VISUAL DATA MINING OF TIME SERIES AGRICULTURAL DATA

BY

HUMAIRA IKHLAQUE

(318- FAS/MSCS/F06)

Supervised by:

PROF DR. AHSAN ABDULLAH

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University Islamabad
2008

16/7/10

EN

Data mining
Data base searching
Visual analytics

MP

D.E
20-12-10

بسم الله الرحمن الرحيم

In The Name of

ALLAH ALMIGHTY

The Most Merciful, The Most Beneficent

A dissertation submitted to the
Department of Computer Science,
Faculty of Applied Sciences,
International Islamic University, Islamabad, Pakistan,
as a partial fulfillment of the requirements for the award of the degree of

MS Computer Science

To
The Holiest Man Ever Born,
Prophet Muhammad (صلى اللهعيه وسلم)
&
To
Our Parents and Families
We are most indebted to our parents and families, whose affection has always been the source of encouragement for us, and whose prayers have always been a key to our success.
&
To
Those Holy Seekers
Who give away their lives to make the stream of life flow
Smoothly and with Justice.
&
To
Our Honorable Teachers
Who have been a beacon of knowledge and a constant source of inspiration,
for our whole life span.

Declaration

We, hereby declare that "Visual Data Mining of Time Series Agricultural Data" neither as a whole nor as a part thereof has been copied out from any source. I have developed this project and the accompanied report entirely on the basis of my personal efforts made under the sincere guidance of my supervisors. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institution of learning.

Humaira Ikhlaque

318-FAS/MSCS/F06

Acknowledgement

All acclamation to Allah Almighty who has empowered and enabled me to accomplish the task successfully

First of all I would like to thank Allah Almighty who helped me out in every problem during the research. I would like to express my serious and humble gratitude to Almighty whose blessings, help and guidance has been a real source of all achievements in life. I would like to admit that the completion of my thesis is due to my loving parents.

I would like thank my Supervisor Dr Ahsan Abdullah for his sincere efforts to guide me throughout this project and for providing access to aggregate ADSS Macro data (www.agroict.org). I also wish to express my appreciation to my co-supervisor Sir Imran Saeed who helped me a lot and provided his full cooperation. I would like to acknowledge all team members of C@IR (Center for Agro-Informatics Research) FAST.

I would like to thank my friend for their cooperation and encouragement.

Humaira Ikhlaque

318-FAS/MSCS/F06

PROJECT IN BRIEF

Project Title	**Visual Data Mining Of Time Series Agriculture Data**
Undertaken By	Humaira Ikhlaque
Supervised By	Prof Dr. Ahsan Abdullah,C@IR FAST-NU
Co-Supervised By	Mr Imran Saeed
Starting Month	January 2008
Ending Month	May 2008
Software Used	C# (Visual Studio 2005) c s/h
Environment Used	MS Window XP Professional
System Used	Pentium IV

ABSTRACT

In this thesis, the primary focus is on Agro-Informatics, which is the development of Information Technology based solutions for collecting, managing and analyzing data produced by agriculture sector. Agriculture data exist in very complex form. It is difficult to analyze the data, thus the interpretation of data into the simpler form is required. The techniques used in this thesis are Query Dependent Spiral (Pixel Oriented Technique) and Shape Coding (Icon Based Technique). Both mention techniques are combined in a way to find the effects of pesticides on pest and predators. The effects are shown in the form of color coded spiral pixels with respect to time dimension. The effects of pest and predators are dependent on query results and time dimensions.

Department of Computer Science

International Islamic University, Islamabad

Date: 30th-08-08.

Final Approval

It is certified that we have read the project titled "Visual Data Mining of Time Series Agricultural Data" submitted by **Miss Humaira Ikhlaque Reg. No. 318-FAS/MSCS/F06**. It is our judgment that this project is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the degree MS in Computer Science.

COMMITTEE

PROFESSOR
Department of Computer Science
COMSATS Institute of Information Technoogy
ISLAMABAD

External Examiner:

Prof Dr. Maqbool uddin shaikh

Internal Examiner:

Tehmina Amjad
International Islamic University, Islamabad

Supervisors:

Prof Dr. Ahsan Abdullah
C@IR FAST-NU, Islamabad

Mr Imran Saeed
Internartional Islamic University, Islamabad

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Data mining is use to discern valuable relationships, samples in huge data using refined data investigation tools [1]. It allows user to view and critically analyze the data and their associations by finding correspondence between data from a huge set of data. Application of Information technology on agricultural data in a way that it results into some useful output is known as agro informatics.

Agriculture is the spinal column and main driving force for the economy of Pakistan and accounts for more than 24% of our GDP. [2]. Agriculture industry is very important to every person in a country, as food is the basic necessity of live. Policy makers and decision makers need modern IT tools for solving problem related to agriculture.

A number of visualization techniques e.g. Chern-off Faces have been proposed for data visualization in agro informatics but the major concern of this thesis will be the query dependent spiral and shape coding techniques that will be used to visualize the effects of pesticides on pest and predators.

Pesticides are toxic insecticides used to control species that have destructive impact on human activities. Human life is susceptible to adverse effects of agricultural pesticides used all over the world. Excessive use of pesticides may become hazardous not only to human life but also to the environment. While using pesticides, it is vitally important to get the job done carefully by both controlling pest populations and at the same time protecting the environment from potential adverse effects of pesticides use. Excessive use of pesticides results in water contagion problems and can also affect the quality of soil.

1.2 Objectives

1. To explore the useful clusters with respect to time dimension.
2. To suggest a solution or strategy if possible to overcome the side effect.

3. To Implement data mining techniques to provide insight into data to non-IT people

1.3 Organization of Study

The thesis has been organized in the following manner so as to give smooth effects to the reader. Chapter 2 introduces the background of the topic: what is data mining and visual data mining, why we use data visualization techniques. Chapter 3 describes the literature review. Chapter 4 describes the problem definition. Chapter 5 describes the design methodology and proposed solution Chapter 6 is the testing of the system and different test cases and their analysis. Results are also discussed in this chapter. Finally conclusion and future work is stated in Chapter 7.

CHAPTER 2

VISUAL DATA MINING

2.1 Definitions and Terminology

Some important terminologies related to data mining are stated as under:

Information

The data which is meaningful is called information. The trends, associations or relationships among all this *data* can provide *information.* [3]

Knowledge

Knowledge is reliable information that can be put to work in the service of all men, and which can be communicated in comprehensible ways so that people everywhere can become more self-reliant and self-sufficient.

Data mining (also known as Knowledge Discovery) technology helps businesses discover hidden data patterns and provides predictive information which can be applied to benefit the business. [4]

Data Warehouse

Data warehousing is the only viable solution for providing strategic information. It is an informational environmental that [5]

- provides an integrated and total view of the enterprise
- Makes the enterprise's current and historical information easily available for decision making
- Renders the organization's information consistent

2.2 Type of Relationships in Data Mining

Four types of relations in data mining are as follows

- **Classes**: Are store in predetermined groups. Stored data is used to locate data in predetermined groups. For example, a Bank could mine customer preferences who are taking loans for various projects. This information could be helpful in catching customers in future [6].

- **Clusters**: Similar data are organized in one group. And the dissimilar data is organized in other group. A cluster does not have predetermined groups. Clustering means forming groups [5]. Data items are grouped according to logical relationships or consumer preferences. For example, data mining can show which group of people did greatest shopping in month of Eid in Super Mall.

- **Associations**: Data can be mined to identify associations. Associations are affinities between items. Association discovery algorithm finds combinations where the presence of one item suggests the presences of another [5].
- **Sequential Patterns**: As the name implies, it discover patterns where one set of item follows another specific set [5]. Data is mined to anticipate behavior patterns and trends. For example the results obtained from the forecasting can predict is the rainfall predicted in the next 24 hour or not, what will be the humidity level in next 24 hours etc.

2.3 Stages of Data mining

The process of data mining consists of four stages:

- Extract the data from different database sources; transform the data into one format, for instance if in one database the format for male and female is 0 and 1. In another database the format is 'm' and 'f' then use only one standard format. Finally load data at one centralized location that is called data warehouse.
- Data is store in multi-relational database system.
- Perform analysis on the data.
- Present the data in a form which can easily be identified by user, such as a picture, image graph or table.

2.4 Data Visualization

Data visualization is use to represent raw data in the form of images, graphs, icons, and colors which facilitate a user in understanding the data in few seconds. Graphs and images give a comprehensive view of the data. Better the visualization effects, better is the data understanding. The design and implementation of visualization techniques should be such that it is easily understandable by human beings and gives more insight of the data.

CHAPTER 3

LITERATURE REVIEW

There are three types of data including uni-variate, bivairate and multivariate.

3.1 Classification of Data Visualization Techniques

3.1.1 Uni-variate data

As the name sounds uni-variate data only represents one variable at time. The standard method for uni-variate data are Histogram and Pie chart.

3.1.1.1 Histograms

Are use to show continuous data in equal intervals. The equal intervals are shown on x-axis while the y-axis shows the frequency.

3.1.1.2 Pie charts

In it the data is represented in from of circle, the circle is divided into segments called sectors of the circle. Each sector represents a distinct data by different color. The size of the sector varies depending upon the size of the data.

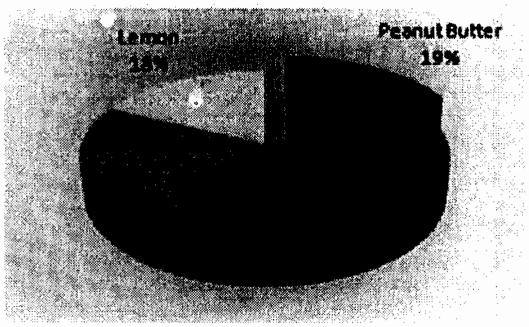


Figure3.1: Pie chart

Problem

The problem occurs when large amount f data comes, and it is difficult to identify data among so many sectors.

3.1.2 Bi-variate data

Bi-variate data consist of paired samples of only two quantitative variables. Two-dimensional data can be visualized in different ways. A very common visualization form is the scatter plot.

3.1.2.1 Scatter plot:

A number of variables are graphed to 2 to 3 axes. The simplest case is where each variable has its own axis. Color can be used to encode additional variables. Scatter plots can use additional visual coding. Given an m-dimensional matrix defining the data where 2 or 3 dimensions should be projected [7]. Scattered plot is displayed by using following data set.

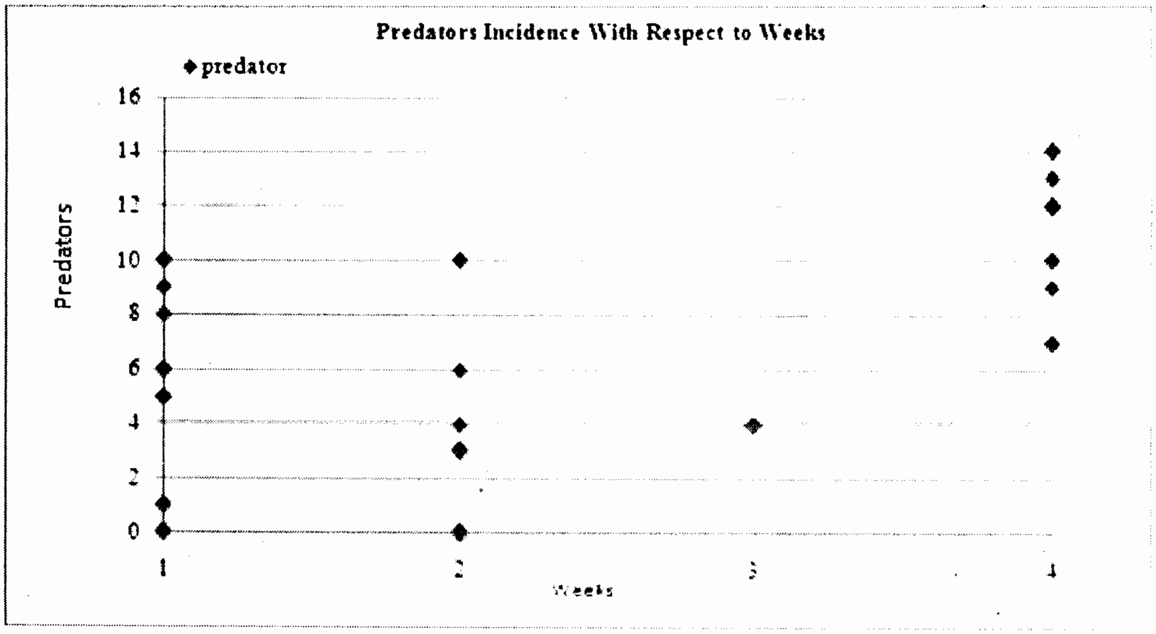


Figure3.2: Scattered Plot Showing Predator Incidence w.r.t Weeks for Month of May

Source of Data: C@IR FAST-NU, Islamabad

Problem

In case of two or more data sets being displayed in the same coordinate system different colors can be used to distinguish between the distinct plots. A problem with this way of displaying data arises when the amount of data points gets very high as the points become too dense.

3.1.2.2 Line graph

Another important visualization technique for two-dimensional data is the line graph. The line is drawn between the points on x-axis and y-axis.

