Chapter 3

Technologies Use for Developing Newspaper Circulation and Sales Forecasting System

3.1 Introduction

Previous chapter describe about available literatures related to the research domain. This chapter mainly concern about technologies that use for developing the Newspaper Circulation and Forecasting System. Mainly used PHP server side scripting language for developing the Circulation System and MySQL Relational Database Management System (RDBMS) for maintain the system database. WAMP Server is use as web server to run this application. This research used Weka data mining software for forecast the net sales of the newspapers and identifying sales patterns. Weka contains several data mining algorithms such as preprocessing, Time series analysis algorithms, evaluation algorithms.

3.2 About PHP

PHP originally called Personal Home Page but now it's called PHP Hypertext Preprocessor [21]. PHP is an open source server side scripting language with embedded with HTML. It's used to create dynamic web pages [22]. PHP is originally created by Rasmus Lerdorf in 1994 and current major version is PHP5 [21].

PHP programming language has many advantages such as

- PHP is embedded with HTML through PHP special tags there for developers can used both HTML codes and PHP codes in one interface. Its helps to reduces number of codes in HTML [22].

- PHP codes are executed on the server therefore, client cannot see the PHP source code [22].

- PHP can communicate through network by using IMAP, SNMP, NNTP, POP3, HTTP [22].

- PHP is high performance language that can save millions of hits per day by using single inexpensive server [21].

- PHP provide interfaces to integrate many database systems apart for MySQL such as Oracle, PostgreSql, etc; in addition using ODBC user can connect any other databases that support ODBC driver. Therefore, Microsoft products also used in PHP [21].

- PHP have many built in libraries that help to performing many web related activates. Such as send emails, generate PDF documents, etc; [21].

- PHP can download by free of charge [21].

- PHP is easy to learn [21].

- PHP is strong object oriented support language [21].

- PHP support many operating systems therefore, any application written in PHP works many operating systems without any modification [21].

- PHP source code freely available therefore developers modify or add to the language [21].

Consider above advantages this research used PHP scripting language for develop the web base Circulation System.

3.3 About MySQL

MySQL Relational Database Management System (RDBMS) is a fast and robust database management system that invented in 1979 but it's publicly available in 1996. It is most popular open source database that's awarded several times in Linux Journal Readers' Choice award. MySQL is available under free open source license and commercial license. MySQL uses Structured Query Language (SQL) and it ensure the efficient data store, search, sort and retrieve data. It assure the concurrently access to the multiple users as well as ensure the fast and secure access (authorized users) for the users, because MySQL consider as multiuser, multithreaded server [21].

MySQL RBDMS have some advantages consider the other databases such as Oracle, MS SQL Server and PostgreSQL such as [21];

- High performance [21].
- MYSQL and PHP work with many major operating systems as well as minor ones [21].
- MySQL is available free open source license or low price commercial license. Therefore, software available in low cost [21].
- MYSQL is easy to used and setup [21].

Consider above factors this research used MySQL RDBMS as database for Newspaper Circulation system.

3.4 WAMP Server

WAMP Server is a free open source application that works on windows. It's normally used in web server environments. It provides key essentials of web server, such as operating system, database, web server and web scripting software. Combination of these elements is knows as server stack. Microsoft windows works as a operating system, Apache works as a web server, MySQL is the database and PHP, Python, or PERL are the scripting languages [23].

Therefore, WAMP Server is use as web server to run this Newspaper Circulation system.

3.5 Preprocessing

Data Preprocessing is one of the important steps in data mining. It's used to remove unnecessary information that contain in data set that use for forecasting. It also removes or minimizes some noises, incomplete and inconsistent data. Therefore, some time add some data for missing values (add mean value related to the missing field). Later this preprocesses data set further processed by data mining algorithm [13]. Sometime in data set contain data it's not relevant to the common data Patten. As a example newspaper sales suddenly increase in one day due to some reasons such as bomb explosion, some marketing actives, etc; therefore, this increase not consistent in longer time period. This situation consider as outliers. Depend on the interest of the user outliers can remove or not in data preprocessing. Identify the outliers first similar data values need to be group and then it can identify by human or computer. Remove outlier's first need to be calculating the mean value and standard deviation. In draw the normal distribution carve and considers the 68% of measurements that include the range of $\mu-\sigma$ to $\mu + \sigma$. Remove out of range values not between $\mu-\sigma$ to $\mu + \sigma$ and insert mean value when users' interest about outliers [13].

3.6 Forecasting Techniques

Newspaper sales forecasting data present in series of time depend data points. Statistical techniques that called Time Series Analysis using to modeling and explaining time dependent set of data points. Based on the past events Time series forecasting generate model for forecasts future events. As well as data has natural temporal order in time series data rather than usual data mining or machine learning applications. Consider the newspaper sales also have natural temporal order. Usual data mining applications not much consider ordering of data points within the data set and each data points is independent point of the concept to be learned [24].

