Features Extraction of Growth Trend in Social Websites Using Non-linear Genetic Programming

Umer Khayam, Durre Nayab, Gul Muhammad Khan, and S. Ali Mahmud

Center for Intelligent Systems and Network Research, UET Peshawar
{umerkhayam,nayaab_khan,gk502,sahibzada.mahmud}@nwfpuet.edu.pk

Abstract. Nonlinear Cartesian Genetic Programming is explored for extraction of features in the growth curve of social web portals and establishment of a prediction model. Daily hit rates of web portals provide the measure of the growth and social establishment behavior over time. Non-linear Cartesian Genetic Programming approach also termed as CGPANN has unique ability of dealing with the nonlinear data as it provides the flexibility in feature selection, network architecture, topology and other necessary parameters selection to establish the desired prediction model. A number of socially established web portals are used to evaluate the performance of the model over a span of two years. Efficient performance is shown by the system keeping the fact in consideration that only single independent web portal data is used for training the network and the same network was used for the other web portals for their performance evaluation. The system performance is significantly good as the system selects only the desired features from the features presented as input and achieves an optimal network and topology that produce the best possible results.

Keywords: Neuro Evolution, Artificial Neural Network, Cartesian Genetic Programming, Web Traffic Prediction, Web Portals, Future Demand.

1 Introduction

For the last few years a large number of web portals have been developed which intend to provide various kinds of valuable services to the internet users. Web portals such as social networks, online forums, job portals, blogging platforms, and news websites aim to provide uninterrupted free as well as paid services to their users. These web portals can also be used for business purposes or an advertising platform when its number of users increases and hit rate becomes high.

The forecast of internet traffic is an important issue [1] and accurate forecasting of web portals helps in determining the current growth and future prospect of the portals. It also provides the business department an idea to identify the features and patterns [3] that can lead to a higher hit rate in future, also an insight about the future demand and anomalies [4] [24] in the network can be detected. Software developers can get an idea about the types of browsers that are used to get access to the company website with the aid of forecasting.

The business departments, developers and the network administrators do the task of forecasting internet traffic intuitively with the help of market information about the

L. Iliadis et al. (Eds.): AIAI 2014, IFIP AICT 436, pp. 414–423, 2014.

usual behavior and future number of visitors [6]. This produces only a rough idea of what the future traffic will look like with a little day to day network administration. On the other hand, contributions from the areas of computational intelligence and artificial neural networks have replaced the intuition based forecasting methods [20] [21]. The forecasting results obtained by these neural networks are accurate, reliable and less time consuming compared to the former methods.

Different methods have been used to forecast the internet traffic and several experiments have been performed on real-world datasets using these methods in order to determine their accuracy [1] [2]. In addition, different time scales have been used for forecasting purposes [2] and some methods perform well in short term forecasting, while others perform better in large term forecasts. By accurate prediction and forecasting, the services of web portals can be extended, diminished or altered in order to maintain a high number of visitors.

The main intent of this work is to present the application of artificial neural networks to web traffic forecasting for web portals.

2 Literature Review

2.1 Traffic Forecasting of Web Portals

Forecasting of web traffic is one of a critical task [6] for researchers these days. There are several applications of internet traffic forecasting that aims for efficient traffic engineering, anomaly detections and business tools development [1]. Various techniques have been used to predict the internet traffic accurately. Paulo Cortez used a neural network ensemble approach and two important adapted time series methods such as ARIMA and Holt-Winters [2] in order to determine their accuracy for predicting traffic in TCP/IP based networks. G. Rutka inspected the performance of traffic models for forecasting the future traffic variations as precisely as possible using multilayer perceptron and radial basis function networks based on measured traffic history [7]. Machine learning methods have been used to analyze web traffic using decision trees [8]. A multi scale decomposition approach was used for real time traffic prediction that outperforms traffic prediction using neural network approach [9] and gives comparatively better results.

Many internet service providers use multi-protocol label switching to establish fill mesh of MPLS between pairs of the routers in a network in order to optimize the bandwidth resources in the network [10]. Even if MPLS is not used, with the knowledge of future demand of traffic matrix, the traditional allocation of the routing protocol weights can be done more efficiently [1]. The forecasted web traffic cab be used for anomaly detection [11] [12] in the network by comparing the actual traffic to the forecasted traffic in the network.