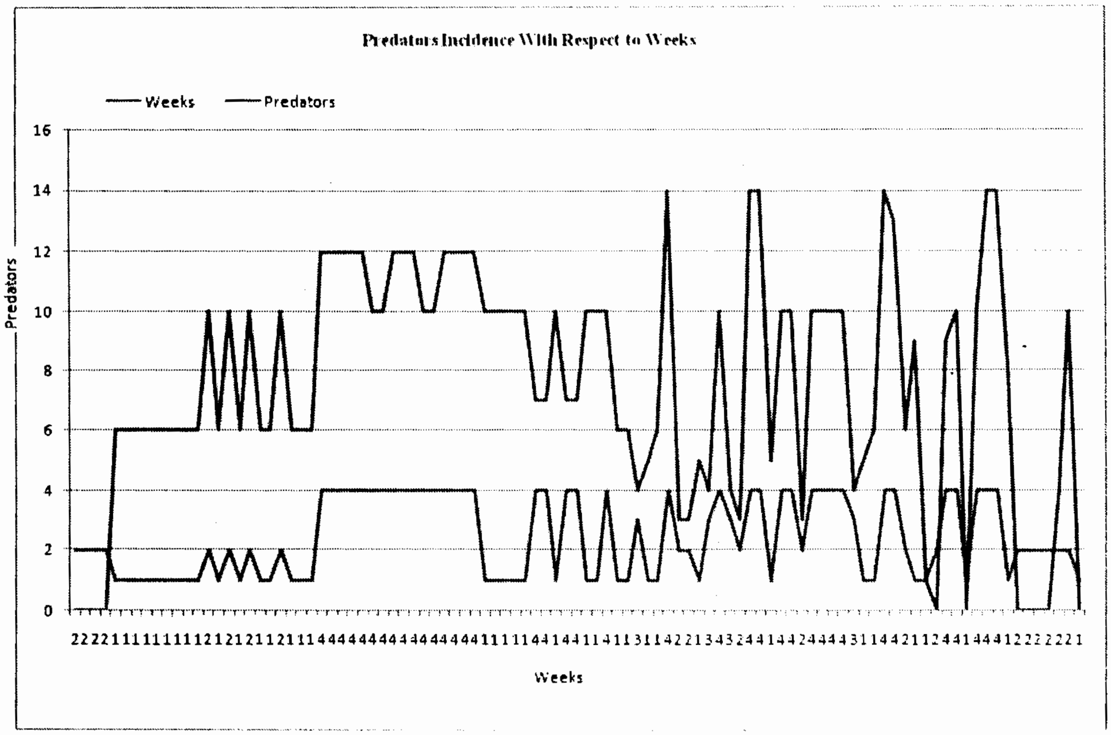


Figure3.3: Line Graph Showing Predator Incidence w.r.t Weeks for Month of May

Source oɾ Data: C@IR FAST-NU, Islamabad

Problem

Lines with different color are use to identify the points, problem arises when huge data set comes and it hides a lot of information

3.1.3 Multivariate Data

Since uni-variate can represent only one and bi-variate can represent only two variables. There are also some short comings of these two methods so multi-varaite is the option for representing multiple variables. Some important methods are icon based and pixel based visualization.

3.1.3.1 Icon-based methods

Icon-based methods are approaches that use icons (or glyphs) to represent high dimensional data. They map data components to graphical attributes. [8]. One of the well known methods is Chern-off Face. It uses different faces and their features to represent data. Each facial feature represents one variable. Figure 3.4 shows a sad face that represents the negative effect of pesticides on pests

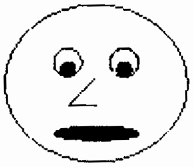


Figure3.4: Chern-off Face

Star glyphs:

The probably most common icon-based technique is the use of star glyphs to denote data points. A star glyph consists of a centre point with equally angled rays. These branches correspond to the different dimensions and the length of the limbs mark the value of this particular dimension for the studied data point. A polygon line connects the outer ends of the spokes [8].

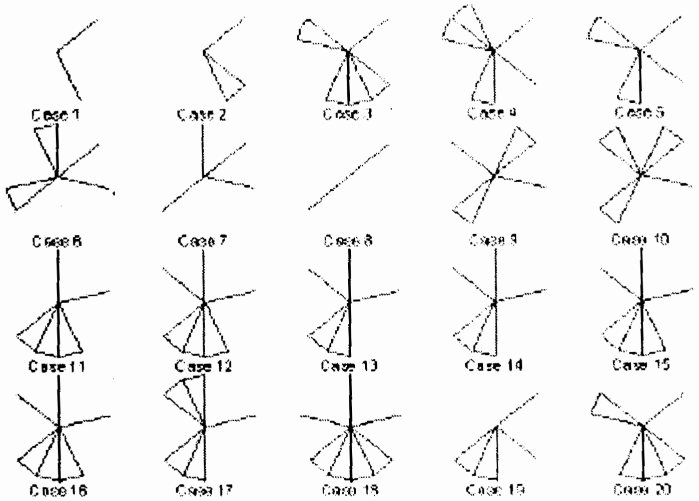


Figure3.5: The stars visualization of the animal's data

Analysis:

These icon-based techniques are very vivid but have several disadvantages. A very severe problem is the organization of the glyphs on the screen as no coordinate system representing two of the dimensions is provided. Even if we decided to use a Cartesian system it would put more weight on these two dimensions and so probably distort the data pattern. Another obstacle is the amount of variables and the size of the data set itself. If the number of rays become too high a distinction between the different spokes and the values they represent is not possible anymore. A similar unclear map emerges if the number of data points exceeds a certain amount [8].

3.2 The Approach and Technique Selected for Research

The technique that is selected for research is query dependent spiral and shape coding with respect to time.

3.3 The Reason for Selection

Following are the reasons for selecting query dependent spiral and shape coding technique for this thesis.

- Agriculture data exist in very complex form. To find effects of pesticides on human health, agricultural data is used. It is difficult to analyze the data, thus the interpretation of data into the simpler form is required. In these thesis results of generated query is shown in the form of color pixels.

- Shape coding developed a unique graphical method to encode and display multivariate data by using pixel icon in the form of array to represent individual records. In the query dependent spiral and shape coding technique multiple variable results are displayed in a single cell with respect to time dimension which is an innovative idea.
- Previously one cell of shape coding technique use to represent only one value of attribute in one cell.
- As opposed to other methods of multivariate data visualization e.g. Chern-off Face, where work is done by reducing the information in the data set to a relatively few variables. It eventually results in the loss of original information and hiding results and findings in the data.
- Query dependent spiral technique can represent thousand of records with different colors, thus avoiding the problem of overlapping and information hiding.
- User can have insight of each attribute selected in "selection predicate"

CHAPTER 4

Problem Statement

4.1 Problem Statement

Agriculture data exist in very complex form. To find effects of pesticides on pest and predators, agricultural data is used. It is difficult to analyze the data, thus the interpretation of data into the simpler form is required. Moreover, charts and tables are commonly used to visually analyze data. These graphics are simple and easy to understand, but charts show only highly aggregated data and present only a limited number of data values while tables often show too many data values. As a consequence, these graphics may either lose or obscure important information, so different techniques are required to monitor complex datasets. For the exploration of large volumes of multi-attribute data, the current charts and tables are not able to show important information such as: Data distribution of multiple attributes, Comparison of correlations and patterns, Instantaneous drilldown to transaction level information (e.g., price and quantity in an invoice).

In the developed query dependent spiral and shape coding technique each record is represented by color coded pixels in spiral shape. The (query dependent spiral and shape coding) technique has the ability to represent huge data set with out reducing and hiding the information. Using this technique i.e. query dependent spiral and shape coding technique user can drill down data up to months and weeks level. User can find patterns, trends, clusters and correlations of a query result with respect to time.

4.2 Problem Solution

For in depth analysis, the purpose of this thesis is to develop a novel technique for mining the multivariate agricultural data with the innovative usage of data visualization. In this thesis two techniques i.e. shape coding (Icon Based Technique) which is used to represent time series data, and query dependent spiral technique (Pixel Oriented) that will be used to visualize only data which is relevant in the context of specific query are merged. The result of query from query dependent spiral techniques will be represented in the time grid. The Y-axis will represent weeks and X-axis will represent months. Each query record will be represented by 4 pixels. The query will operate on numeric fields, for non-numeric values we will just count the occurrence of the text field and assign it rank. For a query user will be able to select six options i.e. >, <, <=,>=, =, <> and in, between. User will be able to select only one table at a time for generating query. The query output will be divided into four parts, as we will represent query results in four weeks on x-axis. One query output will be represented in a separate window. For each query output two windows will be used. One window will represent the attribute distance, while the other will represent the over all distance. The technique i.e query dependent spiral and shape coding uses multiple windows for the different query parts, providing visual feedback for each part of the query and helping the user to understand the overall result. The thesis restricts the number of visualized dimensions to those that are part of the query, i.e. the dimensionality of visualizations corresponds to the number of selection predicates. Following steps are carried out in the described technique.

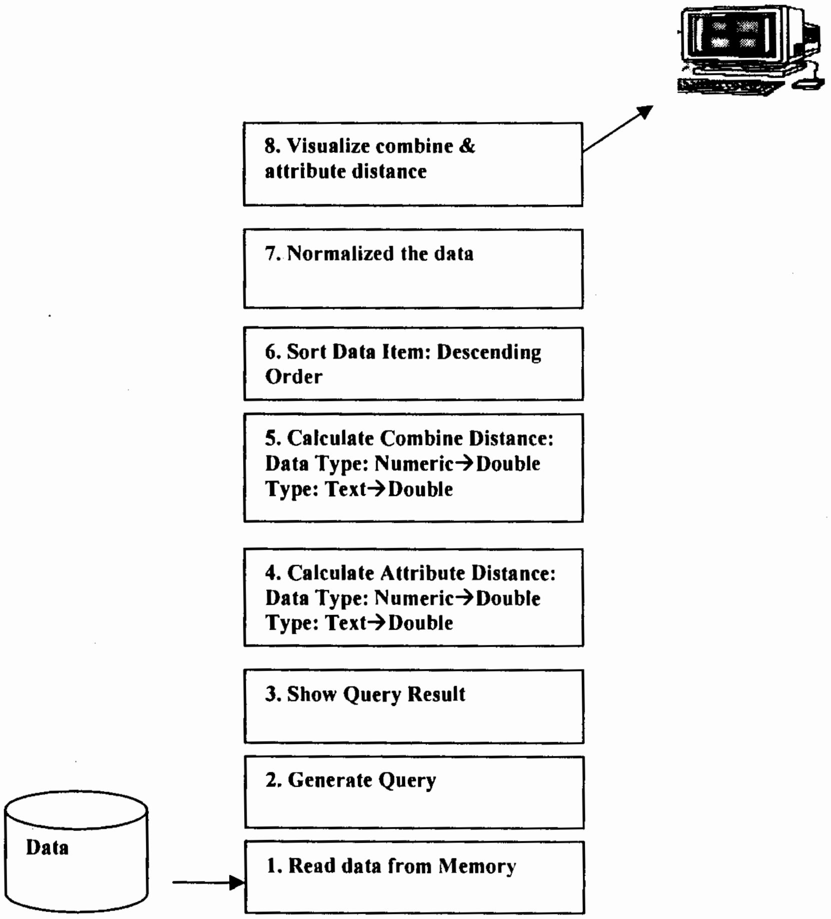

8. Visualize combine & attribute distance
7. Normalized the data
6. Sort Data Item: Descending Order
5. Calculate Combine Distance: Data Type: Numeric→Double Type: Text→Double
4. Calculate Attribute Distance: Data Type: Numeric→Double Type: Text→Double
3. Show Query Result
2. Generate Query
Data
1. Read data from Memory

Figure 4.1: Steps of Query Dependent Spiral and Shape Coding Technique

The main steps carried out during the project are:

- **Data:** Data is obtained from loading database of pesticides and predators from MS Access.

- **Filtering & Preprocessing:** In the filtering step, the raw data is preprocessed and extract information which is used in the mapping step. The filtering step includes operations like interpolating missing data or reduction the amount data. In this step pesticides data is smooth by removing the errors from the data set.

- **Mapping:** After filtering the mapping is the main core of visualization process. In this step preprocessed filtered data is used to transform, and to map different dimensions of data to date dimension.

- **Rendering:** Finally the data is interpreted in the form of color coded spiral shapes.

4.3 Modules

The proposed application has following modules

4.3.1 Module 1: Load Dataþase

The purpose of this module is to load the required database. User loads the database and the list of tables and their corresponding columns appears.

