CHAPTER 3: PATTERN VISUALIZATION AND OPERATIONS FOR NORTH EAST MONSOON RAINS OVER CAUVERY DELTA

Cauvery delta regions in Tamilnadu state of India, fed the entire state with its rich harvest once. But now a high degree of uncertainty prevails over the yielding capacity of the Cauvery delta region thanks to the fact that the river gets dried up to the trickle due to scanty rainfall. A possible solution is to predict the NE monsoon rain taking into account various parameters of weather data. The prediction can result in an n-cube dimensional data model. The on-line analytical processing (OLAP) operations and developing a multidimensional data model for the climate databases will be helpful to facilitate management decision making. For development of climatology a minimum of 30 years data is required as per World Meteorological Organization guidelines. The reliability of result is more if data for more number of years are available. OLAP operations are done on the basis of monsoon data for the past 50 years for the region. Having got the 50 years data of monsoon, a pattern is aimed at, which can attribute the monsoon’s performance. If similar patterns are found on many occurrences in the past, then the same can be used in making predictions for the future. This can be devised by OLAP operations on multidimensional data with their threshold value. In this chapter, the study of OLAP operations on the multidimensional climate dataset is elaborated and rainfall patterns are visualized.
3.1 OLAP OPERATIONS ON THE MULTIDIMENSIONAL CLIMATE DATA MODEL

Meteorologists and weather forecasters base weather prediction mainly on numerical and statistical models (Akaike 1977) apart from their subjective assessment of the dynamic atmospheric system. The history of decision support systems can be traced back to the 1960s. However, the proposal of the construction of large data warehouses for multidimensional data analysis is credited to Codd (Codd et al 1993), who coined the term OLAP for on-line analytical processing. The researchers (Widom 1995) identified several research problems in data warehousing. The data warehouse architects (Kimball and Ross 2002) provide an overview of the deficiencies of SQL regarding the ability to support comparisons that are common in the business world and present a good set of application cases that require data warehousing and OLAP technology. This study reveals that the multidimensional database model for weather prediction has been developed using meteorological data of the Cauvery delta region stations. In order to ensure that all the relevant data are utilized by the data mining techniques, it is important to make use of multi-station data. Therefore, we propose what are the possible OLAP operations on multi-station atmospheric data on this particular region.

3.1.1 MULTIDIMENSIONAL DATA MODEL

A multidimensional data model views data in the form of a data cube (Gray et al 1997; Thomson 1997). A data cube allows information to be modeled and viewed in numerous dimensions. It is defined by scope and facts. Dimensions are the perspectives or entities with deference to which an association wants to keep records. For example a
climate data warehouse in order to keep testimony of the meteorological station’s annotations with respect to the scope of time, precipitation and location track of things like area-wise excess, normal, deficient and scanty rainfall at the end of the monsoon session at which the observations are collected. Each dimension may have a table related with it, called a dimension table, which advance describes the dimension. For example, a dimension table for time possibly will restrain the attributes parameter-name and type. Dimension tables can be precise by users or experts, or robotically generated and accustomed based on data distributions.

A multidimensional data model is normally organized around a central idea, like southwest monsoon data for illustration. A fact table represents this theme (O’Brien and Marakas 2009). Facts are numerical measure. Feng et al direct us to think of them as the quantities by which associations between dimensions are analyzed. Example of essentials for a climate data warehouse consists of rainfall (daily rainfall data in millimeter), temperature (temperature data in Celsius) etc (Feng et al 2001). The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension table. Although cubes are thought of as 3-D geometric structures, in data warehousing the data cube is n-dimensional. Table 3.1 represents the 2-D data cube that is, in fact, a table or spreadsheet for the climate observations for the parameters maximum temperature, minimum temperature, humidity and dry bulb per month in the location of Karaikal. In this 2-D representation, the climate data for the Karaikal station is revealed with reference to the time dimension (organized in quarters) and the temperature dimension (structured according to the Celsius temperature).
Table 3.1. A 2-D view of climate observations

<table>
<thead>
<tr>
<th>Location = Karaikal</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (quarter)</td>
<td>Maximum</td>
</tr>
<tr>
<td>Q1</td>
<td>33</td>
</tr>
<tr>
<td>Q2</td>
<td>34</td>
</tr>
<tr>
<td>Q3</td>
<td>32</td>
</tr>
<tr>
<td>Q4</td>
<td>30</td>
</tr>
</tbody>
</table>

Now, suppose that we would like to analyze the climate data with a third dimension. For instance, suppose we would like to analyze the data on the basis of time, temperature as well as site for the locations Thanjavur and Karaikal. These 3-D data are revealed in Table 3.2. The 3-D data of Table 3.2 are represented as a series of 2-D tables. Theoretically, we may also represent the identical data in the structure of a 3-D data cube. The Table 3.2 describes the view of climate observations according to the proportions time, temperature and location. The temperature is displayed as a measure in concerned tables. The Q1, Q2, Q3 and Q4 in time column of Table 3.2 states the first quarter of the year (January to March), second quarter of the year (April to June), third quarter of the year (July to September) and forth quarter of the year (October to December) respectively.