The forecasted web traffic can be used by the business departments [1] of the company to get an idea about the current popularity and future demand of their web services and applications. The statistical data about the visitors on a particular web portal contain patterns that are viable in determining the future demand. By analyzing statistical data, the factors effecting hit rate of the web portals can be identified, an insight about the future hit rate, types of visitors, and survival of the web portals can be predicted. Several factors have been identified in past that effect the number of visitors

on these portals [1]. These factors include accessibility, browsers compatibility, trust level, flexibility, productive and organized content, and the targeted age, group and gender etc [2]. The business department and the developers of the company can take decisions based on the forecasted results in order to get increased traffic in the future.

Cartesian genetic programming and neural network models are used for forecasting purposes. Today the numbers of structures of the artificial neural networks and training algorithms that are applied in their learning process have evolved and these models are also used to forecast future network traffic. Similarly these models can be used to forecast the traffic patterns of web portals.

2.2 Cartesian Genetic Programming (CGP)

The idea of Cartesian Genetic Programming (CGP) was proposed in 1999 by Julian F Miller [22], that is a genetic programming approach with a two dimensional graphical arrangement of its functional units i.e. nodes. These programs are represented as directed acyclic structures operating in a feed forward fashion. CGP has two formats of network representation i.e. phenotype and genotype. The physical format of array of interconnected nodes represents the phenotypic structural space and its genotype can be represented as an array of integers i.e. genes. The nodes are the functional units that exhibit a node function, inputs, and outputs [23]. CGP provide a general platform for evolving the hybrid structure of any number of networks in any order. It provides a complete Cartesian architecture for their interconnectivity patterns producing less hybrid structures from a pole of networks. CGP is explored in a range of applications producing interesting results [23].

3 Cartesian Genetic Programming Evolved Artificial Neural Network

The strategy employed for the evolution of an ANN plays a vital role in ANNs and hence is a major concern for the researchers these days. The evolutionary strategy used for the evolution of ANN model proposed in this research is CGP. CGPANN is signal processing system that is based on some of the known organizing principles of the human brain [15] [16] [17]. CGP evolve the ANN with its unique architecture that makes it computationally cost effective and efficient. They are computational models that are capable of machine learning and pattern recognition. These systems are represented by a number of independent, simple processors called neurons which are interconnected with each other by weighted connections [18]. Figure 1 shows a typical CGPANN genotype and phenotype with inputs (I_0, I_1, I_2, ... I_9), outputs (O_0, O_1, O_2, ... O_9) , active nodes (0, 1, 2, 4, 5, 7, 8 & 9), inactive nodes (3 & 6) and weights (w_0, w_1, w_2……., $w_{17)}$. The arity (number of inputs per node) of the network is 2 and the number of inputs to the system is 10.

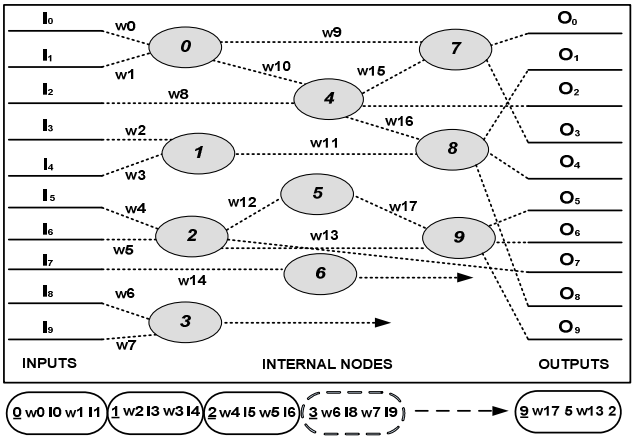

Fig. 1. A typical cgpann phenotype and genotype

The string of numbers underneath the nodal figure is the genotype arranged in boxes represent individual node in terms of its function, inputs and the weights associated with the inputs respectively. The box having dotted lines represents the inactive nodes.