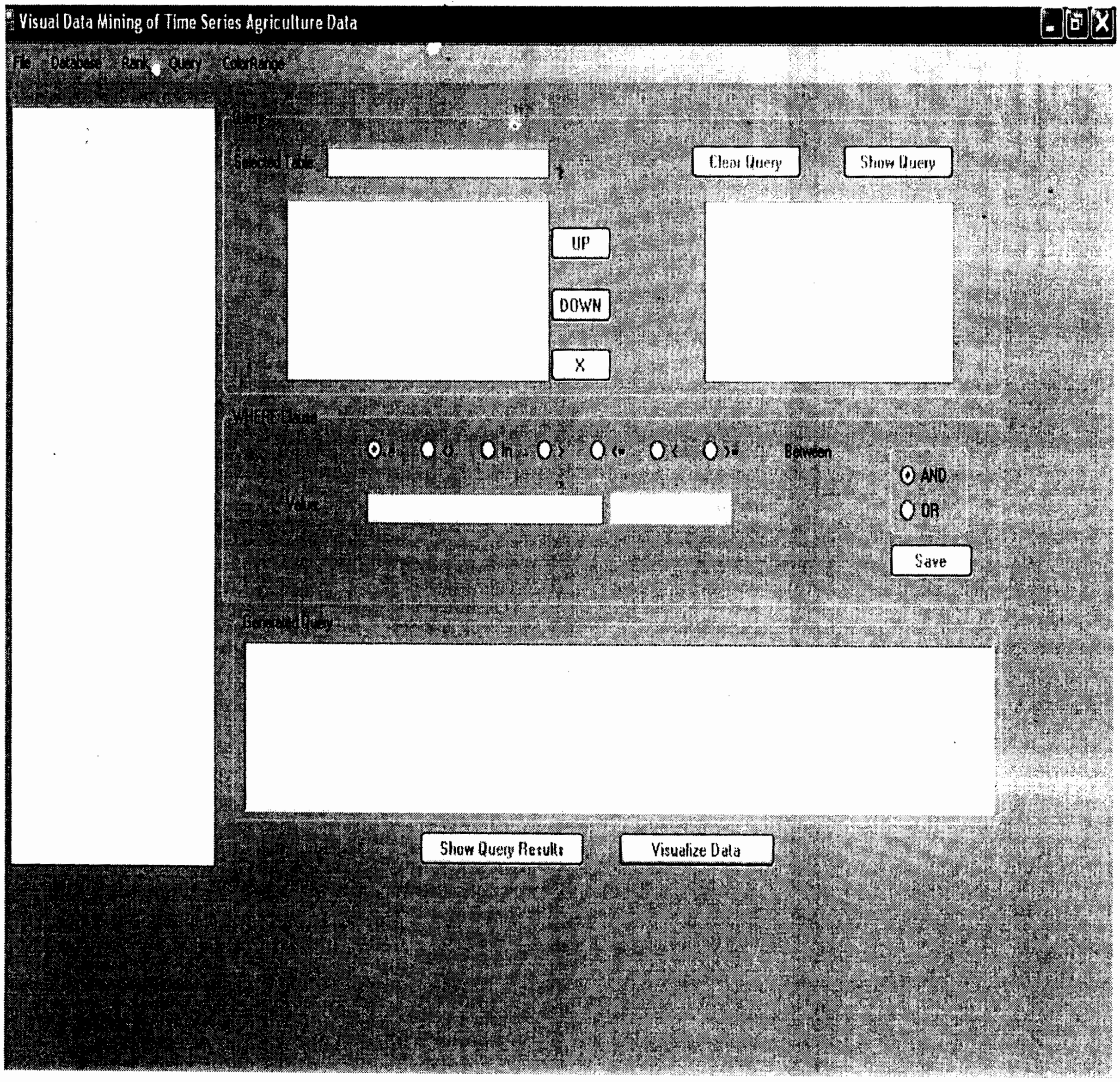
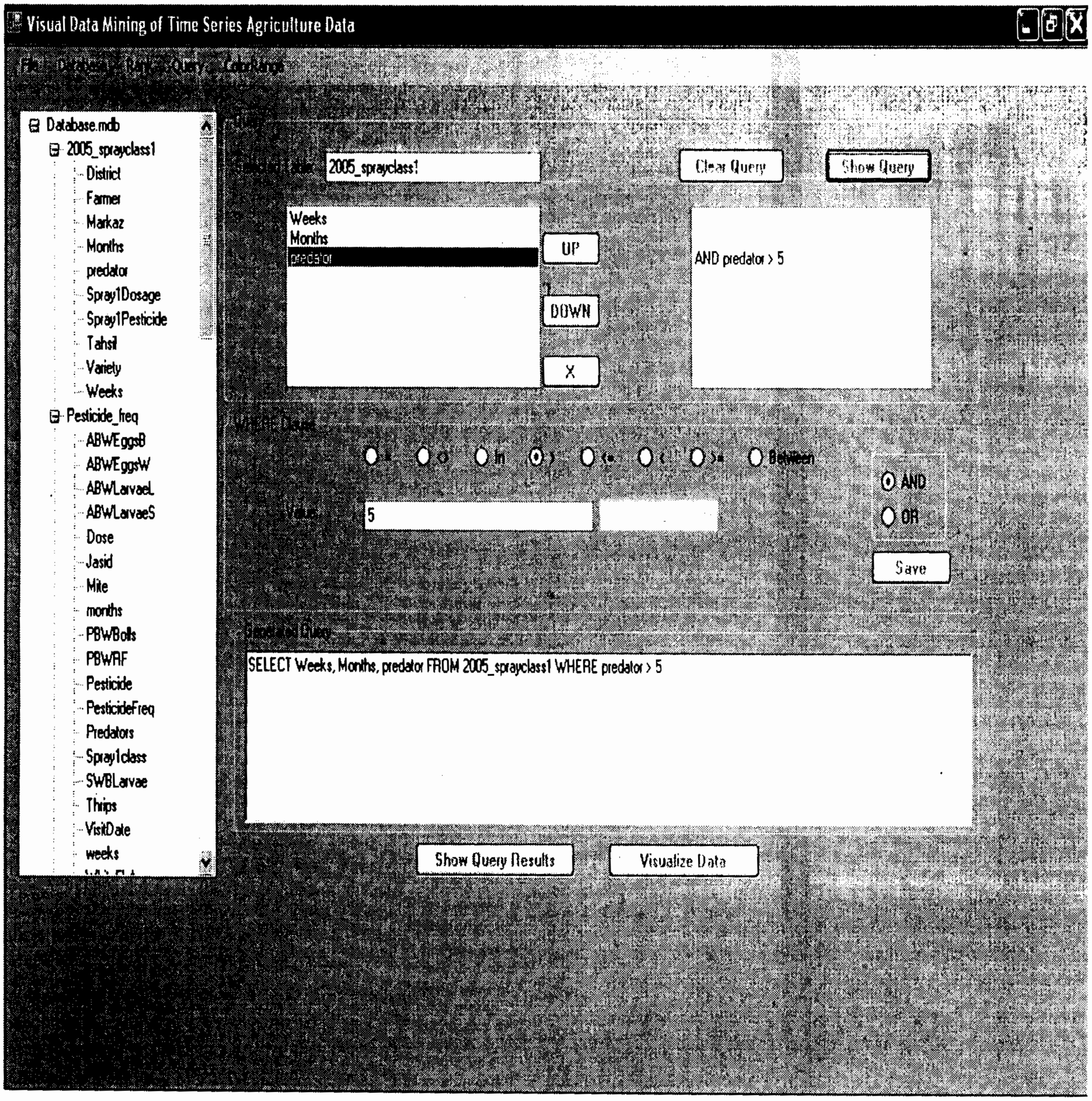


Figure 4.2: Load Database

4.3.2 Module 2: Generating Query using numeric attributes

User selects the table and the columns to be displayed in the query. Select the column name and defines its numeric value. User can also define AND, OR operation. User clicks on save option to save the generated query.



2005_sprayclass1

Weeks
Months
predator

AND predator > 5

● >

Value 5

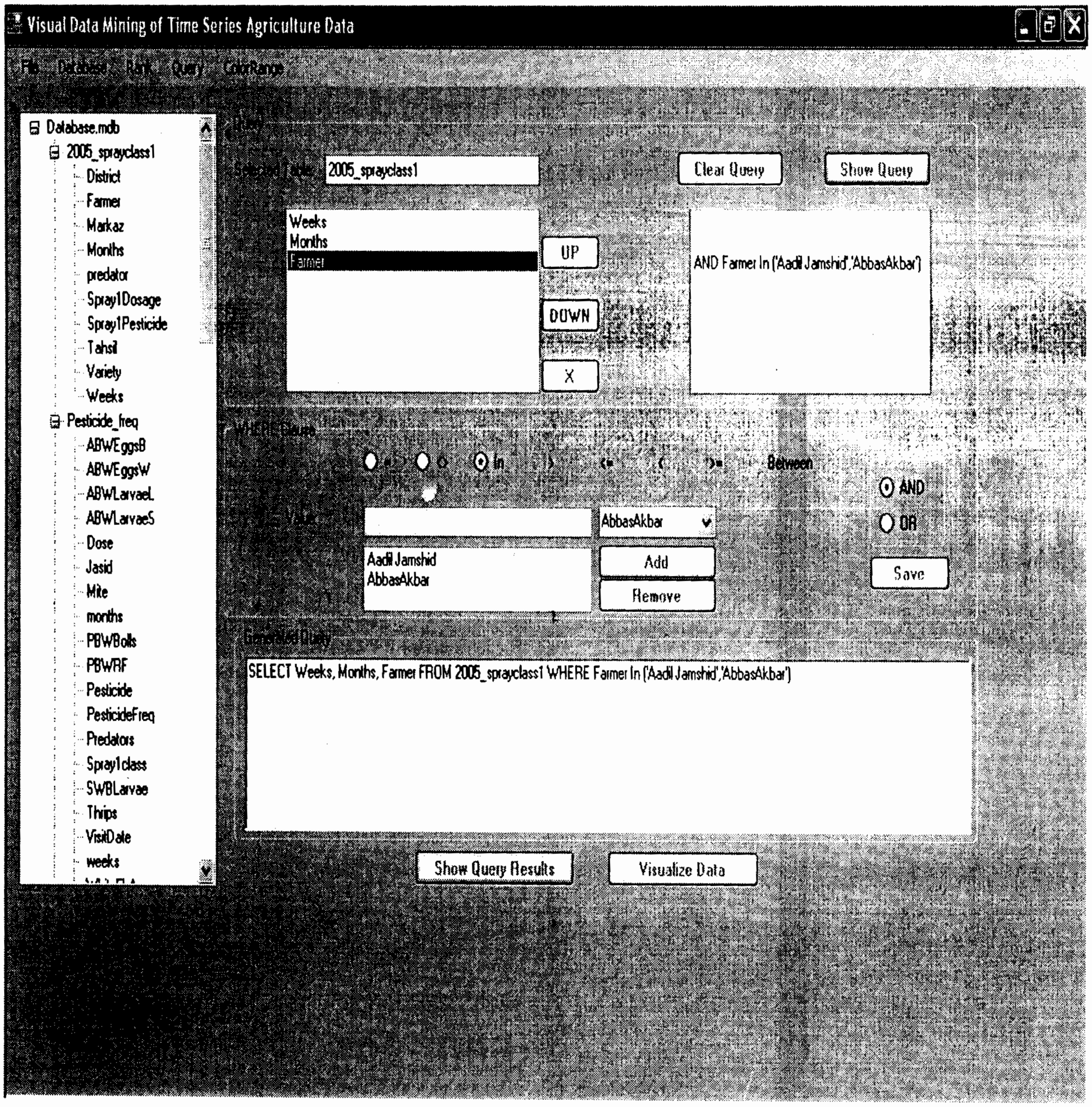
● AND

```
SELECT Weeks, Months, predator FROM 2005_sprayclass1 WHERE predator > 5
```

Figure 4.3: Generating Query Using Numeric Attributes

4.3.3 Module 3: Generating Query using textual attributes

User selects the text columns and selects its value from the drop down list and saves the value.



Selected Table: 2005_sprayclass1

Weeks
Months
Farmer

AND Farmer In ('Aadil Jamshid','AbbasAkbar')

○ ○ ● In

● AND ○ OR

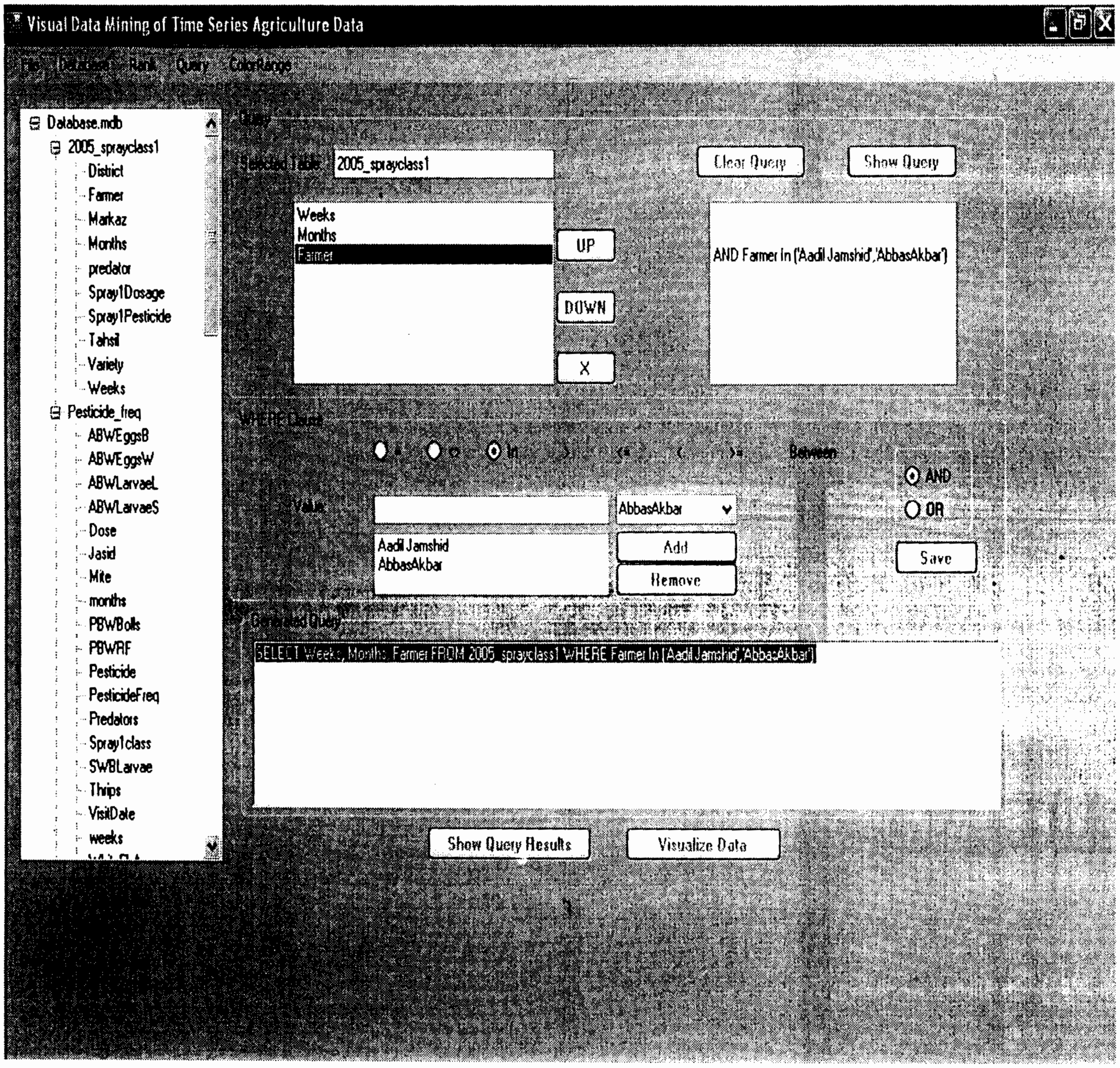
Value: AbbasAkbar

Aadil Jamshid
AbbasAkbar

```
SELECT Weeks, Months, Farmer FROM 2005_sprayclass1 WHERE Farmer In ('Aadil Jamshid','AbbasAkbar')
```

Figure 4.4 Generating Query using Textual Attributes

User clicks on "Show query" and can see the query in the "Generated Query" area.



