly positive* becomes $\{$*positive*, *medium*, *slightly positive*$\}$. The relaxed query is q_2' : $\bigvee_{w' \in \{positive, medium, slightly\ positive\}}$ OPINION $= w'$. It produces the retrieved set $R(q_2') = \{o_1, o_3, o_8, o_{11}\}$. The retrieved subsets of queries $q_1 \wedge q_2$ and $q_1 \vee q_2$ are

$$\begin{aligned} R(q_1' \wedge q_2') &= \{o_1, o_8, o_{11}\}, \\ R(q_1' \vee q_2') &= \{o_1, o_3, o_5, o_8, o_{11}\}. \end{aligned}$$

■

In the relation-based retrieval model, two-level retrieval results are obtained. One is the results from exact matching using the original query q, i.e., $R(q)$. The other is the results from a relaxed query q', i.e., $R(q')$. If reflexive relations are used, we have $R(q) \subset R(q')$. The set $R(q)$ consists of the objects satisfying the query, while $R(q')$ consists of objects which are related to elements in $R(q)$.

4.3 A Neighborhood-Based Retrieval Model

The retrieval method in Pawlak information system uses the exact retrieving mechanism, and the retrieval method in relation-based information systems uses one neighborhood for query relaxation. In many situations, users may expect more informative results. For example, with respect to a certain user-submitted query, the relation-based information system would probably disappoint a user by yielding a result containing too many objects. It becomes difficult to tell which is more important, and which is less. At this moment, users may expect the result to possess certain kinds of preference, which would help them judge the results. Pawlak information systems may probably retrieve a void set for some kinds of queries because they employ the trivial equality relation. At this moment, user might want to know a result that is near the query. To guide users in making more useful selections, in this study a more generalized retrieval model is proposed in the framework of neighborhood-based information systems, $NT = \{O, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}, \{NS(V_a) \mid a \in AT\}\}$.

Consider an atomic query q:

$$q : a = v, \quad where \;\; a \in AT, \;\; v \in V_a, \tag{4.40}$$

Suppose the element v is associated with a neighborhood system:

$$NS_a(v) = \{n_{a,i}(v) \subseteq V_a \mid i \in I\}. \tag{4.41}$$

Following the concept of query relaxation in relation-based information systems, each neighborhood, $n_{a,i}(v)$, corresponds to a relaxed query

$$q'_i : \bigvee_{v' \in n_{a,i}(x)} a = v', \quad where \quad i \in I.$$

Collectively, q is relaxed into a family of queries, $q \rightsquigarrow q''$:

$$q'' = \{q'_i \mid i \in I\}.$$

Clearly, the method of relation-based retrieval which relaxes a query using one neighborhood is a special case.

With respect to a relaxed query q'_i, its retrieved subset is denoted by $R(q'_i)$. Following the relation-based retrieval method, $R(q'_i)$ can be computed by:

$$R(q'_i) = \bigcup_{v' \in n_{a,i}(v)} \{o \mid f_a(o) = v'\}.$$

Let $R(q'_0)$ represent the subset of objects retrieved by the query q, that is

$$R(q'_0) = \{o \mid f_a(o) = v\}.$$

As a result, the retrieved family of subsets for $q \rightsquigarrow q''$ is:

$$R(q'') = \{R(q'_0), R(q'_1), \ldots, R(q'_m)\}, \tag{4.42}$$

where m is the cardinality of $NS_a(v)$. The family $R(q'')$ may be interpreted as a neighborhood system of an element that satisfies the query q. One may compute the $\cap$-closure of $R(q'')$ based upon the method developed in Chapter 2. The results give a ranked list of objects according to their nearness to the element satisfying the original query q.

