

Packet-Loss Prediction Model Based on Historical Symbolic Time-Series Forecasting

by

Hooman Homayounfard

A dissertation submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in the Faculty of Engineering and Information Technology



UNIVERSITY OF TECHNOLOGY, SYDNEY

October 2013

© Copyright 2013

by

Hooman Homayounfard

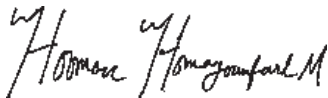
© All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

A handwritten signature in black ink, appearing to read "Homayounfar M.", written in a cursive style.

Acknowledgements

During my PhD program, I have received support and encouragement from a great number of individuals. I express my deep gratitude to my principal supervisor, Dr Paul Kennedy, for his excellent guidance in writing this thesis and the papers which preceded chapters. Dr Paul's professional advice and unyielding support led me to accomplishing this task.

I would like to thank my co-supervisors Prof John Debenham, Prof Robin Braun and Prof Simeon Simoff for advising me on the work of this thesis. I should also acknowledge Prof Barry Jay, Prof Didar Zowghi, Prof Massimo Piccardi, Prof Longbing Cao, Prof Jie Lu, Dr Ante Prodan, Dr Vinod Mirchandani, Dr Priyadarsi Nanda, Dr Masoud Talebian, Andrew Litchfield, Jose Vergara, Nima Ramzani, Debi Taylor and Vahid Behbood. These are a few of many people who had a positive impact on my achievements.

I am in debt to the whole UTS community for challenging my mind over the past years. I do thank Mrs Eilagh Rurenga for polishing chapters in my thesis. I am beyond grateful for the help from Craig Shuard and Phyllis Agius, the FEIT research officers.

Dear family and friends, thank you for your incredible support. I would have been lost without you in this process. My special regard to my wife and mother, Narges and Mahin, for their sincere love, patience and support. The thesis is dedicated to my dear Leila.

Publications

1. **Hooman Homayounfard**, Paul Kennedy, and Robbin Braun, *NARGES: Prediction Model for Informed Routing in a Communications Network*, In J. Pei et al., editor, *LNAI*, volume 7818, pages 327-338. Springer Berlin Heidelberg, 2013.
2. **Hooman Homayounfard** and Paul Kennedy, *HDAX: Historical Symbolic Modelling of Delay Time Series in a Communications Network*, In P. J. Kennedy, K. Ong, and P. Christen, editors, *AusDM09*, volume 101 of CRPIT, pages 129-138, Melbourne, Australia, 2009.

Abstract

Rapid growth of Internet users and services has prompted researchers to contemplate smart models of supporting applications with the required Quality of Service (QoS). By prioritising Internet traffic and the core network more efficiently, QoS and Traffic Engineering (TE) functions can address performance issues related to emerging Internet applications. Consequently, software agents are expected to become key tools for the development of future software in distributed telecommunication environments. A major problem with the current routing mechanisms is that they generate routing tables that do not reflect the real-time state of the network and ignore factors like local congestion.

The uncertainty in making routing decisions may be reduced by using information extracted from the knowledge base for packet transmissions. Many parameters have an impact on routing decision-making such as link transmission rate, data throughput, number of hops between two communicating peer end nodes, and time of day. There are also other certain performance parameters like delay, jitter and packet-loss, which are decision factors for online QoS traffic routing.

The work of this thesis addresses the issue of defining a Data Mining (DM) model for packet switching in the communications network. In particular, the focus is on decision-making for smart routing management, which is based on the knowledge provided by DM informed agents. The main idea behind this work and related research projects is that time-series of network performance parameters, with periodical patterns, can be

used as anomaly and failure detectors in the network. This project finds frequent patterns on delay and jitter time-series, which are useful in real-time packet-loss predictions.

The thesis proposes two models for approximation of delay and jitter time-series, and prediction of packet-loss time-series – namely the Historical Symbolic Delay Approximation Model (HDAX) and the Data Mining Model for Smart Routing in Communications Networks (NARGES). The models are evaluated using two kinds of datasets. The datasets for the experiments are generated using: (i) the Distributed Internet Traffic Generator (D-ITG) and (ii) the OPNET Modeller (OPNET) datasets.