Selected Table: 2005_sprayclass1

Weeks
Months
Farmer

AND Farmer In ('Aadil Jamshid','AbbasAkbar')

○ ○ ● In

● AND
○ OR

Value: AbbasAkbar

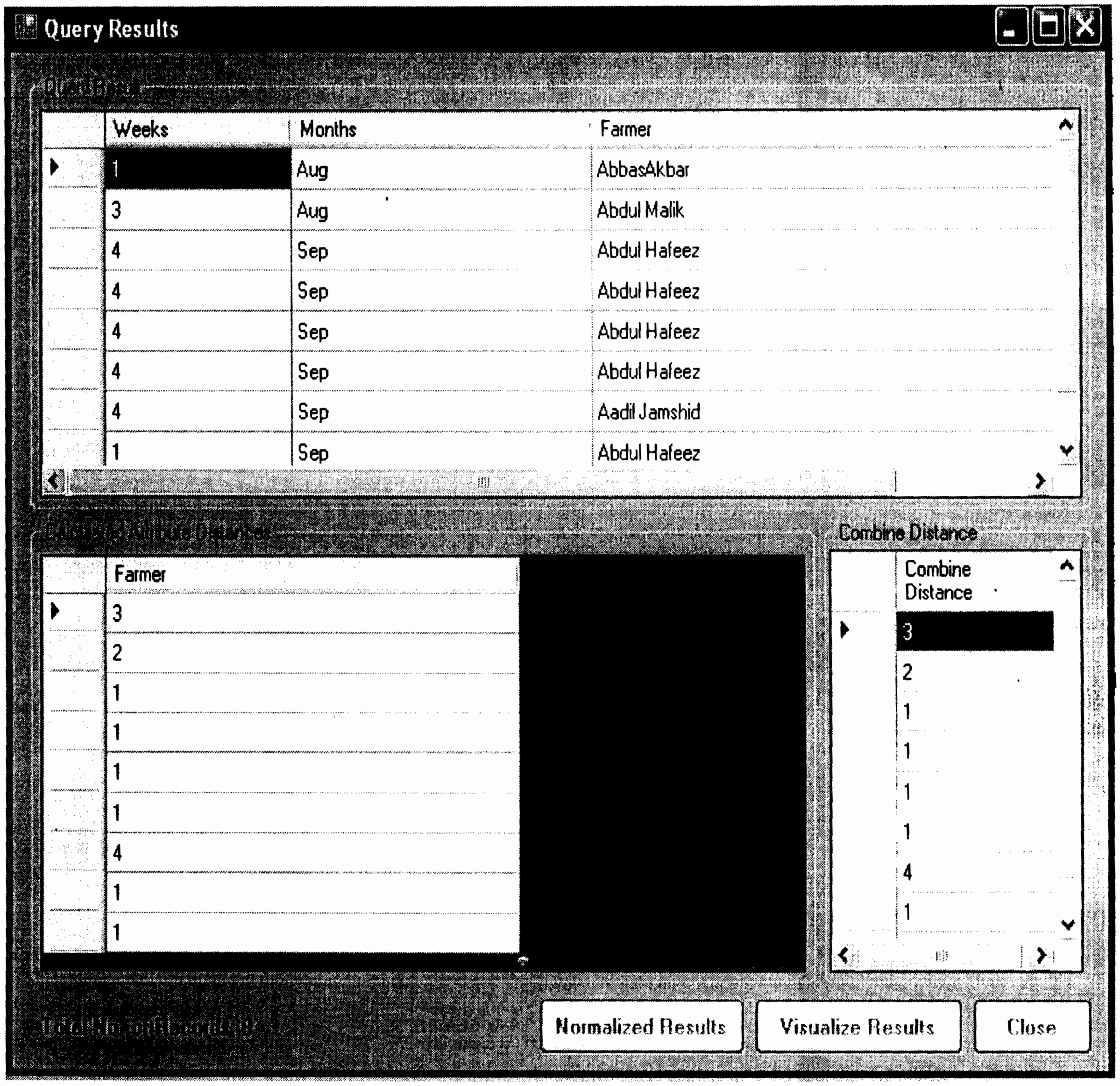
Aadil Jamshid
AbbasAkbar

```
SELECT Weeks, Months, Farmer FROM 2005_sprayclass1 WHERE Farmer In ('Aadil Jamshid','AbbasAkbar')
```

Figure 4.5: Generated Query for Textual Data

4.3.4 Module 3: Query Result, Attribute and Combine Distance

User then clicks on the "Show Query Results" to see the results. User sees the "Attributed Distance" and "Combine Distance".

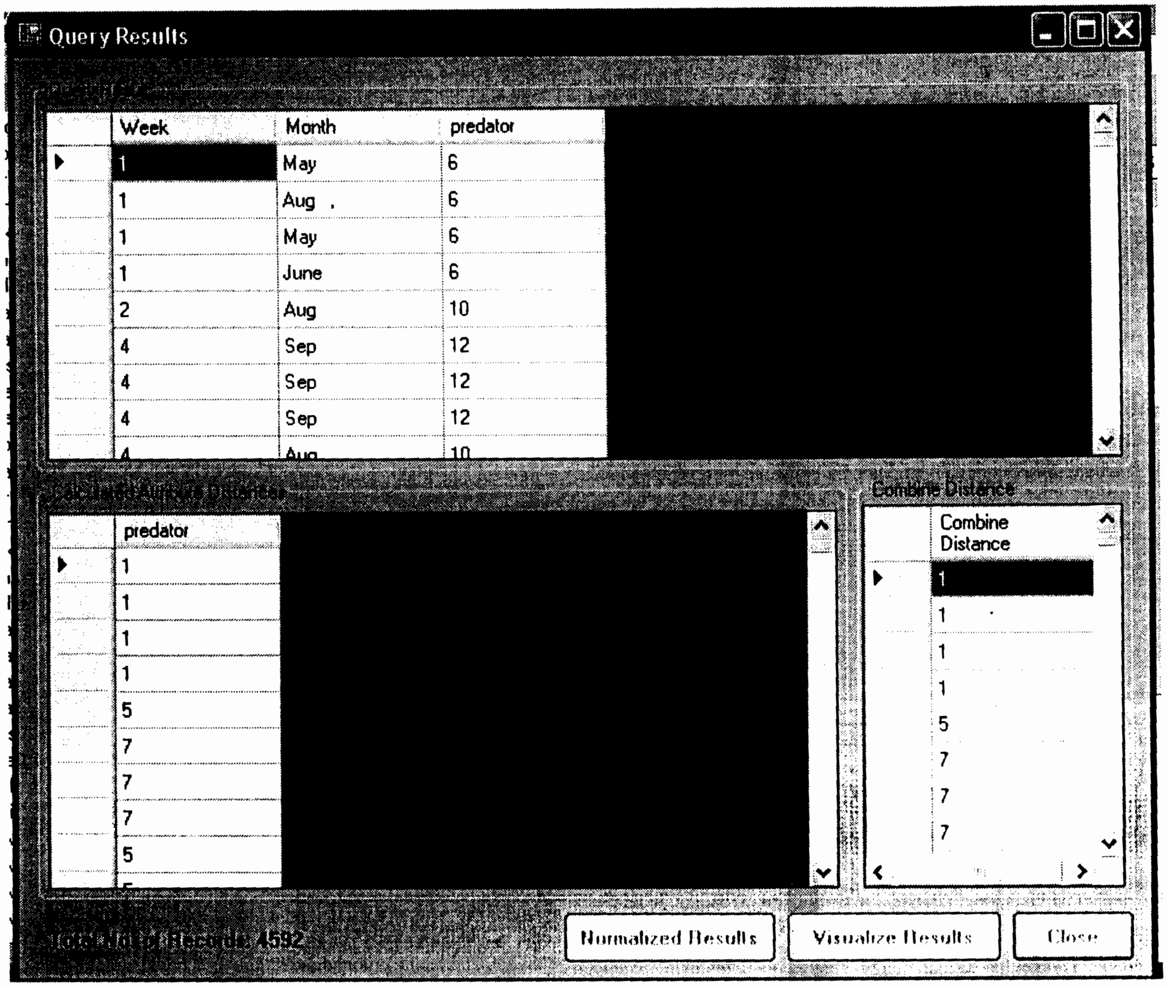


Weeks	Months	Farmer
1	Aug	AbbasAkbar
3	Aug	Abdul Malik
4	Sep	Abdul Hafeez
4	Sep	Abdul Hafeez
4	Sep	Abdul Hafeez
4	Sep	Abdul Hafeez
4	Sep	Aadil Jamshid
1	Sep	Abdul Hafeez

Farmer
3
2
1
1
1
1
4
1
1

Combine Distance
3
2
1
1
1
1
4
1

Figure 4.6(a): Query Result, Attribute and Combine Distance of Textual Data



Week	Month	predator
1	May	6
1	Aug .	6
1	May	6
1	June	6
2	Aug	10
4	Sep	12
4	Sep	12
4	Sep	12
4	Aug	10