Table 3.2. A 3-D view of climate observations

<table>
<thead>
<tr>
<th>Time</th>
<th>Location = Thanjavur</th>
<th>Location = Karaikal</th>
<th>Temperature</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Min</td>
<td>Dry bulb</td>
<td>Wet bulb</td>
</tr>
<tr>
<td>Q1</td>
<td>33</td>
<td>18</td>
<td>34</td>
<td>20</td>
</tr>
<tr>
<td>Q2</td>
<td>34</td>
<td>22</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>Q3</td>
<td>32</td>
<td>21</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td>Q4</td>
<td>30</td>
<td>19</td>
<td>30</td>
<td>24</td>
</tr>
</tbody>
</table>
Time, temperature and location are the three dimensions considered as input for the present study. It is also possible to include other dimensions like wind direction, wind speed, sea level pressure, cloud density, etc. Viewing things in 4-D becomes tricky. However, we can think of a 4-D cube as being a series of 3-D cubes. If research is continued in this way, any n-D data as a series of (n-1) – D “cubes” may be presented. The data cube is a metaphor for multidimensional data storage. The actual physical storage of such data may differ from its logical representation. The important thing to remember is that data cubes are n-dimensional and do not confine data to 3-D.

The above tables Table [3.1] and Table [3.2] show the data at different degrees of summarization. In the data warehousing research literature, a data cube such as each of the above is referred to as a cuboid. Given a set of dimensions, a lattice of cuboids can be constructed, each showing the data at a different level of summarization or group by, which is summarized by a different subset of the dimensions. The lattice of cuboids is then referred to as a data cube. It describes a pattern of cuboids forming a data cube for the dimension time, temperature, location and wind direction.

3.1.2. OLAP OPERATIONS

In the multidimensional representation, data are organized into multiple proportions, and each of which contains multiple levels of abstraction defined by theory hierarchies. This organization allows users with the flexibility to sight data from dissimilar perspectives. A number of OLAP data cube operations subsist to materialize
these dissimilar views, allowing interactive querying and study of the data at hand. Hence, OLAP provides a user-friendly atmosphere for interactive data analysis.

### 3.1.2.1 Roll-up

The roll-up operation is also called drill-up operation. It performs aggregation on a data cube, also by climbing up a concept ladder for a dimension or by aspect reduction. Figure 3.1. shows the outcome of a roll-up operation performed on the middle cube by climbing up the perception hierarchy for given locality. This hierarchy is defined as the whole order station < district < state < country. The roll-up function shown aggregates the data by rising the location ladder from the level of location to the stage of country. In other terms, rather than merging the data by state, the resulting cube groups the data by country. When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube. For example, consider a climate data cube containing three dimensions location time and temperature. Roll-up may be achieved by removing, say the individual station temperature value, resulting in an aggregation of district wise, state wise or region wise temperature value.
Figure 3.1. OLAP Operations on multidimensional climatic data
3.1.2.2. Drill-down

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Figure 3.1 shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as UTC (universal coordination time) < day < pentane (5 days) < month < quarter < year. Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month. The resulting data cube details the total temperature per month rather than summarized by quarter. Since a drill-down adds more detail to the given data, adding new dimensions to a cube can perform it. For example, a drill-down on the central cube of Figure 3.1 can occur by introducing an additional dimension, such as coastal-station.

3.1.2.3. Slice and dice

The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube. Figure 3.1 shows a slice operation where the climate observations are selected from the central cube for the dimension time using the criterion time = “Q1”. The dice operation defines a sub cube by performing a selection on two or more dimensions. Figure 3.1 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions such that location describes
either “Thanjavur” or “Karaikal” and time describes either “Q1” or “Q2” and temperature describes either “maximum” or “minimum’.

3.1.2.4. Pivot (rotate)

Pivot is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Figure 3.1 shows a pivot operation where the temperature and location axes in a 2-D slice are rotated.