HDAX forecasting module approximates current delay and jitter values based on the previous values and trends of the corresponding delay and jitter time-series. The prediction module, a Multilayer Perceptron (MLP), within the NARGES model uses the inputs obtained from HDAX. That is, the HDAX forecasted delay and jitter values are used by NARGES to estimate the future packet-loss value.

The contributions of this thesis are (i) a real time Data Mining (DM) model called HDAX; (ii) a hybrid DM model called NARGES; (iii) model evaluation with D-ITG datasets; and (iv) model evaluation with OPNET datasets.

In terms of the model results, NARGES and HDAX are evaluated with offline heterogeneous QoS traces. The results are compared to Autoregressive Moving Average (ARMA) model. HDAX model shows better speed and accuracy compared to ARMA and its forecasts are more correlated with target values than ARMA. NARGES demonstrates better correlation with target values than ARMA and more accuracy of the results, but it is slower than ARMA.

Contents

Acknowledgements	v
Abstract	vii
1 Introduction	1
1.1 Motivation	5
1.2 Objectives and Key Tasks	6
1.3 Research Design and Methodology	7
1.3.1 Data Network Scenarios	7
1.3.2 Proposed Data Mining Model	8
1.4 Contributions of the Thesis	10
1.5 Structure of the Thesis	11
2 A Review of QoS Time-Series Prediction Models	13
2.1 Overview	14
2.2 Routing in Communications Networks	16

2.2.1	Problems in Decision-Making for Routing	19
2.2.2	Online QoS Routing	21
2.2.3	QoS Routing Parameters	22
2.2.3.1	Packet Delay	23
2.2.3.2	IPDV	23
2.2.3.3	Packet Loss	24
2.3	Telecommunications Data Analysis	24
2.3.1	Knowledge Discovery in Telecommunications	25
2.3.2	Data Network Scenarios	29
2.4	Prediction Methods and Models for QoS Routing	30
2.4.1	Intelligent agents for real time data mining	30
2.4.2	QoS pattern analysis	33
2.4.3	Other Related Work	36
2.5	Time-Series Analysis and Forecasting	37
2.5.1	Quantitative Time-Series Analysis	39
2.5.1.1	Modeling and forecasting with ARMA	40
2.5.1.2	AR Model	40
2.5.1.3	MA Model	41
2.5.1.4	ARMA models	41
2.5.1.5	Non-Stationary Models and ARIMA	42

2.5.2	Qualitative Time-Series Analysis	42
2.5.2.1	Frequent Pattern Mining	43
2.5.2.2	Perception Based Data Mining	45
2.5.2.3	Multilayer Perceptron	47
2.6	Conclusions	47
3	Data Mining Model for Packet Loss Prediction	49
3.1	Overview	50
3.2	Preliminary Descriptions	54
3.2.1	QoS Time Series	55
3.2.2	Pattern Definition	56
3.2.3	Look-up Table	58
3.3	Formal Model Description	59
3.3.1	Forecasting Module: HDAX	60
3.3.1.1	HDAX Training	61
3.3.1.2	HDAX Simulation	62
3.3.2	Predictive Module: Multi-layer Perceptron	63
3.4	Implementation Paradigm	65
3.4.1	Forecasting Module: HDAX	65
3.4.1.1	Training Phase	67
3.4.1.2	Simulation Phase	68

3.4.2	Predictive Module: Multi-layer Perceptron	69
3.4.2.1	Network Design	69
3.4.2.2	Training Algorithm	70
3.5	Conclusions	72
4	Model Evaluation	73
4.1	Evaluation Benchmark and Measurements	74
4.1.1	ARMA Benchmark	74
4.1.2	Error Measurement	75
4.1.3	Speed Measurement	76
4.1.4	Quality Measurement	77
4.2	Datasets	78
4.2.1	D-ITG Datasets	79
4.2.1.1	D-ITG Data Characteristics	81
4.2.1.2	D-ITG Network Test-beds	82
4.2.1.3	D-ITG Software Architecture	84
4.2.1.4	Parameter Settings for D-ITG	85
4.2.2	OPNET Datasets	87
4.2.2.1	OPNET Data Characteristics	88
4.2.2.2	OPNET Network Test-bed	89
4.2.2.3	Parameter Settings for OPNET Modeller	90