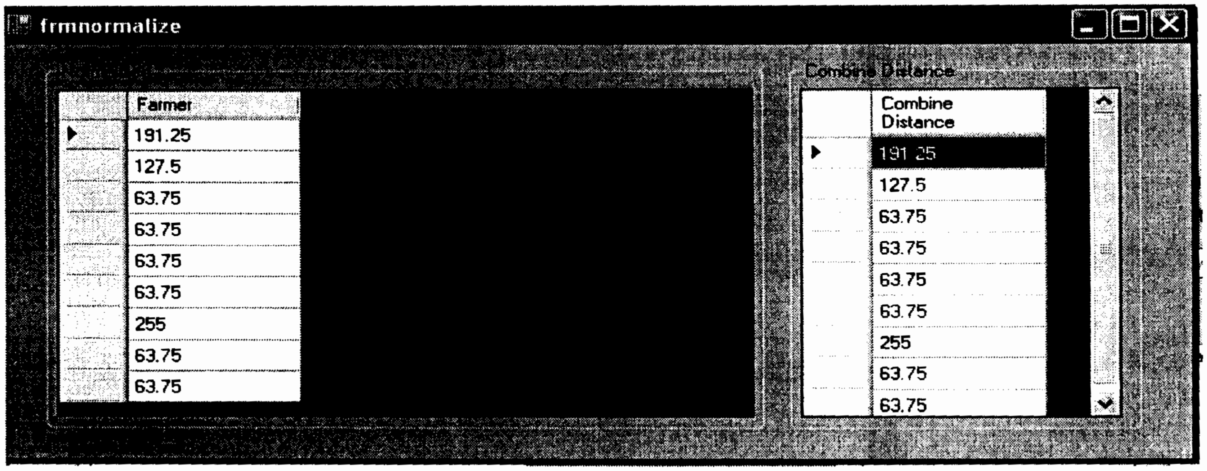
predator
1
1
1
1
5
7
7
7
5

Combine Distance
1
1
1
1
5
7
7
7

Total No of Records: 4592

Figure 4.6(b): Query Result, Attribute and Combine Distance of Numeric Data

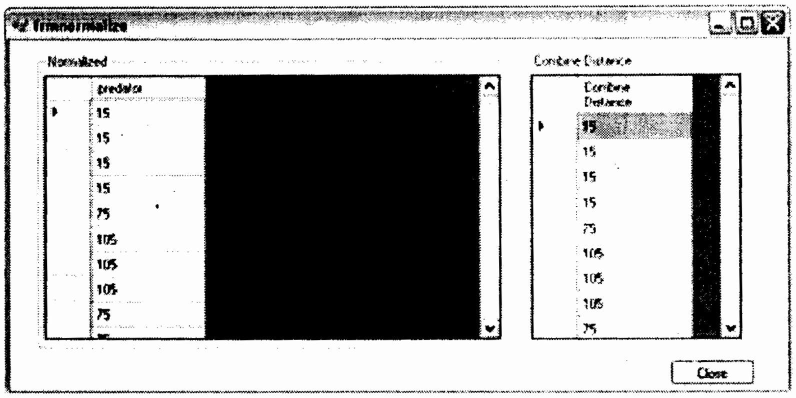
4.3.5 Module 3: Normalized View



Farmer
191.25
127.5
63.75
63.75
63.75
63.75
255
63.75
63.75

Combine Distance
191.25
127.5
63.75
63.75
63.75
63.75
255
63.75
63.75

Figure 4.7(a): Normalized View for Textual Attribute



Normalized

	predalsa
▸	15
	15
	15
	15
	75
	105
	105
	105
	75

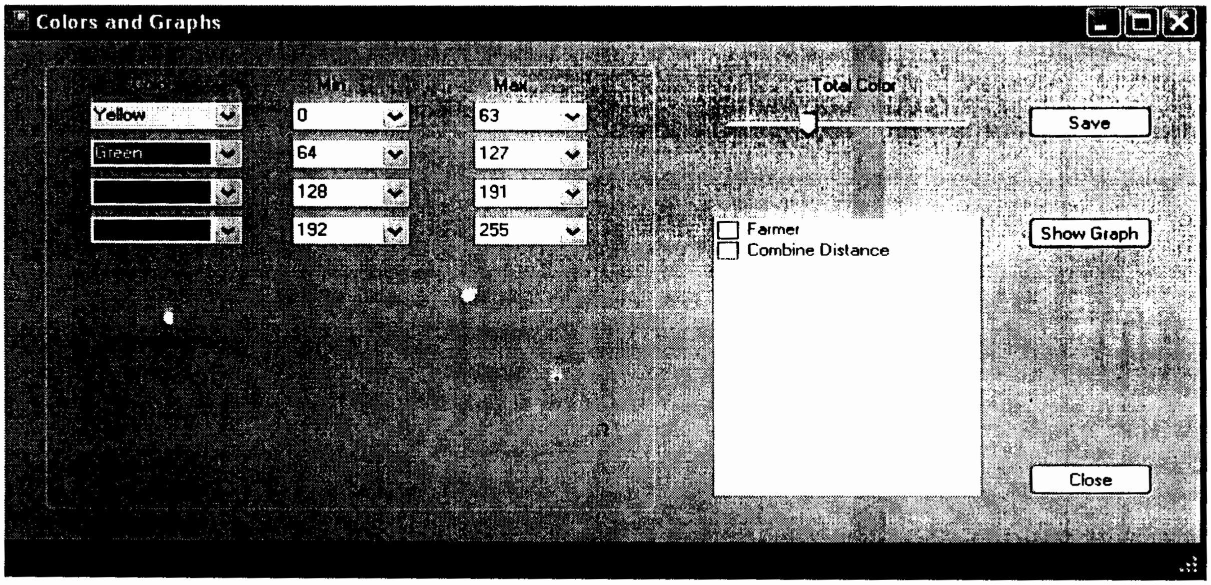
Combine Distance

	Combine Distance
▸	15
	15
	15
	15
	75
	105
	105
	105
	75

Figure 4.7(b): Normalized View for Numeric Attribute

4.3.6 Module 4: Visualization View and Analysis View

This view will be used for the analysis of faces displayed in first module. The analysis will on the basis of selection criteria. Analysis view will also be able to view the attribute distance and combine distance and will show data in cluster which will be represented in color coded pixels. The view represents the data with respect to time dimension and the grain is up to week level.



	Min	Max
Yellow	0	63
Green	64	127
	128	191
	192	255

□ Farmer
□ Combine Distance

Figure 4.8(a): Visualization View Color and Attribute Selection for Textual Attribute

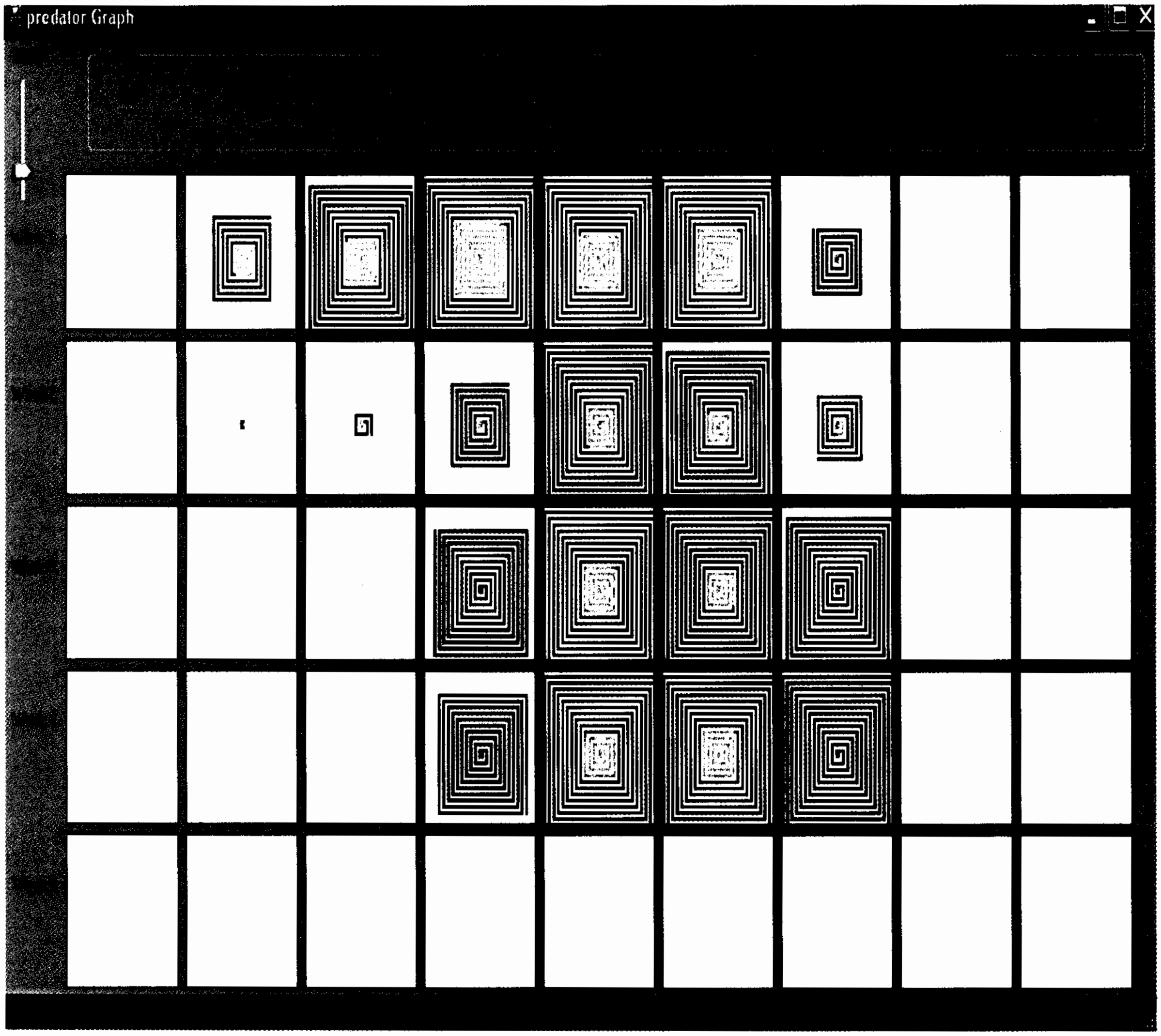


Figure 4.9: Analysis View of Attribute

From figure 4.9 it can be inferred that more occurrence of predators are observed in first week of May, June, July, Aug, Sep and October. The yellow region in the first week of July shows the highest incidence of predator falls in this time period as compared to other weeks and months. The denser the yellow color, more it satisfies the query criteria.. The lowest predator incidence is observed in second week of May. The predator incidence in fourth week of August and September is almost same. This shows similar trend of predator in fourth week of August and September. The predator incidence in second week of September with yellow color shows that only few values either, exactly match the query criteria or lies in the range of 1 to 25. The green color shows the values more close to the query criteria e.g. in range of 26-50 are more in number then that of values in range of 1-25. The blue color shows few values in range of 51-75 then that of green color. Red color shows values in range of 76-100. It can be concluded that in second week of September the values in range of 51-75 are highest as compared to other values of same week and month. The empty cell shows none of the predator existence is observed in these cells.

CHAPTER 5

DESIGN AND METHODOLOGY

5.1 Introduction

This chapter will discuss in detail the methodology adopted for the application development. Further software and hardware requirement are also presented in this chapter

5.2 Assumption/Dependencies

System dependencies are discussed below:

5.2.1 Software/ hardware Requirements

a) MS Visual Studio .Net 2005

b) At least Intel Pentium IV with 1 GB of RAM for server, 2.7 Processor.

5.2.2 Operating systems

a) MS Windows 2005 /XP/Vista

5.2.3 End-user characteristics

a) There will be only one user of this application.

5.2.4 Possible and/or probable changes in functionality

a) Allow user to load "MS ACCESS" databases.

5.3 Proposed Algorithm

The algorithm proposed for visual data mining in this thesis has the following steps.

1. Load the dataset on which the data mining algorithm is to be carried out.

```
Load table schema

LoadTableSchema(tableNode, tableName, conStr)
 Make an OledbConnection
OleDbConnection con = new OleDbConnection(conStr);
Make an arraylist to store table
        objArray = new object[] { null, null, tableName, null };
Open the connection
        con.Open();
Traverse each row in the columns of a table
         foreach (DataRow dr in columns.Rows)
```

```
Add column names to an array list
        info.Add(colName)
```

Table 5.1: Algorithm for Loading Database

2. Select the number of columns from the data tables.

```
Add a tree node
Check if the tree node contains any database and does not contain null.

    (e.Node.Text != "Database.mdb")

    (e.Node.FirstNode != null)

Assign the first table name to the text box
    txtTableName.Text = e.Node.Text
  Assign If the listbox contains column names clear them
    lstboxSelectedColumns.Items.Clear()

 Check if the column is already selected or not.
 If coulmn is already slected
        Print("Invalid column!")
Else
Add Column
  If user selects column without selecting table
    Print ("First select table.")
```

Table 5.2: Algorithm for Selection of Columns

3. Select the columns for "Where Clause" or selection predicate and assign numeric or textual values.

```
Get Column Data from table and selected

  CetColumnData(string tableName, string columnName)

Create a new Dataset by inititalizing an dataset object
Build a query to select distinct column from table
query = "SELECT DISTINCT " + columnName + " FROM " + tableName;

Make an OledbConnection pass it the connection string, open the connection fill
dataset and close the connection

        OleDbConnection con = new OleDbConnection(General.ConStr);
```

```
OleDbCommand cmd = new OleDbCommand(query, con);
OleDbDataAdapter da = new OleDbDataAdapter(cmd);
con.Open();
da.Fill(ds);
con.Close();
```

Table 5.3: Algorithm to Select the Columns for "Where Clause"

4. The user is able to select one table at a time and multiple columns.

User will be facilitated with following options: >, <, <=,>=, =, In, Between, And, Or, text field and drop down list for selecting text attribute.

```
 Step1: User selects one of the following Radio Button for numeric attributes
            if
               clause = "<="; Chek it
            else if
               clause = ">="; Chek it
            else if
               clause = "<>"; Chek it

            else if
  clause = "=";   Chek it
            else if
               clause = ">"; Chek it
            else if
               clause = "In"; Chek it
            else if
               clause = "<"; Chek it
            else if
          clause = "Between"; Chek it

Step2: If user selects "AND" option
            andor = "AND"; Chek it

 else
 andor = "OR"; Chek it

 Step3: User selects one of the following Radio Button for textual attributes
           The text field is enabled

           if (clause is ("In") OR clause("Between"))
              User enter values in text field
              Else if
           Print("Please enter values")
```

```
        Else user selects value from drop down list
```

Table 5.4: Algorithm for Selection of Operations

5. Query is generated and user can view it e.g **SELECT Month, Week, predator FROM 2005_sprayclass1 WHERE predator > 5**

6. Click on show query button, query is generated

```
        Make a string for "Select" command
           string query = "SELECT ";

        Traverse through all items of listbox
           for (int i = 0; i < lstboxSelectedColumns.Items.Count; i++)

        Store the selected column from list box into a temporary string
        and split the column names with ","

          temp = lstboxSelectedColumns.Items[i].ToString().Split(' ');
        Store Queryparameters in the first index of temporary string

       query += temp[0];
                   QueryParameters.Columns.Add(temp[0]);

       Make a string for "FROM" command
               query += " FROM " + TABLENAME;
       Make a string for "WHERE" command

               query += " WHERE ";
          Make an array list to store "AND", "OR" operations.

                       ArrayList tempList = new ArrayList();

                tempList.Add(andor);
```

Table 5.5: Algorithm for Query Generation

7. It will calculate the attribute distance and only in selection predicates.

```
       Make an array list to calculate attribute distance

       ArrayList CalculeteAttributeDistance()
       Initalize and atore distancedata in an arraylist
            ArrayList distanceData = new ArrayList();
```

```
Count each coulmn items given in the "where clause"
for (int i = 0; i < QueryParameters.Columns.Count; i++)

 ArrayList temp=(QueryParameters.WhereClause[i]);

If the count is not equal to zero store the column into another arraylist

            if (temp.Count > 0)
        ArrayList tempList = new ArrayList();

    tempList=GetColumnData(QueryParameters.Columns[i]);

Call a funtion to calculate attribute distance
ArrayList attDistance =CalculateDistance(tempList,temp,i);

 If the result of where clause is a textual data, count the frequency  of each item

ArrayList attDistance = CalculateFrequencyDistance(freqList, temp, tempList)

Store Months and weeeks into a temporary arraylist  as attribute distance fot these
two columns is not calculated

colName = QueryParameters.Columns[i].ToString().ToLower();
 if (colName.Equals("week") || colName.Equals("weeks"))

Week = tempList;

else if (colName.Equals("month") || colName.Equals("months"))

Month = tempList;
```

Table 5.6: Algorithm for Calculating the Attribute Distance

8. It will calculate the combine distance using attribute distance. The combine distance is also calculated for the attributes which shows the attribute relevancy. The combine distance is calculated by assigning priority to the attributes which are on left side of selection predicate and multiplying it with 2 and adding the attributes.

```
Initialize power=0
Count the number of column in a grid view that displays "attribute distance"
for (int i = 0; i < dgvAttDist.Columns.Count; i++)
```

```
If cloumns exists in gridview then
Add one to the power
power++;
Check if column is attribute column, then assign 2 the power.

    tempVal = Math.Pow(2, p--)
```