The generic expressions for the CGPANN model can be given in the following equations. Equation (1) shows the system outputs (y_i) as the summation of system inputs (x_i) when the weights and node functions are not associated with them. Equation (2) and (3) shows the network outputs (y_i) with the weights (wi) and functions (f^i) are associated with the system inputs.

$$\boldsymbol{y_i = \sum_{i=1}^{N} x_i} \tag{1}$$

$$\boldsymbol{y_i = \sum_{i=1}^{N} x_i w_i} \tag{2}$$

$$\boldsymbol{y_i = f^i(y_i) = f^i(\sum_{i=1}^{N} x_i w_i)} \tag{3}$$

Equation (4) and (5) shows the range (i) of the inputs, functions and weights respectively. As can be seen from Eq (4) the range (i) belongs to natural numbers (N) where N begins from 1 to NT. The weights belong to real numbers and are chosen in the range [-1, 1].

$$\{\boldsymbol{i | i \epsilon N 1 \leq i \leq N_T}\} \tag{4}$$

$$\boldsymbol{w_i = \{w_i | w_i \epsilon R \ -1 \leq w_i \leq 1\}} \tag{5}$$

Equation (6) represent the single generation obtained from the network in terms of inputs (I), system outputs (y_i's) and ultimate output (O_p).

$$\boldsymbol{G_k = \{I, y_i, y_{i-1}, \ldots \ldots, y_1, O_p\}} \tag{6}$$

Equation (7) shows the ultimate system output in terms of individual system outputs. Equation (8) shows the individual system output in terms of nodal outputs (y_i) and system inputs (I's) with weights (w_i's) associated with them.

$$\boldsymbol{O_p = \frac{1}{n} \sum_{i=1}^{N} Oi} \tag{8}$$

$$\boldsymbol{Oi = (f(\sum(y_j w_j + y_{j-1} w_{j-1} + \ \ldots \ldots + y_1 w_1 + I w_k))} \tag{9}$$

3.1 Mutation in CGPANN

The optimal network of the CGPANN is achieved during the process of evolution. During evolution mutation in the genotype takes place and optimal network is achieved by the selection and promotion of the fittest genotype. Figure 2 (a, b, c) show the mutation process of the CGPANN network during evolution process. Figure 2a shows the original phenotype and genotype of the network and Figure 2b and 2c demonstrate the process of mutation in the function (***f2***) and input (***I_3***) genes respectively.

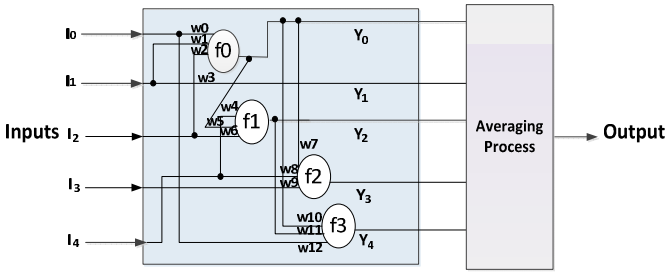

Fig. 2a. Original cgpann phenotype

Genotype1= ***f0*** $I_0 w_0 I_1 w_1 I_2 w_2$, ***f1*** $f_0 w_7 I_4 w_4 I_2 w_6$, ***f2*** $f_0 w_7 I_4 w_8 I_3 w_9$, ***f3*** $f_0 w_{10} f_1 w_{11} I_0 w_{12}$

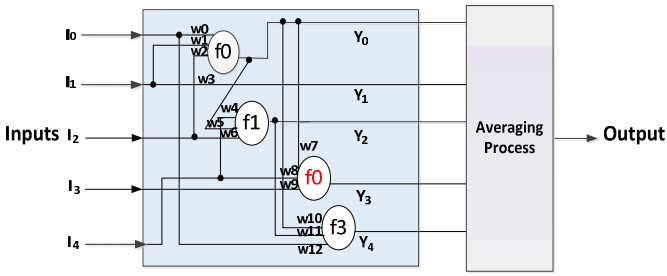

Fig. 2b. Mutation in the function of the network

Genotype2= ***f0*** $I_0 w_0 I_1 w_1 I_2 w_2$, ***f1*** $f_0 w_7 I_4 w_4 I_2 w6$, ***f0*** $f_0 w_7 I_4 w_8 I_3 w_9$, ***f3*** $f_0 w_{10} f_1 w_{11} I_0 w_{12}$