4.3	Evaluation Methodology	91
4.4	Experiments and Results	92
4.4.1	Experiment 1: Approximating Delay Time-Series with HDAX	93
4.4.2	Experiment 2: Impact Analysis of End-to-End Path with Various Network Congestion Level on Model Predictions	97
4.4.2.1	Model Results with D-ITG Datasets	97
4.4.2.2	Model Comparison	98
4.4.2.3	Discussion on the Quality of Results	101
4.4.2.4	HDAX	101
4.4.2.5	NARGES	104
4.4.3	Experiment 3: Impact Analysis of Network Queuing Policies on Model Prediction	106
4.4.3.1	Model Results with OPNET Datasets	106
4.4.3.2	Model Comparison	107
4.4.3.3	Discussion on the Quality of Results	109
4.4.3.4	HDAX	109
4.4.3.5	NARGES	111
4.5	Summary of Model Performance	113
4.6	Conclusions	116
5	Conclusions and Future Work	117

5.1	Conclusions	118
5.2	Limitations	121
5.3	Future Work	121
A	List of Acronyms	125
B	ARMA Parameter Estimation	131
B.1	Preliminary Estimation	131
B.2	Maximum Likelihood Estimation	135
C	Implementation of Loss Predictor - Source Code	137
C.1	NARGES Implementation	137
C.2	HDAX Implementation	150
C.3	HDAX Functions	156
C.4	Error function	160
C.5	Running Experiments	162
C.6	ARMA Implementation	168
	Bibliography	187

List of Tables

2.1	Comparison between routing strategies	19
2.2	A qualitative comparison between WSP and SWP routing strategies (adapted from Marzo et al. (2003))	22
3.1	Deterministic Mapping Function (DMF), the scale of time-series trends used for mapping numerical traces to the categorical (linguistic) terms.	61
3.2	Description of the fields used for the pattern lookup-table implementation.	67
4.1	Characteristics of the end-to-end paths for the data obtained from Distributed Internet Traffic Generator (D-ITG).	82
4.2	Comparison between OPNET, OMNET and NS2	87
4.3	Types of Queueing Policies for the data obtained from OPNET.	88
4.4	Accuracy of Historical Symbolic Delay Approximation Model (HDAX) and Autoregressive Moving Average (ARMA) (benchmark) on first phase of simulation runs together with speed of algorithm.	96
4.5	Accuracy of HDAX and ARMA (benchmark) in the phase two of simulation runs together with speed of algorithm.	96

- 4.6 Normalised root mean square error (NRMSE) together with algorithms speed and cross-correlation coefficients of HDAX and ARMA forecasts for D-ITG delay and jitter time-series. 98
- 4.7 Normalised root mean square error together with speed of calculation and cross-correlation coefficients of NARGES and ARMA predictions for D-ITG packet-loss time-series. 99
- 4.8 Average rankings as calculated using Friedman test for the results of the algorithms for accuracy, speed and cross-correlation (Cross-Correlation Function (CCF)) over delay, jitter and packet-Loss time-series. The algorithms with **bold** rank number have better ranking in each row. 100
- 4.9 Holm / Hochberg Table for $\alpha = 0.05$ (**bold** *algorithm* names). 100
- 4.10 Normalised root mean square error (NRMSE) together with algorithms speed and cross-correlation coefficients of HDAX and ARMA forecasts for OPNET Modeller (OPNET) delay and jitter time-series. 107
- 4.11 Normalised root mean square error (NRMSE) together with algorithms speed and cross-correlation coefficients of Data Mining Model for Smart Routing in Communications Networks (NARGES) and ARMA forecasts for OPNET packet-loss time-series. 107
- 4.12 Average Rankings of the algorithms; Note that in testing the algorithms for accuracy, speed and cross-correlation (CCF) over Delay, Jitter and Packet-Loss 108
- 4.13 Holm / Hochberg Table for $\alpha = 0.05$. Note that in testing the algorithms for accuracy, speed and cross-correlation (CCF) over Delay, Jitter and Packet-Loss models printed in **bold** are statistically significantly better. . 108