Table 5.7: Algorithm for Calculating the Combine Distance

9. The attribute distance for non-numeric attribute is calculated on basis of attribute occurrence, it is then assigned different ranks in descending order. These ranks are then assigned to each occurrence of that attribute. The text values are converted to numeric values because the application is able to handle only numeric values

```
Count Item frequency of each item for each column
Assign the frequency to each distinct item of a column in descending order
Assign rank form one to onwards to each item whose frequency has been
calculated and assigned
Assign these ranks to the corresponding items in attribute distance column
```

Table 5.8: Algorithm for Calculating the Frequency of Textual Data

10. Query result show all attributes in the "Select" clause along with month, year and week.

11. Normalize the attribute and combine distance. In this thesis normalization is a process in which each column data looks consistent i.e in the range of 0-255. As, the RGB color ranges are from 0-255.

```
Get maximum value from each column of attribute distance
Divide each row of each column with the maximum value
Multiply the result obtained in above step by 255.
for (int j = 0; j < colData.Count; j++)
{
(colData[j]) / maxValInColumn * 255
}
Round the decimal number to maximum two figures
```

```
The values are normalized
```

Table 5.9: Algorithm for Normalizing Data

12. Select the color ranges and attribute for which graph is to be displayed.

```
  Set four Default Color of combo box
  comboBox1.SelectedIndex = 138;//Yellow
  comboBox4.SelectedIndex = 51;//Green
  comboBox7.SelectedIndex = 9;//Blue
  comboBox10.SelectedIndex = 113;//Red

  save ranges of colors in combo boxes
  initilize nrange=4
  initialize varaible step, which gives rages for first four by default seelcted colors
  initialize min=0;
  initialize max=255;

  int step = (max - min) / nRanges);

Set Color Combo using RGB color function "color"
  for (int i = 0; i < nRanges; i++)

ComboBox combobox = getComboBox("comboBox" + (i * 3 + 1));
Color color = Color.FromName(combobox.SelectedItem.ToString)

Set Min Combo boxes
 combobox = getComboBox("comboBox" + (i * 3 + 2));
Set Max Combo boxes
 combobox = getComboBox("comboBox" + (i * 3 + 3));
```

Table 5.10(a): Algorithm for Selecting Color Ranges

13. The output will be in the form of color coded pixels, different colors shows how close or far is the attribute from the desire value. The spiral shape of the pixels represents the relevancy of the attributes.

```
Set y-axis to 12 values representing the months.

for (int m = 0; m < 12; m++)
Set x-axis to 5 values representing the weeks.


for (int w = 0; w < 5; w++)
```

```
        Call drawspiral function and sot the data
        DrawSpiral((ArrayList)data.SortedData[m + "-" + w], ((m - 1) * 10 + w));
        Check if the data exists to draw spiral
        if (data == null)
        PictureBox pBox = null;
Start drwing spiral in the center of window. As the cell size is 100*100. Point of drawing is 50, 50
        Point p = new Point(50, 50);
Each record is represented by 4 pixels.
                int diff = 2;// 2*2=4 Pixels
Set the direction in which spiral is to be drawn
        Direction { Down = 0, Left = 1, Up = 2, Right = 3, Same = 4 }

Create a generic list of strings to store colors
            ArrayList colors = new ArrayList();
Get the color names from the Known color enum
Iterate through each string in the colorNames array
Cast the colorName into a KnownColor
Check if the knownColor variable is a System color
Add it to our list
```

Table 5.10(b): Algorithm for Drawing Spirals

14. The width of spiral pixel is 4, thus it has the ability to show maximum values a cell. Thus avoiding the overlap.

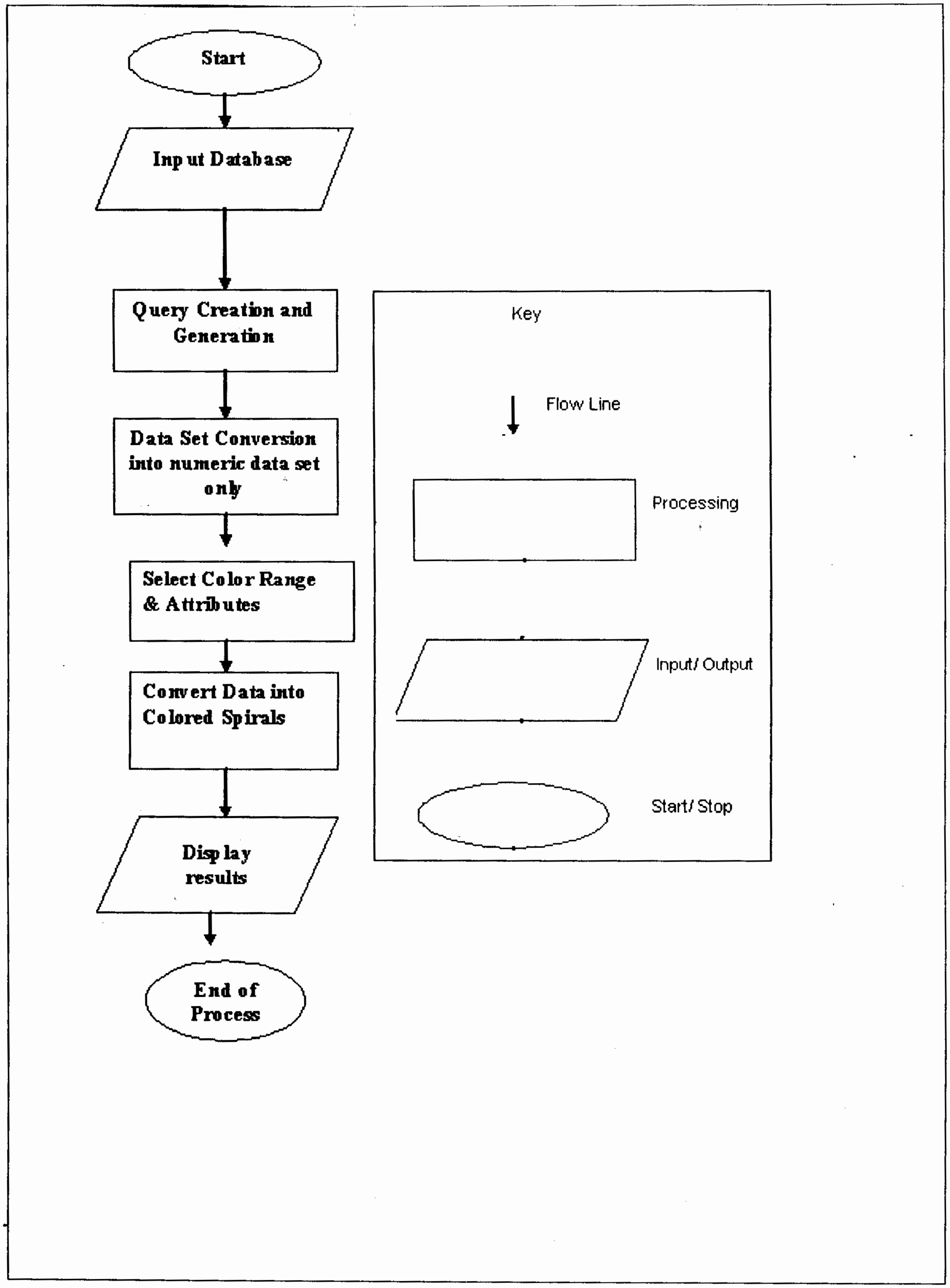

Start
Input Database
Query Creation and Generation
Data Set Conversion into numeric data set only
Select Color Range & Attributes
Convert Data into Colored Spirals
Display results
End of Process
Key
Flow Line
Processing
Input/ Output
Start/ Stop

Figure 5.1: System Model

5.4 Flow Charts

The code for this technique is written in c#. The steps of execution are shown in the flowchart below. Input is provided to it in the form of data tables. First of all it checks the availability of data base. Select the table from the list. Load the maximum columns from the selected table.

5.5 System Architecture

5.5.1 Actor Identification

Actors	Type	Description
User	User	User will have administrative rights, and will have full access to the application. User will be able to perform all activities and tasks, i.e. loading, visualize loaded files, apply visualization techniques on un-clustered data, view the original data values and display results in the form of color coded spiral shape of pixels.
System	System	System will perform all backend calculations and generate results to be displayed. It will respond to the any action/input from user and act accordingly. The major tasks of system will be to uploaded table, perform preprocess calculations on it, display data in form of grid, apply algorithms like converting text data into numeric and calculating attribute and combine distance. And displaying results.

Table 5.11: Actor Identification

5.6 Use case diagram of the System

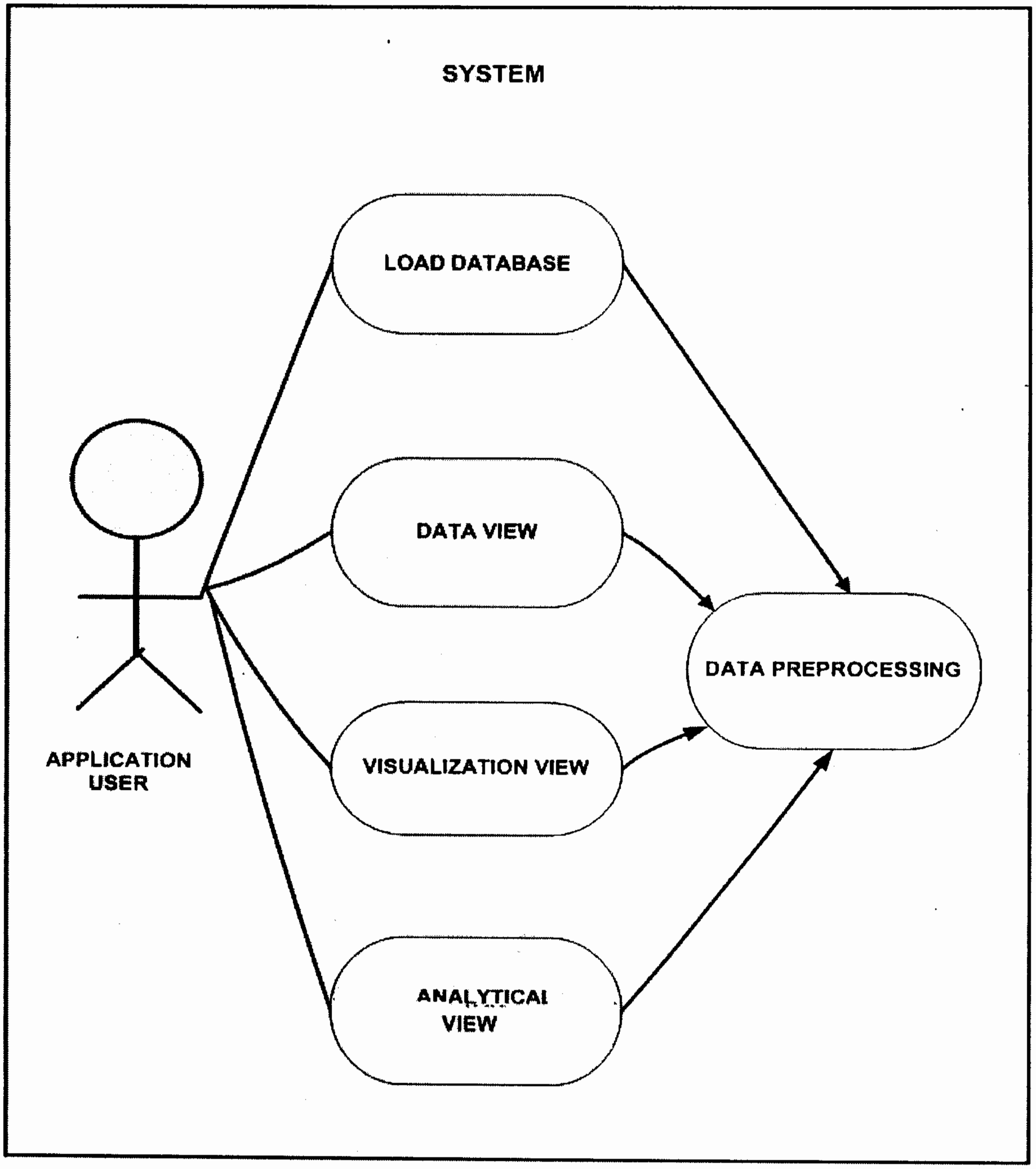


Figure 5.2: Use case diagram of the System

5.7 Use Case Descriptions

5.7.1 Load Database

UC – 01	
Name	Load Database File
Description	After starting application named as 'Visual Data Mining of Time Series Agriculture Data' user will go to Database menu and choose to load database file option and then specify the data tables to be loaded, if valid the selected table will be loaded and available to use.
Actors	1. User
Pre Conditions	1. User starts application.
	2. User is shown main screen.
	3. User opens database menu.
	4. User selects: *Load Database*
	5. User finds and select data table
Post Conditions	1. User will successfully select and open the database table

