Pattern Mining in Telecom Data

Ali Daud[1], Muhammad Akram Shaikh[1], and Faqir Muhammad[2]
[1]*Department of Computer Science &Technology, Tsinghua University, Beijing, China*
[2]*Department of Mathematics & Statistics, Allama Iqbal Open University, Islamabad, Pakistan*
ali_msdb@hotmail.com, akramshaikh@hotmail.com, aioufsd@yahoo.com

Abstract

This paper is concerned with the trend analysis for finding and predicting interesting patterns on telephone calls data of Pakistan Telecommunication Company Limited. Recent years, mining interesting patterns from large scale repositories has become one popular research issue. Telecom industry desires to invest in new technologies for Customer Relationship Management, for example, to retain their customers by performing competitive analysis. Our work in this paper is aimed at identifying useful trends in telephone calls by using statistical techniques. The trends and forecasts can be used for making future decisions for business. By using autoregressive moving averages model fruitful patterns are obtained. Empirical analysis for forecasting data trend shows that the predicted results match well with the actual data from real world. In addition, by doing box plot analysis it is found that the maximum number of calls was made on Sundays. Analysis shows that the found patterns are helpful for assisting decision making.

1. Introduction

Pattern analysis, trend finding and forecasting on time series data is an appealing research area for business companies like telecom [3, 13], finance [19] and large scale fast food restaurants [18]. Our data is from Pakistan Telecommunication Company Limited (PTCL) of the city of Faisalabad. This Telecom Company gathers tremendous amount of data in its repositories. The data includes calls detail data, which describes calls that traverse telecommunication networks, network data, which describes hardware and software components in the network as well as network traffic, sales data, which shows the revenue earned by a company, and data about the customers who are using telecommunication services.

Telecommunication companies are investing in new technologies of trend analysis, forecasting, and data mining. Objectives include reviewing their business progress, identifying maximum and minimum network usage days to improve their network services and launch some special offers, finding fraudulent phone calls to remove loopholes in their revenue collection, performing churn analysis to find why their customers are switching to some other companies by using customers profiles, monitoring sales for revenue collection and maximizing their profits. We observed that tariff policies also play an important role in trend finding and forecasting.

Our main focus is to find some useful trends in call details data of Telecom Company. We are interested in finding the behavior of this data by applying time series models and using trend finding utilities. We investigate telephone calls on different days of week, month, and their timings. The following three different types of call data were used for study 1) Number of calls on weekdays of April 2005 2) Number of calls on each day of April 2005 3) Calls of different durations measured in minutes of April 2005.

Specifically, we conducted the trend analysis in the following steps. First, required variables were extracted form call details data collected during business operations of the company. Then the extracted variables from PTCL databases were loaded into Teradata data warehouse. After that, to make data quality better and convert data in the form required for trend finding and forecasting, data cleaning and transformations were performed according to requirements.

When the data set was ready for trend analysis we first checked whether the data is stationary or not. Stationary is best assessed by making a time series plot, if it fluctuates around a constant value and the spread is almost same every where then the data is said to be stationary. Data set used here was found stationary by using trend analysis utility of mini-tab software. We employed stationary autoregressive moving averages model for time series. We have found that ARMA (2, 2) is the most appropriate model for data and was thus used to forecast calls per day for the next thirty days with maximum and minimum call limits. From

forecasts it became clear that the company's business was going well and no unexpected rises or fall were found in number of calls per day. By using Box-Plots, we found a trend that Sundays had higher number of calls than other weekdays. Thursdays, Fridays and Saturdays have more dispersed distribution of number of calls above the median. Mondays have more disperse distribution for number of calls below the median. However we can clearly see that maximum number of calls was made on Sundays due to special tariffs offered by the company. We have also found that during the middle ten days of the month maximum number of calls was made. So the network usage and customer likeliness of calls in a month are middle ten days. We must keep network in the best condition for at least those ten days to provide better services to the customers.

Our main contributions in this paper include: 1) several useful trends found from telephone calls data of PTCL, Faisalabad. These trends can be used to assist higher level management to make future business/management decisions and 2) Experiments and analysis also suggested that the time series ARMA model is useful for trend finding and forecasting on PTCL data.