Let q_1 and q_2 be two queries. They may be relaxed into:

$$q_1 \rightsquigarrow q''_1,$$

$$q_2 \rightsquigarrow q''_2,$$

respectively. The corresponding retrieved families of subsets are:

$$\begin{aligned} R(q_1'') &= \{R(q_{1,0}'), R(q_{1,1}'), \ldots, R(q_{1,k}')\}, \\ R(q_2'') &= \{R(q_{2,0}'), R(q_{2,1}'), \ldots, R(q_{2,l}')\}, \end{aligned}$$

For queries $q_1 \wedge q_2$ and $q_1 \vee q_2$, they are relaxed into:

$$\begin{aligned} q_1 \wedge q_2 &\rightsquigarrow q_1'' \wedge q_2'', \\ q_1 \vee q_2 &\rightsquigarrow q_1'' \vee q_2''. \end{aligned}$$

By using the operations of neighborhood systems, the retrieved families for $q_1'' \wedge q_2''$ and $q_1'' \vee q_2''$ can be computed by:

$$\begin{aligned} R(q_1'' \wedge q_2'') &= R(q_1'') \sqcap R(q_2'') \\ &= \{R(q_{1,i}') \cap R(q_{2,j}') \mid 0 \leq i \leq k, 0 \leq j \leq l\}, \end{aligned} \tag{4.43}$$

$$\begin{aligned} R(q_1'' \vee q_2'') &= R(q_1'') \sqcup R(q_2'') \\ &= \{R(q_{1,i}') \cup R(q_{2,j}') \mid 0 \leq i \leq k, 0 \leq j \leq l\}. \end{aligned} \tag{4.44}$$

4.4 An Example

An example is presented to illustrate the ideas developed in this chapter. In this example, the information system given by Table 3.1 is used. Consider a query:

$$q_a : \ \text{AGE} = 24. \tag{4.45}$$

With respect to a neighborhood system defined by:

$$\begin{aligned} NS_a(24) &= \{n_{a,1}(24), n_{a,2}(24)\}, \\ &= \{\{24, 23\}, \{24, 25\}\}, \end{aligned} \tag{4.46}$$

q_a is relaxed into a family of two queries, $q_a \rightsquigarrow q''_a$,

$$q''_a = \{q'_{a,1}, q'_{a,2}\}, \tag{4.47}$$

where

$$\begin{aligned} q'_{a,1}: &\quad (\text{AGE} = 24) \ \vee \ (\text{AGE} = 23), \\ q'_{a,2}: &\quad (\text{AGE} = 24) \ \vee \ (\text{AGE} = 25). \end{aligned} \tag{4.48}$$

Consequently, one has the retrieved sets:

$$\begin{aligned} R(q'_{a,1}) &= \{o \mid f_{\text{AGE}}(o) = 24\} \cup \{o \mid f_{\text{AGE}}(o) = 23\} \\ &= \{o_8\} \cup \{o_5, o_{11}\} \\ &= \{o_5, o_8, o_{11}\}, \\ R(q'_{a,2}) &= \{o \mid f_{\text{AGE}}(o) = 24\} \cup \{o \mid f_{\text{AGE}}(o) = 25\} \\ &= \{o_8\} \cup \{o_1\} \\ &= \{o_1, o_8\}. \end{aligned}$$

The retrieved subset $R(q'_{a,0})$ of the original query q_a is:

$$\begin{aligned} R(q'_{a,0}) &= \{o \mid f_{\text{AGE}}(o) = 24\}, \\ &= \{o_8\}. \end{aligned}$$

The retrieved subsets for the relaxed queries q''_a are:

$$\begin{aligned} R(q''_a) &= \{R(q'_{a,0}), R(q'_{a,1}), R(q'_{a,2})\}, \\ &= \{\{o_8\}, \{o_5, o_8, o_{11}\}, \{o_1, o_8\}\}. \end{aligned} \tag{4.49}$$