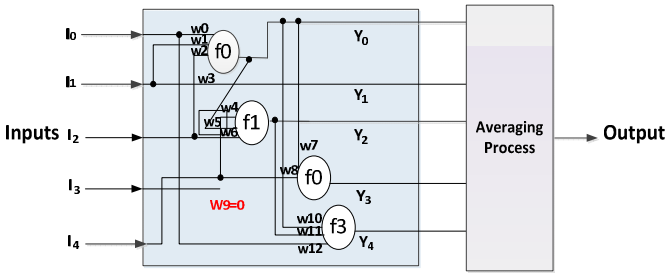

Fig. 2c. Mutation in the input to the node

Genotype3= ***f0*** $I_0 w_0 I_1 w_1 I_2 w2$, ***f1*** $f_0 w_7 I_4 w_4 I_2 w6$, ***f0*** $f_0 w_7 I_3 w9$, ***f3*** $f_0 w_{10} f_1 w_{11} I_0 w_{12}$

3.2 The Forecaster Model

The learning process of our forecaster model consists of adaptive modifying of the connection weights to improve the overall functionality of the neural network as the parallel signal processing system. The system learns the patterns and selects the optimal features in the historical data to predict the future values. The performance of the forecasting model is measured by Mean Absolute Percentage Error (MAPE) as it is a common metric in forecasting applications [6].

The forecaster model consists of single row and multiple columns of neurons trained by altering the values of the connections between the neurons. A nonlinear CGP is used for translating the network as it handles the nonlinear data such as that of web portals efficiently. The statistical data of six web portals for a period of one year is fed to the forecasting model and model is trained. The network is capable of feature selection i-e selecting optimal (not all) number of inputs and nodes for evolving the final network. The neural network model is trained on a one year hourly spaced datasets of Linkedin.com. The trained model is then used to forecast the internet traffic of web portals other web portals. The mean absolute percentage error is calculated between the forecasted values and the actual values.

4 Experimental Setup

A collection of time ordered dataset is taken for six web portals including Linkedin.com, Time.com, Tumblr.com, Answers.com, Hubpages.com and Collegehumor.com for a period of two years starting from October 2011 to October 2013. The data is obtained and verified by Quantcast.com that is a web portal used for monitoring the visitor hit rate on the portals available on the internet. The obtained data was the average daily hit rates of the six specified web portals for a period of two years.

During the training phase daily global hit rates of LinkedIn.com is used for training the system. Initially a random population of ten networks is produced for five independent seeds. The system parameters are defined and initialized. The number of offspring per generation is 9 under the $1+\lambda$ evolutionary strategy where λ is set to be 9. The initially generated genotype is mutated to produce nine offspring. The best genotype among these genotypes is selected and mutated again. The process is repeated unless the best network is achieved. The mutation rate (μr) is set to be 10%. The number of system inputs is set to be ten and that is the daily hit rates data and the arity of the system i.e. number of nodal inputs is set to be 5. Log sigmoid is taken as the activation function. These system parameters are chosen based on previous performance of CGPANN [5, 14] and evolutionary performance. The network is trained for variable number of inputs and outputs. The network takes seven days and fourteen days instances of the daily dataset as inputs to the network and forecasts the hit rate of the web portal for the next single day, seven days and fourteen days as outputs of the network.

Performance of the system is evaluated with MAPE (Mean Absolute Percentage Error) that is calculated by comparing the actual value of the hit rate with that estimated by the model.

The mathematical expression for MAPE is given below:

$$\text{MAPE} = \frac{1}{\text{N}} \sum \text{i} = (1 \text{ to N}) \left(\frac{|\text{L}_{\text{Fi}} - \text{L}_{\text{Ai}}|}{\text{L}_{\text{Ai}}} \right) \times 100$$

Where LAi is the actual value, LFi is the estimated value and N is the number of days.

5 Results and Analysis

During the training phase the system is been trained for one million generations and the best trained network is then tested for its performance. The system is trained on one web portal data i.e. LinkedIn.com and is tested on a different dataset of five web portals. The data with known hit rate values are estimated during the testing phase and compared with the actual values to evaluate the performance of the model.

The CGPANN forecaster model proposed in this work has been trained using daily averaged web traffic data of LinkedIn.com for one year, starting from October 2011 to October 2012. The results for the training session of each network are given in the following Table I. The output of the system is evaluated from the average of five independent evolutionary runs for each network with mentioned combination of inputs and outputs. The fittest network is achieved for the network with 7 inputs and 1 output during the training phase with the MAPE value of **1.1149.**