3.1.3. LIMITATIONS AND CHALLENGES

Many of the individuality of OLAP systems, such as the use of a multidimensional data model and perception hierarchies, the association of measure with dimensions, and the ideas of roll-up and drill-down, also exist in prior work on statistical databases. A statistical database is a database system that is intended to support statistical applications. Similarities between the two types of systems are rarely discussed, mainly due to differences in terms and application domains.

OLAP and SDB systems, however, have individual differences (Shoshani 1997). While SDBs tend to focus on performance metrics, OLAP has got additional facilities like pivot, slice, dice, drill-down and roll-up operations which will help in certain specific applications like the display of meteorological data for climate analysis as well as environmental data analysis to arrive applications where multi correlation analysis is
required. Finally, unlike SDBs, OLAP systems are planned for handling huge amounts of data efficiently.

Information processing using OLAP, based on queries, can find useful information. However, answers to such queries reflect the information directly stored in databases or computable by aggregate functions. They do not reflect complicated patterns or regularities hidden in the database. Therefore, information processing is not data mining. The functionalities of OLAP and data mining can be viewed as disjoint. OLAP is a data summarization or aggregation tool that helps to simplify data analysis, while data mining allows the automated detection of implicit patterns and attractive knowledge hidden in large amounts of data. OLAP tools are embattled toward simplifying and supporting interactive data analysis, whereas the goal of data mining tools is to computerize as much of the procedure as possible, while still allowing users to guide the procedure. In this sense, data mining goes one step beyond established on-line analytical processing. Yet according to this vision, data mining covers a much broader spectrum than simple OLAP operations because it achieve not only the data summary and assessment but also association, classification, prediction, time-series analysis, and other data analysis tasks.

3.2. PATTERN VISUALIZATION ON METEOROLOGICAL DATA

Regional rainfall forecast is an important task for meteorologists. While statistical methods are usually in vogue, for this, the concept of data mining is applied to see the
efficacy. The objective of this work is to develop a forecast method for the rainfall of the Northeast (NE) season over the Cauvery delta region of South India. There are three main parts in this work. The first part is in loading the meteorological data into centralized server so that it can be extracted and filtered according to needs. Secondly, graphical visualization has been applied by using gnuplot and finally the association rules have been mined from those former outputs. From these results the monthly and seasonal rainfall for the future can be assessed.

Cauvery river flows in the peninsular India from west to east, which is very vital for agricultural production in Tamilnadu state of South India. Besides this region gets the main seasonal rainfall during the northeast monsoon months of October to December. In summer months from May to August as well as September also thunderstorm rains occur here. In India the period from mid June to September is Southwest (SW) monsoon or summer monsoon. With the Cauvery waters getting depleted in the downstream regions due to the construction of many dams, the Cauvery delta farmers have to depend upon the seasonal rainfall to a great extent since the nineties of the last century. In the late eighties many townships also came up in delta region. Hence seasonal rainfall estimate over the delta and its prediction will go a long way to help the agriculturists, hydrologists and water managers.
With the vastly varying terrains in India, no clear cut correlation between Southwest (June to September) and Northeast (October to December) monsoon rainfall could be established though some attempts on possible link with global phenomenon have been made recently (Nachiketa Acharya et al 2011; Raj 2003; Samuel Selvaraj and Raajalakshmi Aditya 2011,2012). There is a spatial variability in monsoon performance within Tamilnadu state of India. Hence at sub regional level where there is fairly a uniform terrain; study has to be conducted with latest tools and methodologies to predict the Northeast (NE) monsoon rainfall from the earlier rains. Such a study has been attempted for the Cauvery delta basin. There is no earlier study specifically for this region. Usually the sowing operations start during July to August. The SW monsoon is about 35% of annual rainfall over this region, but the NE monsoon is about 60% of annual rainfall that is found to show large variations. Hence the water managers have to adjust the dependable rainfall realized by August with the operation schedule allowing for certain threshold rainfall by November end. This needs a model for prediction of NE monsoon rain with good chances of success. Graphical mining and data mining technique are some computer-based tools (Dunham 2005; Han and Kamber 2008), which have been successfully used for predicting parameters in certain fields in the recent past. It is being tried here for rainfall prediction for the first time.