Now, consider further a query:

$$q_b: \ \text{OPINION} \ = \ \textit{slightly positive}. \tag{4.50}$$

With respect to a neighborhood system: [1]

$$\begin{aligned} NS_{\text{OPINION}}(s.pos) &= \{n_{b,1}(s.pos), n_{b,2}(s.pos)\}, \\ &= \{\{s.pos, med\}, \{s.pos, pos\}\}, \end{aligned}$$

q_b is relaxed into a family of queries, $q_b \rightsquigarrow q_b''$:

$$q_b'' = \{q_{b,1}', q_{b,2}'\}, \tag{4.51}$$

where

$$\begin{aligned} q_{b,1}' : &\quad (\text{OPINION} = s.pos) \vee (\text{OPINION} = med), \\ q_{b,2}' : &\quad (\text{OPINION} = s.pos) \vee (\text{OPINION} = pos). \end{aligned} \tag{4.52}$$

The corresponding retrieved subsets are:

$$\begin{aligned} R(q_{b,1}') &= \{o \mid f_{\text{OPINION}}(o) = s.pos\} \cup \{o \mid f_{\text{OPINION}}(o) = med\} \\ &= \{o_8\} \cup \{o_1\} \\ &= \{o_1, o_8\}, \\ R(q_{b,2}') &= \{o \mid f_{\text{OPINION}}(o) = s.pos\} \cup \{o \mid f_{\text{OPINION}}(o) = pos\} \\ &= \{o_8\} \cup \{o_3, o_{11}\} \\ &= \{o_3, o_8, o_{11}\}. \end{aligned}$$

The retrieved subset $R(q_{b,0}')$ of the original query q_b is

$$\begin{aligned} R(q_{b,0}') &= \{o \mid f_{\text{OPINION}}(o) = s.pos\}, \\ &= \{o_8\}. \end{aligned}$$

[1] e.neg: extremely negative; h.neg: highly negative; neg: negative; med: medium; e.pos: extremely positive; h.pos: highly positive; pos: positive.

The retrieved subsets for the relaxed queries q_b'' are:

$$\begin{aligned} R(q_b'') &= \{R(q_{b,0}'), R(q_{b,1}'), R(q_{b,2}')\}, \\ &= \{\{o_8\}, \{o_1, o_8\}, \{o_3, o_8, o_{11}\}\}. \qquad (4.53) \end{aligned}$$

More insights may be obtained by examining the retrieval process. Suppose q_1 is a query defined by

$$q_1 = q_a \wedge q_b. \qquad (4.54)$$

One may relax this query as, $q_1 \leadsto q_1''$, i.e.,

$$q_1 = q_a \wedge q_b \quad \leadsto \quad q_1'' = q_a'' \wedge q_b''.$$

According to equation (4.43), we further obtain:

$$\begin{aligned} R(q_1'') &= R(q_a'' \wedge q_b'') \\ &= R(q_a'') \sqcap R(q_b'') \\ &= \{R(q_{a,i}') \cap R(q_{b,j}') \mid 0 \leq 2 \leq k, 0 \leq j \leq 2\}, \\ &= \{\{o_8\}, \{o_8, o_{11}\}, \{o_1, o_8\}\}. \qquad (4.55) \end{aligned}$$

Equation (4.55) represents a neighborhood system of element o_8. The nearness of an object to $\{o_8\}$ can be interpreted as the nearness to the original query q_1. The $\sqcap$−closure of this neighborhood system is depicted as in Figure 4.1. Consequently, one obtains the ranking structure of the $\sqcap$−closure as depicted in Figure 4.2. By analyzing the ranked list in Figure 4.2, the retrieved results are expressed in the following preferences:

$$\{o_8\} \prec \left\{ \begin{array}{c} \{o_1\} \\ \{o_{11}\} \end{array} \right\} \prec \{o_2, o_3, o_4, o_5, o_6, o_7, o_9, o_{10}, o_{12}\}.$$