The paper is structured as follows. Section 2 introduces related work. Section 3 describes data preprocessing. Section 4 explains the process of model fitting. Section 5 presents experimental results and discussions. Conclusion is given in section 6.

2. Related work

A few efforts have been placed for trend analysis and forecasting on telecom data. For example, Hilas et al [13] used different forecasting models including ARIMA for monthly outgoing calls in a university campus. Ait-Hellal [3] presented a robust identification and employed this in analyzing telecommunication traffic data by using ARMA processes.

Considerable work has been done for trend analysis in other fields. For example, Bhattacharyya et al [18] fitted time series Autoregressive Integrated Moving Averages (ARIMA) model and showed the daily demand trend of perishable ingredients for the franchise. By using their results the franchise can maintain their inventory in future. Onado [24] used ARIMA models to forecast the quality of French cheese.

Tse [27] tested the pattern of the real estate prices empirically by employing the ARIMA analysis. ARIMA was used to model the linear pattern of oil price time series [26]. A Generalized Space-Time ARMA Model was used for Unemployment Analysis [11]. Channouf [9] developed and evaluated time-series models including ARIMA of call volume to the emergency medical service. Allard [2] used time series analysis for trend finding in infectious disease surveillance.

Poddig and Huber [25] found that ARMA-models seem to be valuable forecasting tools for predicting turning points. Madden and Tan [22] compared different linear models for forecasting and concluded that for the short horizon forecasting (1-period ahead), the statistics suggest that the ARIMA is best model. Financial time series forecasting is done by combining ARMA and GRNN models [19]. Huang and Shih [15] aimed at SHORT-TERM load forecast at predicting system load over a short time interval and assured performance of ARMA model, improving the load forecast accuracy significantly.

ARMA models have been used for characterizing LAN traffic [7, 10], video codec sources in ATM networks [12, 23], and ATM traffic [16]. In the latter, the ARMA traffic model has been shown to perform better than a two-state MMPP (Markov Modulated Poisson Process). Comparisons with other traffic models can be found in [21]. In [4], a robust control approach has been used for the design of rate-based flow controls for ABR traffic, in the presence of exogenous traffic modeled by ARMA processes. Yang [28] used ARMA and Hopfield model for intrusion detection in network time series data.

3. Preprocessing

Telephone calls data of Pakistan Telecommunication Company Limited (PTCL), Faisalabad has multiple dimensions such as item called, time called, and location called. One can do trend analysis and forecasting on the data according to different items called by customers, different locations called by the customers and the number of calls made with respect to time. Large amount of data length can be suitably met by using the hierarchical organization of dimensioned data that is often employed in data warehouses [8]. Nowadays, many companies are willing to improve their business by performing Customer Relationship Management (CRM). So with the emergence of critical future decision making they are investing in data warehouses and data-marts to keep their data safe and usable for decision-making [17].

3.1. Data extraction and loading

Data variables such as telephone number, call date, call start time, and call end time were extracted from

repositories of the company. Data variables were then stored in .txt file and then loaded into Teradata data warehouse by using fast data loading script for the purpose of data cleaning and transformations.

3.2. Data cleaning

Cleaning data from impurities is an integral part of data processing and maintenance. Muller and Freytag [20] presented a survey of data cleaning problems, approaches, and methods.

In this work, we performed data cleaning as follows. First, the calls data that were made in April 2005 were selected for study. Secondly, repeated calls with same attributes were removed to minimize redundancy. Thirdly, calls having time less than four seconds were removed because PTCL, Faisalabad does not include calls of duration less than four second in billing. Such calls are usually considered as drop calls.

3.3. Data transformation

The goal of data transformation is to make the calls data well ready for analysis. Calls data was available with call start and end time. We calculated the duration of each call by taking difference of end time and start time. The calls were then sorted and summed according to each day of month (1, 2, 3,…,30). This data was used for model fitting and predicting number of calls. In addition, calls were also sorted and summed according to the weekdays (Monday, Tuesday… Sunday) and the data was used to plot box plot to find weekly trends in the data. Finally, calls were sorted and summed according to number of calls per minutes (1, 2, 3,…,30) and the data was used in trend analysis for number of calls per minutes.