List of Figures

2.1	Stages of the theoretical framework of the Knowledge Discovery in Databases (KDD) process in a communications network (adapted from Rocha-Mier et al. (2007))	27
2.2	Process model of a communications network TSDM (adapted from Rocha-Mier et al. (2007))	31
2.3	Initial and pattern time-series for a network variable a) target time-series, b) pattern time-series of slope values (adapted from Rocha-Mier et al. (2007))	32
2.4	Outliers (discords) are particularly attractive as anomaly detectors (adapted from Keogh et al. (2005))	33
2.5	PDL pattern structures from active QoS measurement (adapted from Miloucheva et al. (2003))	35
2.6	Using Spatio-analyser for automation of the DM tasks (adapted from Miloucheva et al. (2003))	36
3.1	Conceptual Framework for NARGES Data Mining Model	51
3.2	A schema of NARGES data mining model	54
3.3	D-ITG Graphical User Interface (GUI)	56

3.4	Basic Patterns	57
3.5	Basic patterns assigned to triplet trends.	58
3.6	The training phase uses a time-series dataset values to recognise $i - j - k$ patterns and train the look-up table that maps each of these patterns to a respective frequency. The table is then used for forecasting the k trend at time $t + 1$ in the simulation phase.	62
3.7	Multi-layer Perceptron	64
3.8	A sample string of symbolic values trend time-series $\{P, SI, P, I, SD, P\}$	66
3.9	Impact of the number of Hidden Layer's Neuron on the Multilayer Perceptrons (MLP) Accuracy and Performance	71
4.1	Target dataset (target delay, jitter and packet-loss time-series) together with the HDAX forecasted data (forecasted delay and jitter time-series) are used in NARGES model to predict future packet-loss.	80
4.2	D-ITG framework	82
4.3	The University of Naples Federico II (UNINA) experimental test-bed used to generate D-ITG traces. (adapted from Botta et al. (2008))	83
4.4	Trace Route between the two nodes used for D-ITG QoS data generation .	84
4.5	D-ITG GUI setup	85
4.6	OPNET network design used for QoS data generation	89
4.7	Applications profile setting on OPNET Modeller	90

4.8	Target (solid line), HDAX predicted (star-dashed line) and ARMA predicted (dot-dashed line) delay values for simulation run 1.	94
4.9	Target (solid line), HDAX predicted (star-dashed line) and ARMA predicted (dot-dashed line) delay values for simulation run 2.	95
4.10	Target (solid line), HDAX predicted (star-dashed line) and ARMA predicted (dot-dashed line) delay values for simulation run 3.	95
4.11	Target (solid line), HDAX predicted (star-dashed line) and ARMA predicted (dot-dashed line) delay values for simulation run 3.	96
4.12	Boxplots of distributions of target delay time-series for dataset 13 together with those for outputs of HDAX and ARMA	102
4.13	Stemplots of cross-correlation of HDAX forecasts and target delay time-series for dataset 13 together with those of ARMA	103
4.14	Boxplots of distribution for target jitter time-series within dataset 13 together with those for outputs of HDAX and ARMA	103
4.15	Stemplots of cross-correlation of HDAX forecasts and target jitter time-series for dataset 13 together with those of ARMA	104
4.16	Boxplots of distributions for target packet-loss time-series within dataset 13 together with those for outputs of NARGES and ARMA	105
4.17	Stemplots of cross-correlation of NARGES predictions and target packet-loss time-series for dataset 13 together with those of ARMA	105
4.18	Boxplots of distributions of target delay time-series for dataset 41 together with those for outputs of HDAX and ARMA	110

4.19 Boxplots of distributions of target jitter time-series for dataset 41 together with those for outputs of HDAX and ARMA 110

4.20 Stemplots of cross-correlation of HDAX forecasts and target delay time-series for dataset 41 together with those of ARMA 111

4.21 Stemplots of cross-correlation of HDAX forecasts and target jitter time-series for dataset 41 together with those of ARMA 112

4.22 Boxplots of distributions of target packet-loss time-series for dataset 41 together with those for outputs of NARGES and ARMA 112

4.23 Stemplots of cross-correlation of NARGES predictions and target packet-loss time-series for dataset 41 together with those of ARMA 113

4.24 Error (NRMSE) of HDAX and NARGES vs ARMA together with speed of algorithm and cross-correlation coefficients are shown in the column (a) to (c), respectively. The first and second rows are the HDAX results and the last row shows the whole model (NARGES) results. In the twin bar charts, the left gray bars shows HDAX (in the first two rows) and NARGES (in the last row) while the right bar filled with wide downward diagonal pattern denotes ARMA outcomes. 115