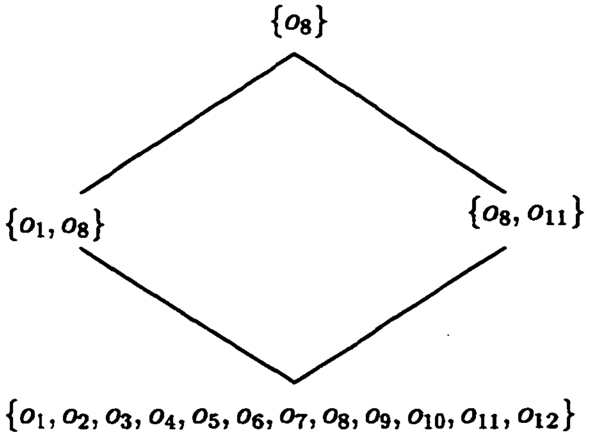

$\{o_8\}$
$\{o_1, o_8\}$
$\{o_8, o_{11}\}$
$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8, o_9, o_{10}, o_{11}, o_{12}\}$

Figure 4.1: $\cap$−closure of Retrieval Results of Query $q_a \wedge q_b \rightsquigarrow q_a'' \wedge q_b''$

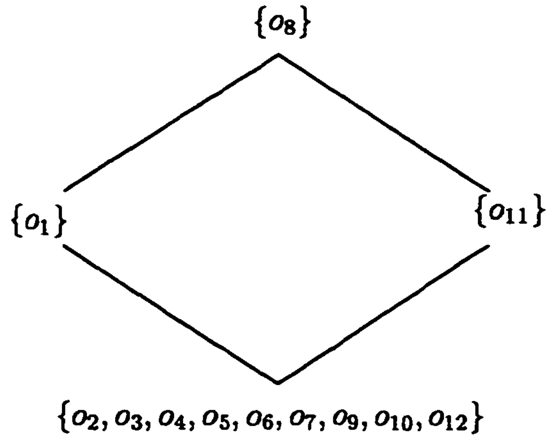


Figure 4.2: Ranked List of Query $q_a \wedge q_b \rightsquigarrow q_a'' \wedge q_b''$

Note that the nearest subset to query q_1 is $\{o_8\}$, which exactly satisfies the query q_1. That is $f_{\text{AGE}}(o_8) = 24$ and $f_{\text{OPINION}}(o_8) = s.pos$, and they satisfy q_a : AGE $= 24$ and q_b : OPINION $= s.pos$. The next nearest object subset to q_1 is either $\{o_1\}$ or $\{o_{11}\}$, which do not exactly satisfy the query q_1 but nearly do so. With respect to attribute AGE, $f_{\text{AGE}}(o_1) = 25$ and $f_{\text{AGE}}(o_{11}) = 23$, so the object o_1 is a bit older than that requested by the query AGE $= 24$, and o_{11} is a bit younger. With respect to attribute OPINION, $f_{\text{OPINION}}(o_1) = med$ and $f_{\text{OPINION}}(o_{11}) = pos$, object o_1 is a bit more negative than requested by the query OPINION $= s.pos$, and o_{11} is a bit more positive. The rest of objects $\{o_2, o_3, o_4, o_5, o_6, o_7, o_9, o_{10}, o_{12}\}$ are ranked last, which are further away from q_1.

Consider the results of query:

$$q_2 = q_a \vee q_b. \tag{4.56}$$

It can be relaxed into $q_2 \rightsquigarrow q_2''$, i.e.,

$$q_2 = q_a \vee q_b \quad \rightsquigarrow \quad q_2'' = q_a'' \vee q_b''.$$

Following equation (4.44), one obtains:

$$\begin{aligned} R(q_2'') &= R(q_a'' \vee q_b'') \\ &= R(q_a'') \sqcup R(q_b'') \\ &= \{R(q_{a,i}') \cup R(q_{b,j}') \mid 0 \le 2 \le k, 0 \le j \le 2\}, \\ &= \{\{o_8\}, \{o_1, o_8\}, \{o_3, o_8, o_{11}\}, \\ &\quad \{\{o_5, o_8, o_{11}\}, \{o_1, o_5, o_8, o_{11}\}, \\ &\quad \{\{o_3, o_5, o_8, o_{11}\}, \{o_1, o_3, o_8, o_{11}\}. \end{aligned} \tag{4.57}$$