4. Our approach

Different time series models can be employed on data depending upon the kind of data like autoregressive, moving averages, autoregressive moving averages, also called Box-Jenkins model [5]. Each model is based on some assumptions. If the data does not fulfill model requirements the model cannot be used. We checked if the data is stationary or not. We plotted a time series graph for numbers of calls on each day and found that the data was almost stationary. This is because there was no huge unexpected rise or fall in trend line visible in fig. 1.

Now some stationary time series model has to be selected to fit on the data. The data was subjected to trend analysis and model building by applying univariate Autoregressive Moving Averages (ARMA) [6] model using mini-tab software. The model ARMA assumes that time series is stationary. Box and Jenkins recommend differencing non-stationary time series one or more times to achieve stationary data. Doing so produces an ARIMA model, where the "I" standing for "Integrated". Since we checked and found that data was stationary so there was no need to take differences.

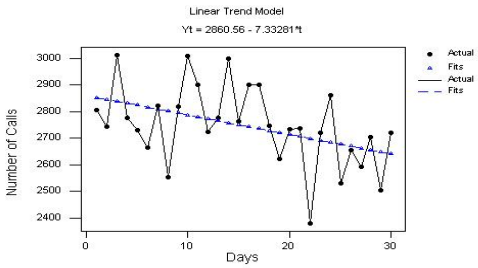

Figure 1. Trend analysis of calls per day

The Box-Jenkins ARMA model is a combination of the AR and MA models.

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q} \qquad (1)$$

where the terms in the equation have the same meaning as given for the AR and MA model. If this process is stationary then it must have a constant mean μ for all time periods. Box-Jenkins models are quite flexible due to the inclusion of both autoregressive and moving average terms.

There are three primary stages in building a Box-Jenkins time series model. In model identification the first step of developing a Box-Jenkins model is to determine if the series is stationary and if there is any significant seasonality that needs to be modeled. We have checked that and there was no significant seasonality have to be addressed, the next step is to identify the order (i.e., the p and q) of the autoregressive and moving average terms.

From table 1 it is clear that at AR (2) the probability is maximized. From fig. 2 we can see that after the second lag PACF is rising so the order of AR is 2.

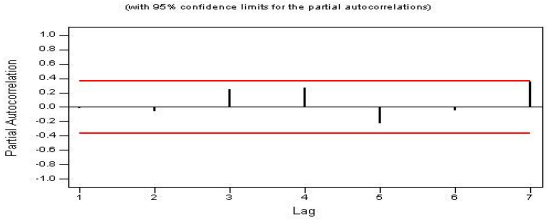

Figure 2. PACF of residuals for calls per day

The autocorrelation function of a MA (q) process becomes zero at lag q+1 and greater, so we examine the sample autocorrelation function to see where it

essentially becomes zero. We do this by placing the 95% confidence interval for the sample autocorrelation function on the sample autocorrelation plot. Minitab software generated the autocorrelation plot with mentioning confidence interval.

Table 1. AR (p) Process

ARMA	Coefficient	SE Coefficient	p-value
AR1	0.0867	0.1946	0.660
AR2	0.0096	0.1950	0.961
AR3	0.3102	0.1942	0.123
AR4	0.2594	0.2092	0.226

From table 2 it is clear that at MA (2) the probability is maximized. From fig. 3 we can see that after second lag ACF was rising so the order of MA is 2. So ARMA (2, 2) was found the best fit on data.