3.2.1. DATA COLLECTION

Monthly rainfall data for available stations of the composite Thanjavur district of Tamilnadu in India in downstream Cauvery basin for 25 years since 1985 have been used
for analysis. The area covered is shown in Figure 3.2. The sample stations in the Figure 3.2 are Grand Anaicut(1), Thanjavur(2), Papanasam(3), Kumbakonam(4), Valangaiman(5), Nannilam(6), and Nagapattinam(7) which is close to Karaikal of Pondichery Union Territory of South India. The monthly rainfall data is collected from the Department of Public Works Department of Irrigation section, Thanjavur Circle, Tamilnadu, India.

![Figure 3.2. Rain gauge stations in the Cauvery delta](image)

1. Grand Anaicut
2. Thanjavur
3. Papanasam
4. Kumbakonam
5. Valangaiman
6. Nannilam
7. Nagapattinam
3.2.2. PROPOSED APPROACH

A few attempts on meteorological applications using data mining techniques have been made (Chen 2007; Cofina et al. 2003; Dhanya and Nagesh Kumar 2009). However, forecast for the monthly and seasonal rainfall for this Cauvery delta region is needed. This has been attempted using the techniques of association rules in graphical mining, it is necessary to filter the raw data and convert into the format that is accepted by visualization tool, gnuplot (Janert 2010). The proposed approach is described in Figure 3.3 that emphasizes the steps involved in the knowledge discovery using graphical mining of data mining technique.

Figure 3.3. Steps involved in the knowledge discovery using graphical mining
The procedure for the knowledge discovery using graphical mining is as follows:

Step 1: Loading climatic data to SQL Database.
Step 2: Filtering the noises and unwanted data from the dataset.
Step 3: Transform data as required for the visualization software.
Step 4: Generating graphs using the visualization tool.
Step 5: Studying of patterns on the generated results.
Step 6: Mining rules from those patterns.

3.2.3. RESULTS AND DISCUSSIONS

Using the graphical mining and data mining technique, rainfall realizable in the NE monsoon was assessed from the rainfall realized during Southwest monsoon months of June, July and August months with association rules. The running five-year mean rainfall is visualized using Gnuplot tool (Janert 2010) for different stations in Figures 3.4 to 3.10. In these figures the total monthly rainfall of June, July and August months of SW monsoon, September month of SW monsoon and October, November and December months of NE monsoon have been described. The total rainfall of the Southwest monsoon months and Northeast monsoon months were compared and the total monthly rainfall within the Northeast monsoon months were also compared.
Figure 3.4. Five year mean rainfall of Grand Anaicut

Figure 3.5. Five year mean rainfall of Kumbakonam
Figure 3.6. Five year mean rainfall of Nagapattinam

Figure 3.7. Five year mean rainfall of Nannilam
Figure 3.8. Five year mean rainfall of Papanasam

Figure 3.9. Five year mean rainfall of Thanjavur
After the analysis, the following rules are mined out from the above graphs:

\[
\text{RF} \ (X, \ "\text{SW}_C < \text{SW}_P") \Rightarrow \text{NE}_C (X, \ "\text{N}_E\text{P}_C") \\
\text{RF} \ (X, \ "\text{OCT} < 200\text{mm}") \Rightarrow \text{NOV} \ (X, \ "\text{OCT} \text{R/F} + 50\text{mm}") \\
\text{RF} \ (X, \ "\text{OCT} > 200\text{mm}") \Rightarrow \text{NOV} \ (X, \ ">400\text{mm}")
\]

The equation 3.1 describes the rule that when the total rainfall (RF) of current year of South West (SW\text{C}) monsoon is less than the previous year of South West (SW\text{P}) monsoon then the rainfall (RF) of current year of North East (NE\text{C}) monsoon is greater than the previous year of the total rainfall (RF) of North East (NE\text{P}) monsoon. The equation 3.2 describes the rule that when the total Rainfall (RF) of October month is less
than 200mm then the total rainfall of the successive November month is most probably
the actual October month rainfall with an excess of 50mm. The equation 3.3 describes the
rule that if the total rainfall of October month is greater than 200mm then the total rainfall
of the successive November month will have the precipitation of more than 400mm.

3.3. SUMMARY

The study reveals that in the Cauvery delta basin, the rainfall pattern of many
stations obey the association rule except Nagapattinam. Nagapattinam being coastal in
nature may perhaps be the reason. The rule suggests a correlation between the Southwest
and Northeast monsoon rains over this area when considered in relation to previous years
performance. It is also possible to assess the November month rain from the rains of
October using the same rule. Frequent pattern mining leads to the discovery of
associations and correlations among rainfall occurrence in large climate data set. This
methodology can help to forecast rainfall during the months of Northeast monsoon for
the Cauvery delta region of Thanjavur district of South India.