Its $\sqcap$−closure and ranked list are given in Figure 4.3 and Figure 4.4, respectively. It is straightforward to generate the preferences as follows:

$$\{o_8\} \prec \left\{ \begin{array}{c} \{o_1\} \\ \{o_{11}\} \prec \left\{ \begin{array}{c} \{o_3\} \\ \{o_5\} \end{array} \right. \end{array} \right\} \prec \{o_2, o_4, o_6, o_7, o_9, o_{10}, o_{12}\}.$$

Object o_8 satisfies both queries q_a : AGE $=$ 24 and q_b : OPINION $=$ *s.pos*, and thus satisfies $q_2 = q_a \vee q_b$. Either object o_1 or object o_{11} follows o_8. The selection between o_5 and o_3 depends on user preference. From the sequences:

$$23 \nearrow 24,$$

$$s.pos \nearrow pos \nearrow h.pos,$$

one may intuitively say that with respect to the query $q_2 = q_a \vee q_b$, the object o_5 may be selected. From the sequences:

$$22 \nearrow 23 \nearrow 24,$$

$$s.pos \nearrow pos,$$

with respect to the query $q_2 = q_a \vee q_b$, one may alternatively say that object o_3 should be selected. Objects $\{o_2, o_4, o_6, o_7, o_9, o_{10}, o_{12}\}$ are ranked last.

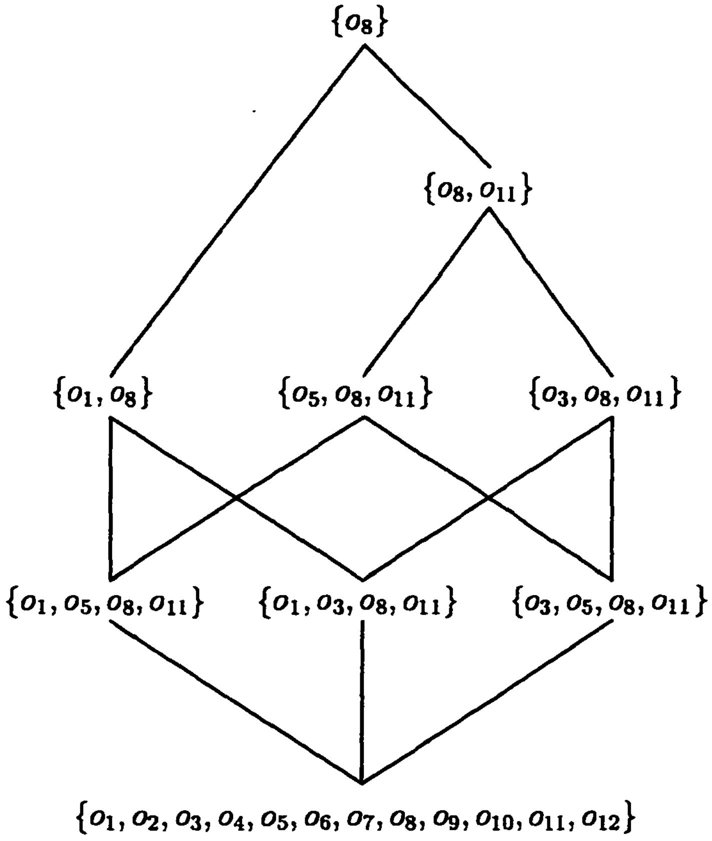

$\{o_8\}$
$\{o_8, o_{11}\}$
$\{o_1, o_8\}$
$\{o_5, o_8, o_{11}\}$
$\{o_3, o_8, o_{11}\}$
$\{o_1, o_5, o_8, o_{11}\}$
$\{o_1, o_3, o_8, o_{11}\}$
$\{o_3, o_5, o_8, o_{11}\}$
$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8, o_9, o_{10}, o_{11}, o_{12}\}$