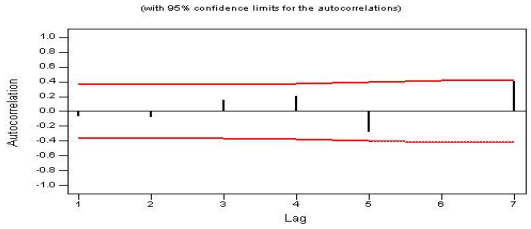

Figure 3. ACF of residuals for calls per day

Table 2. MA (q) Process

ARMA	Coefficient	SE Coefficient	p-value
MA1	-0.1752	0.1925	0.371
MA2	0.0088	0.2024	0.966

The main approaches for model estimation of Box-Jenkins models are non-linear least squares and maximum likelihood estimation [1]**.** Maximum likelihood estimation is generally the preferred technique. Estimating the parameters for the Box-Jenkins models is a quite complicated non-linear estimation problem. For this reason, the parameter estimation should be left to a high quality software program that fits Box-Jenkins models.

We used mini-tab software for model estimation using graphs. Since model validation assumptions for a stationary univariate process were confirmed and found satisfactory, so proposed model was, hence confirmed to be appropriate for telephone calls data.

5. Experimental results and discussions

We used telephone number, call date, call start time, and call end time as variables in our experiments. By using the mentioned variables three relations were created in the data warehouse for trend finding and forecasting. The first relation consists of number of calls per day for the month. This relation was used for model determination, trend finding and forecasting. The second relation consisting of number of calls in week days (Monday, Tuesday,…,Sunday) of month, and was used for trend finding in weekdays. The third relation containing calls of different durations measured in minutes (1, 2, 3,…,30) for a month and was used for finding trends in call durations.

5.1. Number of calls per day

Fig. 4 shows the time series plot with stationarity and forecasts of calls per day for the next 30 days. It also shows forecasts with maximum and minimum calls limits.

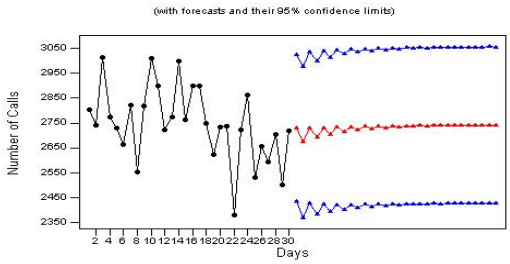

Figure 4. Time series plot for number of calls per day with forecasts

From fig. 4 we can see that forecast of number of calls for the next month is also stationary because there are no sudden rises or falls in the forecast. This also confirms that selected model is appropriate for the given data. We can see that our network usage or number of calls is continuously increasing, which concludes that the business of PTCL is going well.

Hilas et al [13] used ARIMA for monthly outgoing calls in a university campus and used outcome to predict future demands for telecommunication network of the university. ARIMA model is also used for trend finding and forecasting by [2, 11, 24, 25, 26, 27] and they all found useful results by using this model.

Fig. 4 also shows that during first 10 days of a month normal number of calls were made during the middle 10 days of a month maximum number of calls were made and during the last 10 days minimum number of calls were made. So, the network usage and customer likeliness of making calls in a month are middle 10 days. The network must be kept in good condition for at least those ten days to provide better services to the customers and if there are problems like dropped calls or poor signals during those days the

company must try to get rid of them quickly to earn maximum revenue and satisfy customers.

We can get better results from ARMA model by putting different values of p and q for different data values like here we checked from AR (1), AR (2), … , AR (4) and MA (1) MA (2), MA(3) and found that ARMA (2,2) was best fit for the data. We recommend ARMA model for this type of stationary data for trends finding and forecasting in future.

5.2. Number of calls on weekdays

Visual representation of number of calls for weekdays is represented by the box plot in fig. 5.

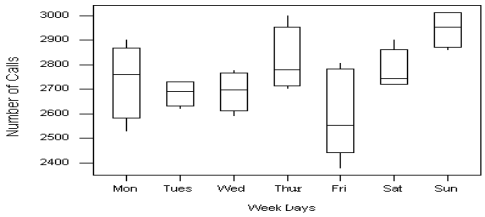

Figure 5. Box plots for number of calls on weekdays

Box-Plot provides information on dispersion by using quartiles. It is clear from fig. 5 that Thursday, Saturday and Sunday have higher number of calls than other weekdays. Thursday, Friday and Saturday have more dispersed distribution of number of calls above the median. Monday has more disperse distribution for number of calls below the median. Tuesday, Wednesday and Sunday have a little less dispersed distribution for number of calls. But we can clearly see that maximum number of calls was made on Sundays. PTCL offers half rates on Nation wide dialing on Sundays probably this was the main reason of maximum number of calls on Sundays. Company decision makers can provide similar kind of offers on some other days or special tariffs for night timings in order to increase the business of company.