Table 1. Training results of CGPANN for web portal forecasting model

Nodes	**7 in 1 out**	**7 In 7 out**	**14 In 7 out**
50	0.11675	0.13302	0.132575616
100	0.11686	0.13457	0.132009816
150	**0.11493**	0.13291	0.13219517
200	0.11581	0.13402	0.132067988
250	0.11624	0.13377	0.133859813
300	0.11557	0.13290	0.132211834
350	0.11666	0.13318	0.132714417
400	0.11619	0.13349	0.133092481
450	0.11925	0.13347	0.13273484
500	0.11659	0.13286	0.13262616

For evaluation and testing, a new dataset is fed to the network for its validation. The new dataset is the statistical data of the visitor hit rate of Time.com, Tumblr.com, Answers.com, Hubpages.com and Collegehumor.com for a period of two years from ranging from October 2011 to October 2013. The testing phase results for the hit rates forecasting model are tabulated in Table II, III and IV for the mentioned combination of inputs and outputs. The proposed forecaster model takes seven days hit rates as input to the system and predicts the eighth day hit rate in the first case. In the second

case it takes seven daily hit rates as input and predicts seven future daily hit rates as output of the system. In all the combinations of the inputs and outputs the forecaster model showed higher accuracy hence the model is proficient both in terms of single instant prediction and multiple instants predictions. The best MAPE values achieved for each web portal are highlighted in each table respectively.

The best MAPE value achieved by the model is 9.053% given in Table 2 where as other works done the web portal forecasting has achieved accuracies in the ranges such as 12-23% [1], 13–22% [2], 72.74 and 12.04 % [19].

Table 2. Testing results for one day hit rate as input with seven day's data as output

No. Of Nodes	Times	Tumblr	Answers	Hubpages	Collegehumor
50	0.1463	**0.09053**	0.14202	0.09647	0.1553
100	0.1433	0.09261	0.14445	0.09671	0.1541
150	0.1549	0.09426	0.14266	0.10058	0.1692
200	0.1537	0.09373	0.14592	0.10016	0.1652
250	0.1603	0.09414	0.14548	0.10192	0.1730
300	0.1430	0.09072	0.14254	0.09622	0.1507
350	0.1469	0.09361	0.14625	0.09828	0.1565
400	0.1545	0.09301	0.14100	0.09983	0.1677
450	0.1686	0.10202	0.15097	0.10875	0.1873
500	0.1424	0.09478	0.14661	0.09819	0.1542

Table 3. Testing results for seven days hit rates as input with seven days data as output

No. Of Nodes	Time	Tumblr	Answers	Hub pages	College humor
50	0.1442	0.1159	0.19272	0.10967	0.15613
100	0.1426	0.1152	0.19391	0.10884	0.15269
150	0.1270	0.1141	0.18955	0.10495	0.13731
200	0.1496	0.1191	0.19283	0.11337	0.16625
250	0.1571	0.1201	0.19474	0.11554	0.17485
300	0.1447	0.1165	0.19327	0.11035	0.15741
350	0.1530	0.1186	0.19458	0.11311	0.16570
400	0.1351	0.1151	0.19247	0.10721	0.14410
450	0.1112	0.1141	0.19209	**0.10121**	0.11393
500	0.1121	0.1134	0.19011	0.10173	0.11769

Table 4. Testing results for 14 days hit rates as input with 7 days data as output

No. Of Nodes	Time	Tumblr	Answers	Hub pages	College humor
50	0.12795	0.11923	0.19835	0.10806	0.14111
100	0.13847	0.11873	0.19897	0.10981	0.15333
150	0.15067	0.12162	0.20324	0.11404	0.16946
200	0.13021	0.11689	0.19627	0.10698	0.14178
250	0.09451	0.11605	0.19751	0.09905	**0.09592**
300	0.15046	0.12001	0.20089	0.11343	0.16898
350	0.14978	0.12036	0.20373	0.11343	0.16571
400	0.14785	0.12161	0.20692	0.11215	0.16292
450	0.15392	0.12095	0.20030	0.11530	0.17346
500	0.14222	0.11855	0.19732	0.11093	0.15814

6 Conclusion and Future Work

We have explored nonlinear CGP for the implementation of forecasting model for global web portals growth analysis and extracting the prominent features. A number of socially established web portals are analyzed for their growth process. The performance of the system revealed that the system is robust that learns the trends and extracts the optimal features responsible for the growth rate of these portals. The network is capable of feature selection i-e selecting optimal (not all) number of inputs and nodes for the evolution of the final network. The proposed system has the ability to obtain an optimal set of features, number of nodes and connections paradigm and morphology for the best possible prediction model for the task at hand. Further work can explore the social behavior on individual portals including: probing the posts and updates, advertisement on web portals for improving the news feeds, analyzing the intent and emotions in user updates, company survival capabilities analysis and business success or failure of start-up firm analysis.