Figure 4.3: $\cap$−closure of Retrieval Results of Query $q_a \vee q_b \leadsto q''_a \vee q''_b$

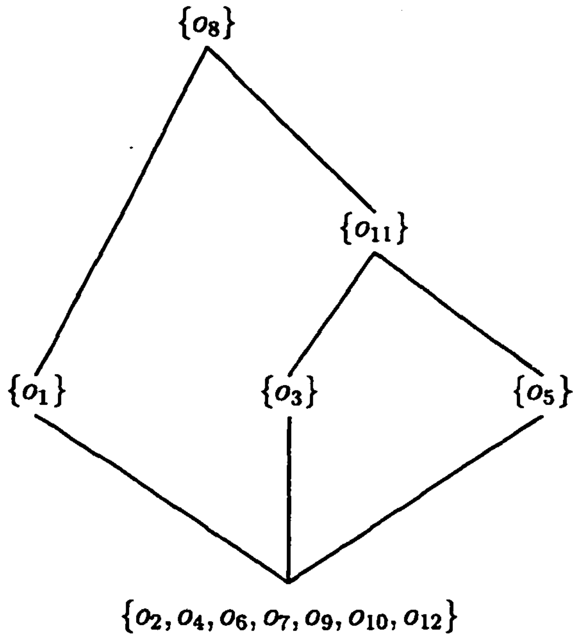


Figure 4.4: Ranked List for Query $q_a \vee q_b \rightsquigarrow q_a'' \vee q_b''$

Chapter 5

CONCLUSION

In this thesis, a more generalized framework of information systems is proposed based on the notion of neighborhood systems. The corresponding retrieval model is developed within the proposed framework. This model is of great interest when conducting retrieval process in databases that do not contain enough information matching the original queries.

The concept of neighborhood systems is a useful and effective tool for representing semantics information. It is a generalization of Pawlak information system. This thesis started from neighborhood systems on domains of attributes in an information system. A Frechet(V) space that contains a family of neighborhood systems is introduced on attribute value. Consequently, more general relationships on objects can be induced through the well-developed notion of neighborhood systems. The use of binary relation on attribute values is a special case such that there is only one neighborhood in a neighborhood system. Neighborhood-based information systems provide the basic environment to conduct the approximate retrieval. The retrieval algorithm has a solid theoretical basis. The proposed retrieval model presents a structured retrieval results, from which a user can search different level of approximations.

The applications of neighborhood systems can be extended into many areas related to information sciences, such as inference, information approximation, machine learning, data mining and data analysis. This thesis is an initial step of application of neighborhood systems. The future work should be on the aspect of developing special types of neighborhood systems for special applications in information sciences.

REFERENCES

Alefeld, G and Herzberger, J. 1983. *Introduction to interval computations.* Academic Press, New York.

Brink, C. 1993. *Power Structures.* Algebra Universalis, **30**, 177-216.

Chu, W. W. and Chen, Q. 1992. *Neighborhood Associative Query Answering.* Journal of Intelligent Information Systems, **1**, 355-382.

Chu, W. W. and Chen, Q. 1994. *A Structured Approach for Cooperative Query Answering.* IEEE Transactions on Knowledge and Data Engineering, **6**, 738-749.

Cuppens, F. and Demolombe, R. 1989. *Cooperative answering: A methodology to provide intelligent access to databases.* In: Proceedings of the 2nd International Conference of Expert Database Systems, 621-643.

Lin, T.Y. 1988. *Neighborhood Systems and Approximation in Relational Databases and Knowledge Bases.* In: Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems.

Lin, T.Y. 1996. *Neighborhood Systems - Application to Qualitative Fuzzy and Rough Sets.* In: Advances in Machine Intelligence and Soft Computing, P. P. Wang, Ed., Department of Electrical Engineering, Durham, North Carolina.