5.3. Calls of different durations in minutes

Time duration of calls was different; say 1.3 minute, 2.5 minute but we rounded them off toward larger number as 1.3 to 2 and 2.5 to 3 minutes. To find trends in number of calls per minute trend analysis utility of Mini-tab software was used.

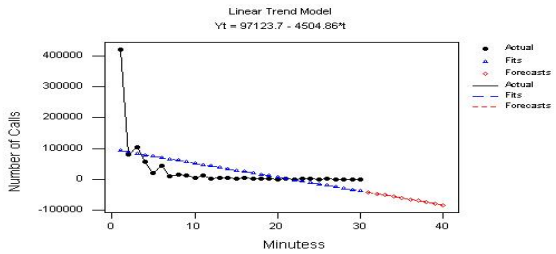

Figure 6. Trends analysis for number of calls per minutes

From fig. 6, it is clear that one-minute calls are above 400000 in number and then for second minute it drops to less than 100000 and then move slightly up from 100000 calls. After that trend line drops continuously and forecast for values from 30 to 40 shows continuous decreasing trend. Minutes of calls can never be negative so we will consider trend line of fig. 6 only above 0 (zero) minutes, trend line clearly shows that as time duration of calls increases number of calls per minute decreases. In nation wide dialing (NWD), maximum calls are of one minute, so we conclude that long distance calls are always of less duration say maximum number of calls in between from one to four minutes. Some special tariffs can be provided to the customers to attract them to make long duration calls to get more business out of it.

6. Conclusions

In this paper, we proposed that autoregressive moving averages model is a good choice for trend finding and forecasting on telephone calls stationary data. Our experimental results conform to our claim as we acquired useful patterns upon forecasting data; we found that business of the Telecom Company is going well. We found that maximum number of calls ware made on Sundays due to special offers. These useful patterns can be used by higher level management to make decisions for future progress of Telecom Company and giving different kind of special tariffs. Finally, we concluded that this model can be used on other stationary time series data for acquiring useful patterns.

References

[1] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, *Proceedings of the 2nd International Symposium on Information Theory*, Academia Kiado, Budapest, 1973, pp. 267–281.

[2] R. Allard, Use of Time Series Analysis in Infectious Disease Surveillance, Bull World Health Organ, 76, 1998, pp. 327-333.