References

1. Cortez, P., et al.: Internet traffic forecasting using neural networks. In: International Joint Conference on Neural Networks, IJCNN 2006. IEEE (2006)
2. Cortez, P., et al.: Multi Scale Internet traffic forecasting using neural networks and time series methods. Expert Systems 29(2), 143–155 (2012)
3. Fausett, L.: Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Pearson Education India (2006)
4. Krishnamurthy, B., et al.: Sketch-based change detection: methods, evaluation, and applications. In: Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement, ACM (2003)

5. Nayab, D., Muhammad Khan, G., Mahmud, S.A.: Prediction of Foreign Currency Exchange Rates Using CGPANN. In: Iliadis, L., Papadopoulos, H., Jayne, C. (eds.) EANN 2013, Part I. CCIS, vol. 383, pp. 91–101. Springer, Heidelberg (2013)
6. Papagiannaki, K., et al.: Long-term forecasting of Internet backbone traffic. IEEE Transactions on Neural Networks 16(5), 1110–1124 (2005)
7. Rutka, G.: Neural network models for Internet traffic prediction. In: Proceedings of Electronics and Electrical Engineering, Lithuania, vol. 4(68), pp. 55–58 (2006)
8. Piramuthu, S.: On learning to predict web traffic. Decision Support Systems 35(2), 213–229 (2003)
9. Mao, G.: Real-time network traffic prediction based on a multiscale decomposition. In: Lorenz, P., Dini, P. (eds.) ICN 2005. LNCS, vol. 3420, pp. 492–499. Springer, Heidelberg (2005)
10. Davie, B.S., Rekhter, Y.: MPLS: technology and applications, vol. 1. Morgan Kaufmann Publishers, San Diego (2000)
11. Krishnamurthy, B., et al.: Sketch-based change detection: methods, evaluation, and applications. In: Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement, ACM (2003)
12. Jiang, J., Papavassiliou, S.: Detecting network attacks in the internet via statistical network traffic normality prediction. Journal of Network and Systems Management 12(1), 51–72 (2004)
13. Khan, M.M., Khan, G.M., Miller, J.F.: Evolution of Optimal ANNs for Non-Linear Control Problems using Cartesian Genetic Programming. In: International Conference on Artificial Intelligence, ICAI, pp. 339–346 (2010)
14. Pacelli, V., Bavelacqua, V., Azzollini, M.: An Artificial Neural Network Model to Forecast Exchange Rates. Journal of Intelligent Learning Systems and Applications, JILSA 3(2A), 57–69 (2011)
15. Fausett, L.: Fundamentals of neural networks. Prentice Hall (1994)
16. Osowski, S.: Neural networks in algorithmic approach. WNT, Warsaw (1996) (in Polish)
17. Zurada, J.M.: Introduction to artificial neural systems. West Publishing Company (1992)
18. Ding, X., Canu, S., Denoeux, T.: Neural network based models for forecasting. In: Proceedings of Applied Decision Technologies Conf. (ADT 1995), Uxbridge, UK, pp. 243–252 (1995)
19. Gluszek, A., Kekez, M., Rudzinski, F.: Web traffic prediction with artificial neural networks. Wilga-DL Tentative. International Society for Optics and Photonics (2005)
20. Makridakis, S., Weelwright, S., Hyndman, R.: Forecasting: Methods and Applications. John Wiley & Sons, New York (1998)
21. Hanke, J., Reitsch, A.: Business Forecasting. Allyn and Bancon Publishing. Allyn and Bancon Publishing, Massachussetts (1989)
22. Rothermich, A.J., Miller, J.F.: Studying the emergence of multicellularity with cartesian genetic programming in artificial life. In: Proceedings of the 2002 UK Workshop on Computational Intelligence, pp. 397–403. GECCO (2002)
23. Miller, J.F.: "Cartesian Genetic Programming," Genetic Programming. Natural Computing Series. Springer, Heidelberg (2011)
24. Jiang, J., Papavassiliou, S.: Detecting network attacks in the internet via statistical network traffic normality prediction. Journal of Network and Systems Management 12(1), 51–72 (2004)