Lipski, Witold 1979. *On Semantic Issues Connected with Incomplete Information Databases.* ACM Transactions on Database Systems, **4**, 262-296.

Lipski, Witold 1981. *On Databases with Incomplete Information.* Journal of the Association for Computing Machinery, **28**, 41-70.

Marek, W. and Pawlak, Z. 1973. *Mathematical Foundations of information storage and retrieval.* Parts 1, 2, 3, CC PAS Reports, 135, 136, 137, Warszawa.

Marek, W. and Pawlak, Z. 1976. *Information storage and retrieval systems.* Theoretical Computational Sciences, 331-354.

Marek, W. and Pawlak, Z. 1981. *Rough Sets and information systems.* ICS PAS Report, 441. Warszawa.

Marek, W. and Rode-Babezenko, I. 1975. *A decompositiom of information systems.* ICS PAS Report, 212. Warszawa.

Marek, W. and Traczyk, T. 1977. *Stochastic information system.* Fundamenta Informatica, **1**, 121-130.

Moore, R. E. 1966. *Interval Analysis.* Englewood Cliffs, New Jersey: Prentice-Hall.

Motro, A. 1990. *FLEX: A tolerant and cooperative user interface to databases.* IEEE Transactions on Knowledge and Data Engeneering, **2**, 231-246.

Orlowska, E. 1985. *Logic of indiscernibility relations.* Lectures Notes in Computer Science, **208**, Springer-Verlag, Berlin, 177-186.

Orlowska, E. 1986. *Semantic Analysis of Inductive Learning.* Theoretical Computer Science, **43**, 81-89.

Orlowska, E. and Pawlak, Z. 1984. *Logical foundations of knowledge representation.* ICS PAS Reports, **537**, 1-108.

Pawlak, Z. 1981. *Information systems - theoretical foundations.* Information Systems, **6**, 205-218.

Pawlak, Z. 1982. *Rough Sets.* International Journal of computer and information science, **11**, No.5.

Pawlak, Z. 1984. *Rough Classification.* International Journal of Man-Machine Studies, **20**, 469-483.

Salton, G and McGill, M. H. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

Sierpinski, W. and Krieger, C. C. 1956. *General Topology.* University of Toronto Press, Toronto.

Slowinski, R. and Vanderpooten, D. 1995. *Similarity relation as a basis for rough approximations.* ICS Research Report 53/95, Institute of Computer Science, Warsaw, Poland.

Vakarelov, D. 1991. *A modal logic for similarity relations in Pawlak knowledge representation systems.* Fundamenta Informaticae, **XV**, 61-79.

Wasilewska, A. 1989. *Conditional Knowledge Representing Systems - Model for an Implementation.* Bulletin of the Polish Academy of Sciences, Mathematics, **37**, 63-69.

Whitesitt, J. E. 1962. *Boolean Algebra and its Applications.* Addidon-wesley Publishing Company, London, England.

Wong, S. K. M. and Yao, Y. Y. 1990. *A Probabilistic Inference Model for Information Retrieval.* Journal of American Society of Information Sciences, **41**, 324-329.

Yao, Y. Y. 1993. *Interval-Set Algebra for Qualitative Knowledge Representation.* In: IEEE, Proceedings of the 5th International Conference on Computing and Information, Sudbury, Ontario, 370-374.

Yao, Y. Y. 1996. *Two Views of the Theory of Rough Sets in Finite Universes.* International Journal of Approximation Reasoning, **15**, 291-317.

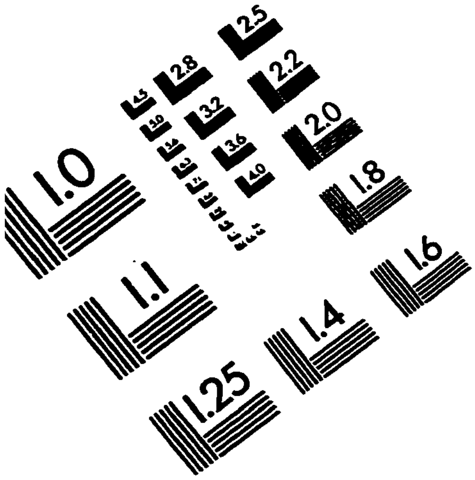
Yao, Y. and Noroozi, N. 1994. *A unified model for Set-based Computations.* Proceedings of RSSC'94.