Table 5.12: Load Database

Load Database File

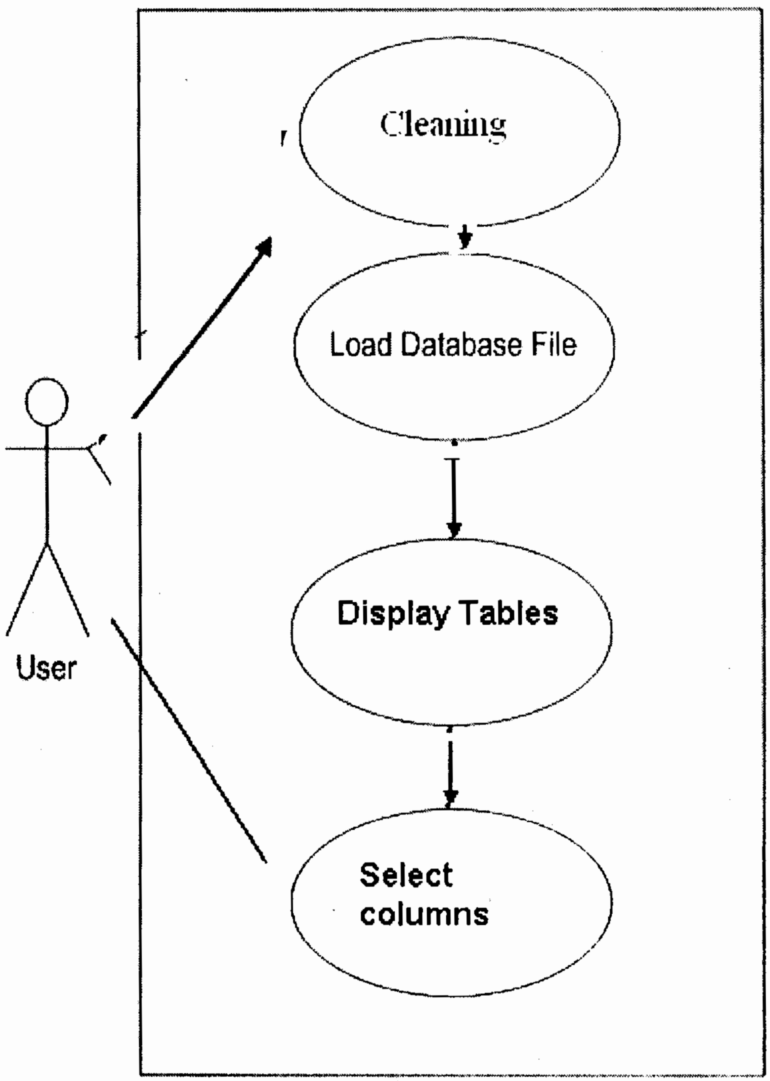


Figure 5.3: Load Database

5.7.2 Data set conversion

UC – 02	
Name	Create Datasets
Description	After the selected table is parsed datasets will be created for each database selected. The data is parsed into numeric format.
Actors	System
Pre Conditions	
Post Conditions	1. The grid view will be populated with available numeric data set.
Steps	1. System takes the table values.
	2. System creates the package of data and other necessary information according to a predefined format.

Table 5.13: Data set conversion

5.7.3 Normalize Values

UC – 03	
Name	Normalize Data
Description	The data will be shown in normalized view. For each column maximum value is selected and each row is divided by it and multiplied with 255.
Actors	System
Pre Conditions	1. Loaded table is authenticated
	2. Table is parsed and stored.
	3. Datasets are created from data.
	4. Datasets are displayed in grid view
Post Conditions	-
Steps	1. System checks for available data in dataset class.
	2. System gets data from attribute distance and combines distance columns.
	3. System stores the data array lists in the current array lists.
	4. System normalizes the values.

Table 5.14: Normalize Values

5.7.4 Color Selection

UC – 04	
Name	Color Selection
Description	The normalized data is shown in range 0-255 using color scales have different color. The maximum colors that can be view are 10. To facilitate user so that he can view query results in maximum color. Rather, then using only 4 colors i.e. yellow, green, blue and red only. Four colors are set by default as query is displayed in 4 weeks and each record is represented by 4 pixels. If the query results in a huge data set user should be facilitated to select more color range to have more clear view of data.
Actors	System
	1 Datasets are displayed in grid view
	2. Table is parsed and stored.
	3. Datasets are created from data.
Pre Conditions	4.Datasets are displayed in grid view
	5.Data is normalized
Post Conditions	-
Steps	1. System checks for available data in dataset class.
	2. System gets data from attribute distance and combines distar

	columns.
	3. System stores the data array lists in the current array lists.
	4. System normalizes the values.
	5. System displays color ranges to be selected

Table 5.15: Color Selection

5.7.5 Spiral/Graph View

UC – 05	
Name	Spiral/ Graph View
Description	Every attribute is displayed in a separate window representing the values in spiral shape with respect to weeks and months
Actors	System
Pre Conditions	1. System displays color ranges to be selected
Post Conditions	-
Steps	1. System checks for available data in dataset class.
	2. System gets data from attribute distance and combines distance columns.
	3. System stores the data array lists in the current array lists.
	4. System normalizes the values.
	5. System displays color ranges to be selected
	6. Select attribute for which visualization is required.

Table 5.16: Spiral/ Graph View

CHAPTER 6

TESTING

6.1 Introduction

Testing is the process of executing software to verify that it satisfies the specified requirement. It is the strategy that recovers as many defects as possible. Since no program or system design is perfect at the final implementation, therefore, testing is an essential requirement. The proposed system is tested by giving datasets as the input to the system. The results show that how effects of pesticides on human life is find by using proposed technique. The test cases for this purpose are given below.

6.2 Results:

Test Case ID	Test Case	References	Result
1	Verify that the database opens successfully. (Table 5.1)	TC1	Passed, Figure 4.1
2	Verify that table of db opens along with columns. (Table 5.2)	TC2	Passed Figure 4.2
3	Verify that the columns are selected successfully the where clause. (Table 5.3)	TC3	Passed Figure 4.3
4	Verify that the operation on columns is performed successfully. (Table 5.4)	TC4	Passed Figure 4.4
5	Verify that query is generated successfully. (Table 5.5)	TC5	Passed Figure 4.5
6	Verify that the result of query is correctly displayed (Table 5.5)	TC6	Passed Figure 4.6(a)
7	Verify that the results on "where" clause are calculated and shows attribute distance successfully (Table 5.6)	TC7	Passed Figure 4.6(a)
8	Verify that the combine distance is	TC8	Passed

	calculated successfully.(Table 5.7)		Figure 4.6(a)
9	Verify that the normalization of attribute distance is performed successfully. (Table 5.8)	TC9	Passed Figure 4.7(a)
10	Verify that the normalization of combine distance is performed successfully. (Table 5.8)	TC10	Passed Figure 4.7(a)
11	Verify that user can select color ranges (Table 5.9)	TC11	Passed Figure 4.8(a)
12	Verify that user can select attributes to be displayed for visualization (Table 5.9)	TC12	Passed Figure 4.8(a)
13	Verify that the results are displayed and spirals shapes with colors are displayed properly in proper place. (Table 5.10)	TC13	Passed Figure 4.9

6.2.1 Case 1

Test Case ID: 1 **Test Engineers:** Humaira

Test Case Reference TC1

Objective Verify that database opens successfully.

Method:

1. Load Database.
2. Open file of "MS-Access" from any location on computer.
3.

Comments:

Passed

Table 6.1 Test Case 1

6.2.2 Case 2

Test Case ID: 2	**Test Engineers:** Humaira
Test Case Reference	TC2
Objective	Verify that table of db opens along with columns
Method: 1. Open table. 2. Specific columns of each table appear successfully.	
Comments: Passed	

Table 6.2 Test Case 2

6.2.3 Case 3

Test Case ID: 3	**Test Engineers:** Humaira
Test Case Reference	TC3
Objective	Verify that the columns are selected successfully.
Method: 1. Open table. 2. Select desire columns from table.	
Comments: Passed	

Table 6.3 Test Case 3

6.2.4 Case 4

Test Case ID: 4	**Test Engineers:** Humaira
Test Case Reference	TC4
Objective	Verify that the operation on columns is performed successfully
Method: 1. Perform actions on columns E.g >,<,=,<>,In, Between, And, Or.	
Comments: Passed	

Table 6.4 Test Case 4

6.2.5 Case 5

Test Case ID: 5	**Test Engineers:** Humaira
Test Case Reference	TC5
Objective	Verify that the query is generated successfully.
Method: 1. The query is generated in the same order as intended by the user.	
Comments: Passed	

Table 6.5 Test Case 5

6.2.6 Case 6

Test Case ID: 6	**Test Engineers:** Humaira
Test Case Reference	TC6
Objective	Verify that the result of query is correctly displayed.
Method: 1. The correct results of query are displayed, along with all the selected columns with specific criteria.	
Comments: Passed	

Table 6.6 Test Case 6

6.2.7 Case 7

Test Case ID: 7	**Test Engineers:** Humaira
Test Case Reference	TC7
Objective	Verify that the results on “where” clause are calculated and shows attribute distance successfully.
Method:	1. The correct results of attribute distance are displayed which selected in “where clause”. 2. The attribute distance is calculated from criteria defined and query range via subtracting.
Comments:	Passed

Table 6.7 Test Case 7

6.2.8 Case 8

Test Case ID: 8	**Test Engineers:** Humaira
Test Case Reference	TC8
Objective	Verify that the combine distance is calculated successfully.
Method:	1. The correct results of combine distance are displayed that is calculated from attribute distance. 2. It is calculated via following formula for each row attrutex^{2}+attributey^{1}+attributez^{0} 3. The attribute selected first h as the highest priority and then so on.
Comments:	Passed

Table 6.8 Test Case 8

6.2.9 Case 9

Test Case ID:9	**Test Engineers:** Humaira
Test Case Reference	TC9
Objective	Verify that the normalization of attribute distance is performed successfully.
Method:	1. The correct results of normalization are displayed. Each column maximum value is selected and each row is divided by this value and multiplied by 255.
Comments:	Passed

Table 6.9 Test Case 9

6.2.10 Case 10

Test Case ID:10	**Test Engineers:** Humaira
Test Case Reference	TC10
Objective	Verify that the normalization of combine distance is performed successfully.
Method:	1. The correct results of normalization are displayed. Each column maximum value is selected and each row is divided by this value and multiplied by 255.
Comments:	Passed

Table 6.10 Test Case 10

6.2.11 Case 11

Test Case ID:11	**Test Engineers:** Humaira
Test Case Reference	TC11
Objective	Verify that user can select color ranges
Method:	1. User select color form color range and can select maximum up to 10 values for visualization.
Comments:	Passed

Table 6.11 Test Case 11

6.2.12 Case 12

Test Case ID:12	**Test Engineers:** Humaira
Test Case Reference	TC12
Objective	Verify that user can select attributes to be displayed for visualization
Method:	1. User select attribute given in where clause for visualization by clicking on check boxes.
Comments: Passed	

Table 6.12 Test Case 12

6.2.13 Case 13

Test Case ID:13	**Test Engineers:** Humaira
Test Case Reference	TC13
Objective	Verify that the results are displayed and spirals shapes with colors are displayed properly in proper place.
Method:	1. User clicks on show graph to have visualization view of query. 2. User can Zoom in and Zoom out.
Comments: Passed	

Table 6.13: Test Case 13

Results of testing are shown in chapter 5 along with the screen shots.

CONCLUSIONS

Query dependent Spiral is a useful pixel based technique which uses features in color spiral the variables to depict multivariate data. Spiral shows clustering on the basis of similar data in term of weeks and months. Although we can perform analysis using conventional charts and graphs but they show a more abstract level of detail of the data using sums and aggregates, in which we can not view all records individually, but using Query dependent Spiral along with shape coding technique we can do the in-depth analysis by viewing each record in form of spiral as shown in Figure 7.1. Figure 7.1 represent 1148 records at different location, while it is not possible to view 1148 records using traditional graphs. The traditional graphs will show over lapping and records will be hidden as shown in Figure 7.2.

Policy makers depend on the sophistication of their decision making process and on their willingness to collect and record the data needed for problem solving. This study gives results in the form of color spirals by using pesticides data. Pesticide data exists in complex form which is difficult to understand by simple users. To facilitate the decision making process of farmers, the data is normalized. For the normalization, the complex data has been preprocessed and converted into simple form. Clustering is performed on the basis of selection criteria of parameters. The parameters used to find the effects of pesticides on predators and pests are Weeks, Months, District Multan, Spray1Pesticide, Spray1Dosage, pesticides, pesticides frequency, dose, Jasid, thrip Variety, Farmer, predators etc. The output shows effect by different color ranges form lighter to brighter image in District Multan. The technique can not only be used for numeric data, but also have the flexibility to deal with any textual data.