3.7 WEKA Data Mining Software

Weka (Waikato Environment for Knowledge Analysis) is data mining software developed at University of Waikato, New Zealand by using Java. This software containing collocation of machine learning algorithms that using data mining such as data pre- processing, classification, regression clustering, association rules and visualization [25]. From Weka version 3.7.3(=> 3.7.3) this data mining software facilitated to developed, evaluated and visualized the time series analysis based forecasting models. Using data mining approach weka time series framework transforming the data into form that standard propositional learning algorithms can process. Weka does this process by removing the temporal ordering of individual inputs values by encoding the time dependency using additional input fields. This field called as "lagged Variables". Various other fields are calculated by automatically and allow the algorithms to model treads and seasonality. To learn the model any of the weka's regression algorithms can be used after the data has transformed. Weka allowed many methods that capable of predicting continuous target, mainly multiple linear regressions, non linear methods mainly support vector machines for regression trees and model tree. Decision trees with linear regression at the leaves consider as model trees. Weka is containing multiple classifier functions that used for forecasting [24]. This research will be used following four classifiers for forecasting the newspaper sales. Such as

- Gaussian Process
- Multilayer perceptron,
- Linear Regression
- SMO Regression

- **Gaussian Process-**

  Regression without hyper parameter tuning this classifier function implements for Gaussian process. Use of global mean/mode values this classifier filled the missing values as well it help to choose the appropriate noise lave easier [26].Gaussian Process is represent as equation (3.1):

  $$X_{t,}t\epsilon\, T, \tag{3.1}$$

- **Multilayer Perceptron-**

  This algorithm uses back propagation to classify instances and this network can developed by manually or algorithm or both. During the training time this network can be monitored and modified [27]. Its shows in equation (3.2):

  $$f(a) = f(a)\grave{\cdot}(1 - f(a)), \tag{3.2}$$

- **Linear Regression-**

  This classifier uses for liner regression for prediction and this function allowed to use weighted instances [28]. Its repercented by equation (3.3), in hear Y denoted by dependent variable and X denoted at experimental variable.

  $$Y=a+bX, \tag{3.3}$$

- **SMO Regression-**

  This function is support for vector machine for regression [29]. SMO Regression shows equation (3.4) and input vector shows in x_i, $y_i \epsilon\{-1, +1\}$ is binary values related to it, as well as α_i are Lagrange multipliers.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} y_i y_j K\left(x_i' - x_j\right) \alpha_i \alpha_j \qquad (3.4)$$

3.8 Evaluation Techniques

After forecast results generated this predicted results must evaluate with actual result and select most accurate forecasting algorithm for newspaper sales forecasting. Assure the accurate of forecasting results have been evaluated and analysis following techniques. These evaluation algorithms are already build in Weka data mining tool. Evaluation techniques are

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Relative Absolute Error (RAE)

- **Mean Squared Error (MSE)-**

  This technique measure how closer data point to the fitted line. Measures the distance vertically from the point to the corresponding Y value on carve fit/the error and square the value from every data point. Then all values of data points add and divided by number of points minus two. Especially the squaring is use for stop cancelling negative values from positive values. The smaller the MSE means accuracy of outputs is high [30]. Equation (3.5) illustrates the Mean Squared Error.

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i' - Y_i\right)^2 \qquad (3.5)$$

$\acute{Y}$ Vector of n Predictions and Y is is the vector of observed values corresponding to the inputs which generated the predictions.

- **Mean Absolute Error (MAE)-**

Its measure the forecast data sets average size of the error without considering their directions. It helps to measures the accuracy of continuous variables. If MAE values in liners score its means all individual differences forecast data are weighted equally in the average [31]. Mean Absolute Error figured out in equation (3.6) and f_i is represent prediction and y_i is represent actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \qquad (3.6)$$

- **Root Mean Squared Error (RMSE)-**

This technique takes square root of the mean square error. Its easily interpreted statistic, when it has same units as the quantity plotted on the vertical axis as well it directly shows in terms of measurement units and its measure more fits than the correlation coefficient. It's also observed variation in measurements of a normal point [30]. It also measures the average magnitude of the error. As well its takes difference between forecast and actual values and calculated each squared value. Then calculated averaged over the sample and takes square root of the average. RMSE normally give high weight to large errors. Therefore, RMSE can used for identify the large errors [31]. Equation (3.7) demonstrates the Root Mean Squared Error.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\acute{Y}_i - Y_i)^2}{n}} \qquad (3.7)$$

$\acute{Y}_i$ Predicted values for t times of a regression dependent variable Y_i computed for different predictions of n.

Mean absolute error and Root Mean Squared Error can both used to identify the error variation in the forecasts data set. Mostly RMSE value is larger or equal to MAE value. If values are equal all the errors are same weight but if values have difference it mean the variance of the indicial errors are high [31].

- **Relative Absolute Error (RAE)-**

  Its relative work as simple predictor and its takes average of the actual values. In hear error is considered as the total absolute error instead of the total squared error. Find out the relative absolute error it takes the total absolute error and normalized [32]. Equation (3.8) illustrate Relative Absolute Error and V denote as some given value and its approximation.

$$\eta = \left|\frac{v - v_{approx}}{v}\right| \tag{3.8}$$

3.9 Summary

Newspaper circulation and forecasting system has two main modules such as circulation module and forecasting module. Circulation module contains three sub module for handling newspaper issue, return collection and payments. Develop Circulation module this research used PHP server side scripting language. MySQL RDBMS is used for database and WAMP server is used for web server. These tools are mostly open source and support to work in many operating systems. Forecasting module basically develop by using WEKA data mining software. This software containing collocation of machine learning algorithms that using data mining such as data pre- processing, classification, regression clustering, association rules and visualization. From version 3.7.3 Weka allowed time series analysis base forecasting. First data set need prepossess by using Weka prepossessing algorithm and remove garbage values then apply several classifiers (SMOreg, Liner regression, etc ;) and forecast the values. Then use of evaluation techniques (MSE, MAE, etc ;) find out accurate forecasting algorithm for used. Next chapter describe how adopt technologies solve the research problem.