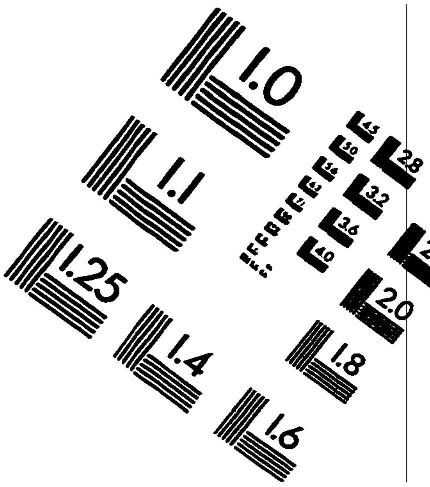
Yao, Y. Y. and Chen, C. X. 1997. *Neighborhood Based Information Systems.* Proceedings of Joint Conference of Information Sciences, North Carolina, Vol. 3, Rough Set & Computer Science, 154-157.

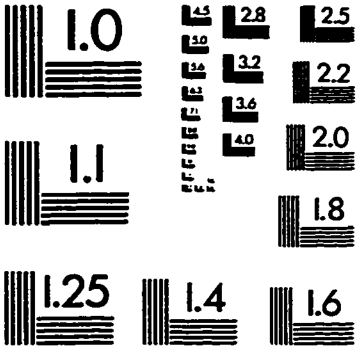
Yao, Y. Y. and Wong S. K. M. 1995. *Generalization of Rough Sets Using Relationships Between Attribute Values.* Proceedings of the 2nd Annual Joint Conference on Information Sciences, 30-33.

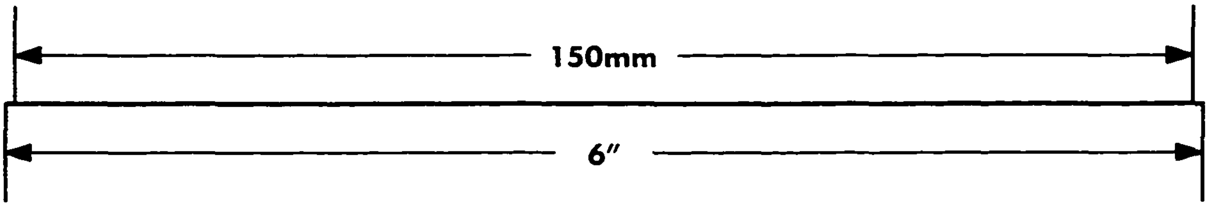
Zakowski, W. 1983. *Approximations in the space* (U, Π). Demonstratio Mathematica, **XVI**, 761-769.

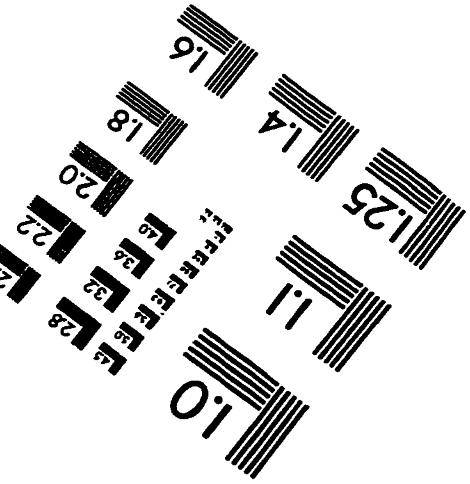
IMAGE EVALUATION TEST TARGET (QA-3)











APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989