[3] O. Ait-Hellal, E. Altman, and T. Basar, *Proceedings of the 38th IEEE Conference on Decision and Control*, Phoenix, Arizona, USA, 3, December 1999, pp. 3071-3076.
[4] E. Altman and T. Bagar, Optimal Rate Control for High Speed Telecommunication Networks, *34th IEEE CDC*, New Orleans, Louisiana, December 1995.
[5] G.E.P. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control (2nd ed.)*, Holden Day, San Francisco, 1976.
[6] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel, *Time Series Analysis: Forecasting and Control, 3rd Edition*. Prentice-Hall, Englewood Cliffs, NJ, 1994.
[7] S. Basu, A. Mukherjee, and S. Klivansky, *Time Series Models for Internet Traffic, in IEEE Infocom '96*, San Fransisco, California, March 1996, pp. 611-619.
[8] S. Chaudhuri and U. Dayal, An Overview of Data Warehousing and OLAP Technology. *ACM SIGMOD Record 26 (1)*, March 1997.
[9] N. Channouf, P.L. Ecuyer, A. Ingolfsson, and A.N. Avramidis, *The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary*, Alberta, Health Care Management Science, 10, November 2006, pp. 25-45.
[10] M. Corradi, R.G. Garroppo, S. Giordano and M. Pagano, Analysis of f-ARIMA. Processes in the Modeling of Broadband Traffic, *IEEE International Conference on Communications*, 3, 2001, pp. 964-968.
[11] V.D. Giacinto, A Generalized Space-Time ARMA Model with an Application to Regional Unemployment Analysis in Italy, *International Regional Science Review*, 29(2), 2006, pp. 159-198.
[12] R. Grnenfelder, J. Cosmas, S. Manthorpe, and A. Odinma-Okafor, Measurement and ARMA Model of Video Codecs in an ATM Environment, *13th ITC*, Copenhagen, 14, June 1991, pp. 981-985.
[13] C.S. Hilas, S.K. Goudos, and J.N. Sahalos, Seasonal Decomposition and Forecasting of Telecommunication Data. A Comparative Case Study, *Technological Forecasting and Social Change***,** 73(5), June 2006, pp. 495-509.
[14] J. Han and M. Kamber, *Data Mining: Concepts and Techniques (2nd ed.)*, The Morgan Kaufmann Series in Data Management Systems, March 2006, pp. 1-36.
[15] S.J. Huang and K.R. Shih, Short-Term Load Forecasting Via ARMA Model Identification Including Non-Gaussian Process Considerations, *IEEE Transactions on Power Systems*, 18 (2), May 2003.
[16] D. Heyman, A. Tabatabai, and T.V. Lakshman, Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM networks, *IEEE Dans. On Circuits and Systems for Video Technology*, 2, 1992, pp. 49-59.
[17] W.H. Inmon, *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professional*. John Wiley & Sons, Inc. New York, 1992, pp. 127-145.
[18] L.M. Liu, S. Bhattacharyya, L.S. Sclove, R. Chen, and J.W. Lattyak, Data Mining On Time Series: An Illustration using Fast-food Restaurant Franchise Data, *Computational Statistics & Data Analysis***,** 37, October 2001, pp. 455-476.
[19] W. Li, J. Liu, J. Le, and X. Wang, The Financial Time Series Forecasting Based on Proposed ARMA-GRNN Model, *Proceedings of International Conference on Machine Learning and Cybernetics*, 4, August 2005, pp. 2005- 2009.
[20] H. Müller and J.C. Freytag, Problems, Methods, and Challenges in Comprehensive Data Cleansing, *Publication Categorizer on Data Cleaning*, 2003. http://www.dbis.informatik.huberlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164 mueller.pdf.
[21] B. Melamed, D. Raychaudhuri, B. Sengupta, and J. Zdepski, TES-Based Video Source Modeling for Performance Evaluation of Integrated Networks, *IEEE 'i9an.s. On Communications*, 42, October 1994.
[22] G. Madden and J. Tan, Forecasting Telecommunications Data with Linear Models, *Telecommunications Policy*, 31, February 2007, pp. 31-44.
[23] M. Nomura, T. Fujii, and N. Ohta, Basic Characteristics of Variable Rate Video Coding in ATM Environments, *IEEE JSAC, 7*, 1989, pp. 752-760.
[24] V. Onado, Application of ARIMA Models to Forecast the Quality of a French Cheese: the 'Comté', *Int J Biomed Comput*, 28, September 1999, pp. 249-267.
[25] T. Poddig and C. Huber, Data Mining for the Detection of Turning Points in Financial Time Series, Advances in Intelligent Data Analysis: *Proceedings of Third International Symposium, IDA-99*, Amsterdam, The Netherlands, 1642, August 1999.
[26] W. Shouyang, Y.U. Lean, and K.K. LAI, Crude Oil Price Forecasting With Tei@I Methodology**,** *Journal of Systems Science and Complexity*, 18(2), 2005, pp. 145-166.
[27] R.Y.C. Tse, An application of the ARIMA Model to Real-Estate Prices in Hong Kong, *Journal of Property Finance*, 8(2), 1997, pp. 152 – 163.
[28] T. Yang, A Time Series Data Mining Based on ARMA and Hopfield Model for Intrusion Detection, *International Conference on Neural Networks and Brain*, 2, October 2005, pp. 1045- 1049.