SELECT Weeks, Months, predator FROM 2005_sprayclass1 WHERE predator > 5

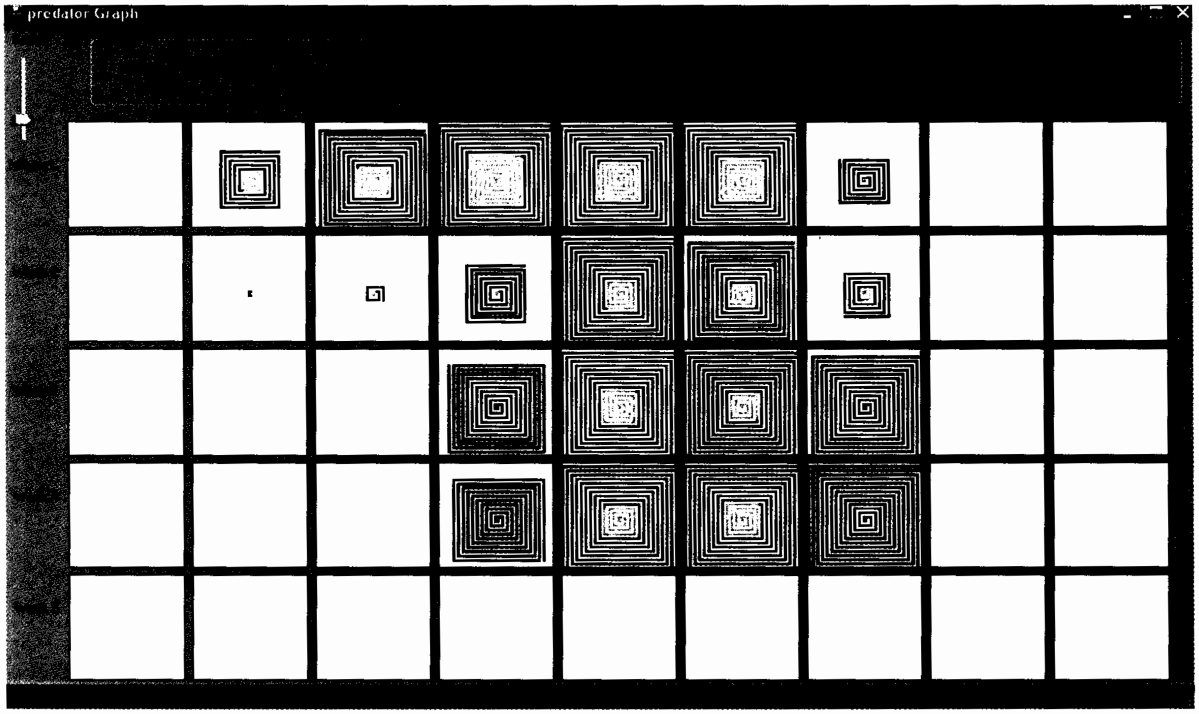


Figure 7.1:Spiral showing Predator on specific months

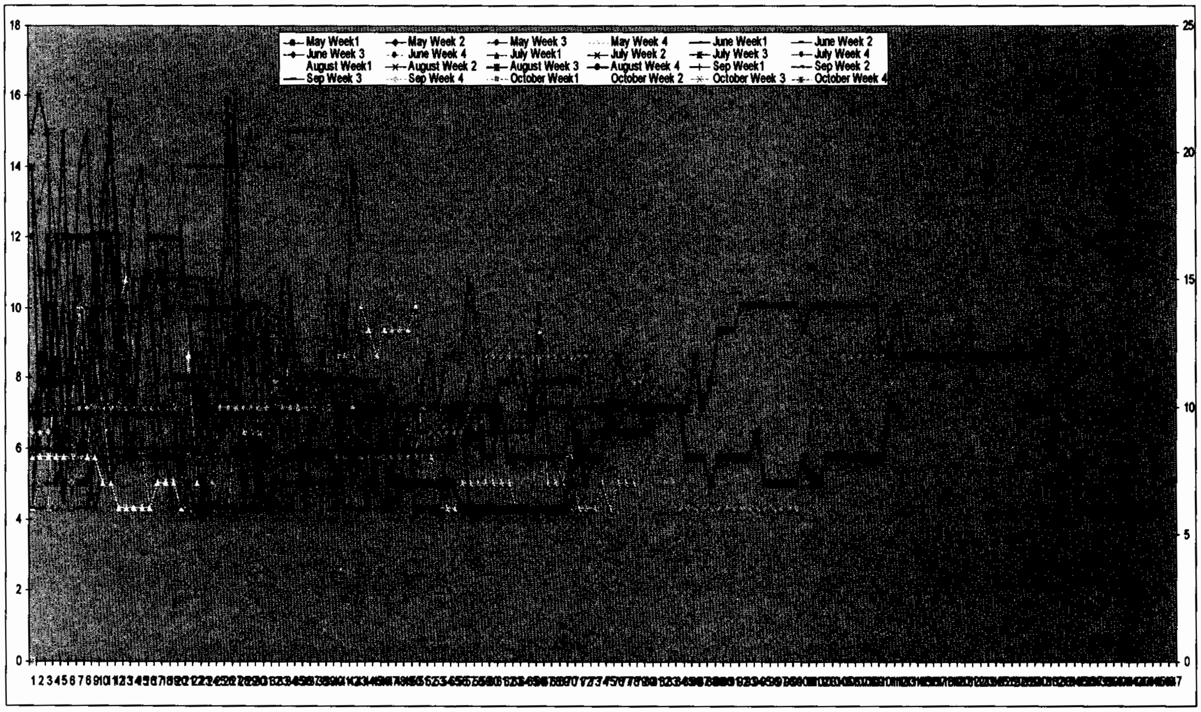


Figure 7.2: "Line Column 2 Axes" Graph

In the future we can extend the technique to more depth i.e. to view query results in the context of days.

In future the project can be extended by storing data in the form of sound, maps, pictures and clips. Moreover, data can also be store in the form of "Meta data".

The objectives stated in Chapter 1 are completely achieved. First objective was to explore useful clusters with respect to time dimension, observe the patterns and correlation between attributes as shown and explained in Figure 4.9. The second objective was to overcome the shortcomings of uni-variate and bi-variate techniques which are explained in Chapter 3. This objective is achieved by developing a proposed technique i.e. "Query Dependent Spiral and Shape Coding". The last objective was to provide an insight of data to non-IT people, which is also achieved by providing analysis of data using different colors scheme.

REFERENCES

1. Jeffrey W. Seifert "Data Mining: An Overview Analyst in Information Science and Technology Policy Resources, Science, and Industry Division"
URL: http://www.fas.org/irp/crs/RL31798.pdf Accessed on 21st Aug 2008

2. http://www.agroict.org/Agro-lnformatics.htm Accessed on 21st Aug 2008

4. Manas "How Shall we define "Knowledge"? Volume XXIV, No. 49 December 8, 1971, pp.3

5. Pulraraj Ponniah "Data ware House Fundamentals, A Comprehensive guide for IT Professionnal" John Wiley & Sons (Asia) Pte Ltd, ISBN 9812-53-012-6 pp, 409,415

6. Bill Palace, 1996 "What is Data Mining?"
http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
Acceesed on 21st Aug 2008

7. Daniel A. Keim, Hans-Peter Kriegel "Visualization Techniques for Mining Large Databases": A Comparison IEEE Transaction on Knowledge and Data Engineering, Vol. 8, No. 6, Dec. 1996.

8. http://www.iwi.uni-hannover.de/lv/seminar_ss05/bartke/iconbased.htm Acceesed on 21st Aug 2008

Appendix A

Formulae

Calculating Attribute Distance

To determine the approximate results distance between the data and query values are calculated. For the numeric type such as integer or real the distance of two values is easily determined by their numerical difference. The distance calculation yield distance tuples, which denotes distance of the data item to the query.

Calculating Combine Distance

Combine distance is the over all distance of the attributes and shows the relevancy of the attribute. It is calculated by using following formula

Combine Distance=attribut1*2 ^2+ attribut2*2^1 attribut3*2^0 -----------Equation 1

Weights are assigned to the attributes depending upon their importance or usage. Attribute1 in "equation 1" is assigned highest weight i.e. 2, "attribut2" is assigned the weight "1" and the "attribute3" is assigned the lowest weight i.e. "0".

Normalization

It is a linear transformation of the range [dmin, dmax] for each selection predicate to a fixed range i.e. 0-255. Normalization is performed so that values calculated by distance functions may be same order of magnitude. After normalization distance values are sorted into ascending order, and are represented in the form of color spiral in their respective assigned cells.

Conversion of Textual Data into Numeric Data

Following steps are carried out to convert textual data into numeric data.

Varity Name
CIM 473
BT 121
BT 121
BT 121
CIM 473
CIM 496
CIM 496

Step 1: Count every occurrence of each variety

Varity Name	Occurrence
CIM 473	2
BT 121	3
CIM 496	2

Step 2: Sort the occurrence in descending order

Varity Name	Sorting
BT 121	3
CIM 473	2
CIM 496	2

Step 3: Assign ranks to the varieties

Varity Name	Ranks
BT 121	1
CIM 473	2
CIM 496	3

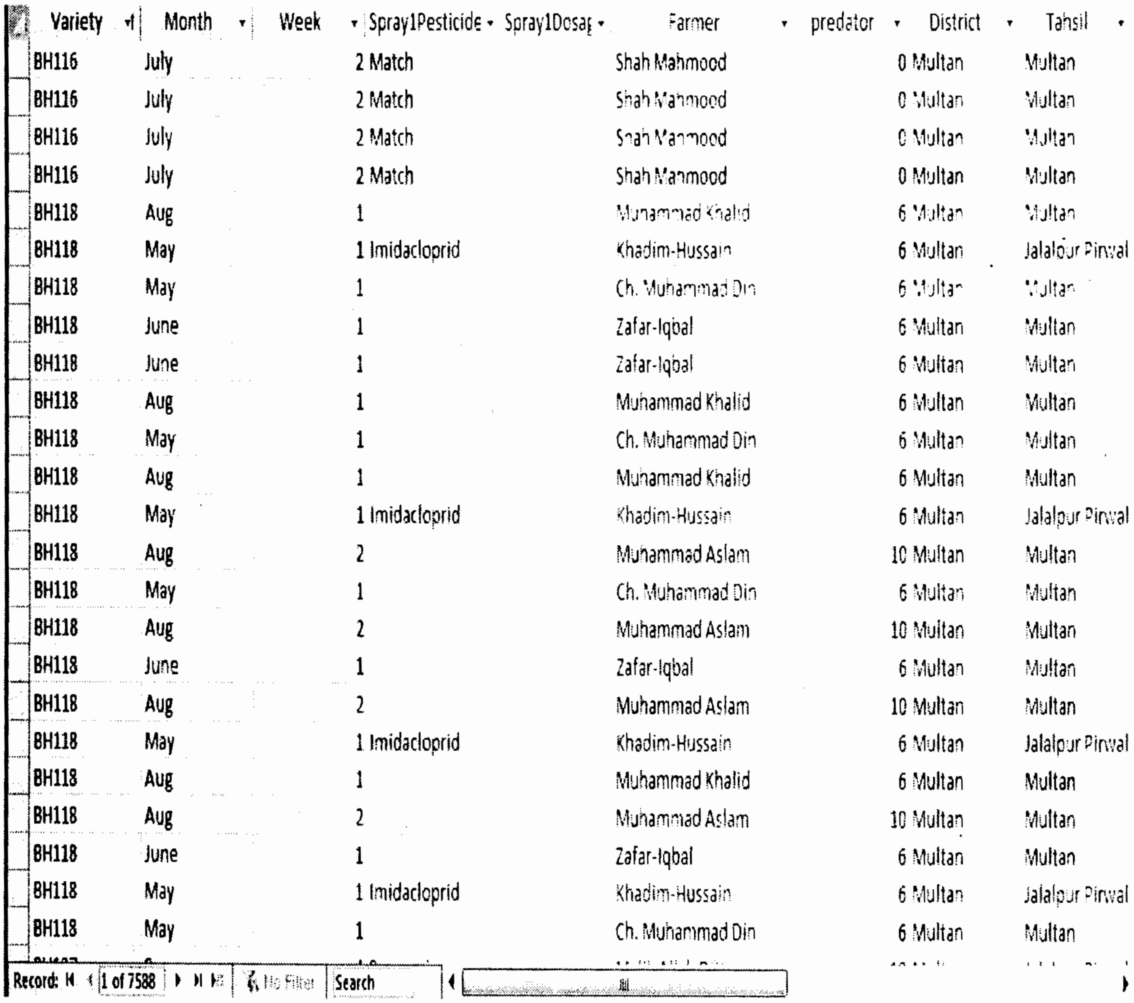
Step 4: Assign ranks to each occurrence of the varieties

Varity Name	Ranks
CIM 473	2
BT 121	1
BT 121	1
BT 121	1
CIM 473	2
CIM 496	3
CIM 496	3

Appendix B

Screen Shot

The screen shot of a table "2005_sprayclass1"from a database used in the project is shown below.



Variety	Month	Week	Spray1Pesticide	Spray1Dosag	Farmer	predator	District	Tahsil
BH116	July	2	Match		Shah Mahmood	0	Multan	Multan
BH116	July	2	Match		Shah Mahmood	0	Multan	Multan
BH116	July	2	Match		Shah Mahmood	0	Multan	Multan
BH116	July	2	Match		Shah Mahmood	0	Multan	Multan
BH118	Aug	1			Muhammad Khalid	6	Multan	Multan
BH118	May	1	Imidacloprid		Khadim-Hussain	6	Multan	Jalalpur Pirwal
BH118	May	1			Ch. Muhammad Din	6	Multan	Multan
BH118	June	1			Zafar-Iqbal	6	Multan	Multan
BH118	June	1			Zafar-Iqbal	6	Multan	Multan
BH118	Aug	1			Muhammad Khalid	6	Multan	Multan
BH118	May	1			Ch. Muhammad Din	6	Multan	Multan
BH118	Aug	1			Muhammad Khalid	6	Multan	Multan
BH118	May	1	Imidacloprid		Khadim-Hussain	6	Multan	Jalalpur Pirwal
BH118	Aug	2			Muhammad Aslam	10	Multan	Multan
BH118	May	1			Ch. Muhammad Din	6	Multan	Multan
BH118	Aug	2			Muhammad Aslam	10	Multan	Multan
BH118	June	1			Zafar-Iqbal	6	Multan	Multan
BH118	Aug	2			Muhammad Aslam	10	Multan	Multan
BH118	May	1	Imidacloprid		Khadim-Hussain	6	Multan	Jalalpur Pirwal
BH118	Aug	1			Muhammad Khalid	6	Multan	Multan
BH118	Aug	2			Muhammad Aslam	10	Multan	Multan
BH118	June	1			Zafar-Iqbal	6	Multan	Multan
BH118	May	1	Imidacloprid		Khadim-Hussain	6	Multan	Jalalpur Pirwal
BH118	May	1			Ch. Muhammad Din	6	Multan	Multan