# NETWORK PROBLEMS DETECTION AND CLASSIFICATION BY ANALYZING SYSLOG DATA

**FIDAA A. M. JARGHON**



**Supervisors:**
**Dr. Mohammad Shamrie Sainin**
**Assoc. Prof. Hatim Tahir**

**2016**

**Network Problems Detection and Classification by Analyzing Syslog data**



**By**

**Fidaa A. M. Jarghon**

**Supervisors:**
**Dr. Mohammad Shamrie Sainin**
**Assoc. Prof. Hatim Tahir**

## ABSTRAK

Rangkaian penyelesaian masalah adalah satu proses penting yang mempunyai bidang penyelidikan yang luas. Langkah pertama dalam prosedur penyelesaian masalah adalah mengumpul maklumat untuk mengenal pasti permasalahan. Mesej syslog yang dihantar oleh hampir semua peranti rangkaian mengandungi sejumlah besar data yang berkaitan dengan masalah rangkaian. Banyak kajian yang dijalankan sebelum ini didapati telah menggunakan menganalisis data syslog yang boleh membimbing untuk masalah rangkaian dan sebab-sebabnya. Mengesan masalah rangkaian akan menjadi lebih efektif jika masalah yang hendak dikesan telah dikelaskan dari segi lapisan rangkaian. Pengelasan data syslog perlu mengenal pasti mesej syslog yang menghuraikan masalah rangkaian untuk setiap lapisan, dan mengambil kira format yang berbeza dari pelbagai syslog untuk peranti vendor. Kajian ini menyediakan kaedah untuk mengelaskan mesej syslog yang menunjukkan masalah rangkaian dari segi lapisan rangkaian. Alat pengenalpastian data kaedah digunakan untuk pengelasan mesej syslog manakala penerangan bahagian atas mesej syslog telah digunakan untuk proses pengelasan. Apabila mesej syslog berkaitan telah dikenal pasti; ciri kemudiannya dipilih untuk melatih penjodoh bilangan. Enam algoritma pengelasan telah dipelajari iaitu LibSVM, SMO, KNN, Naive Bayes, J48, dan Random Forest. Satu set data sebenar yang diperoleh daripada peranti rangkaian Universiti Utara Malaysia (UUM) digunakan untuk peringkat ramalan. Keputusan merumuskan bahawa SVM menunjukkan prestasi terbaik semasa peringkat latihan dan ramalan. Kajian ini menyumbang pada bidang penyelesaian masalah rangkaian, dan pengelasan.



Keywords data teks: Pengelasan, SVM, Pengesanan Kerosakan..

## ABSTRACT

Network troubleshooting is an important process which has a wide research field. The first step in troubleshooting procedures is to collect information in order to diagnose the problems. Syslog messages which are sent by almost all network devices contain a massive amount of data related to the network problems. It is found that in many studies conducted previously, analyzing syslog data which can be a guideline for network problems and their causes was used. Detecting network problems could be more efficient if the detected problems have been classified in terms of network layers. Classifying syslog data needs to identify the syslog messages that describe the network problems for each layer, taking into account the different formats of various syslog for vendors' devices. This study provides a method to classify syslog messages that indicates the network problem in terms of network layers. The method used data mining tool to classify the syslog messages while the description part of the syslog message was used for classification process. Related syslog messages were identified; features were then selected to train the classifiers. Six classification algorithms were learned; LibSVM, SMO, KNN, Naïve Bayes, J48, and Random Forest. A real data set which was obtained from the Universiti Utara Malaysia's (UUM) network devices is used for the prediction stage. Results indicate that SVM shows the best performance during the training and prediction stages. This study contributes to the field of network troubleshooting, and the field of text data classification.



**Keywords:** Classification, SVM, Fault Detection

## TABLE OF CONTENTS

**Title** **Page**

## LIST OF TABLES

Table Page

No table of figures entries found.

## LIST OF FIGURES

CHAPTER One

# INTRODUCTION

Most institutions and organizations, regardless of the business types, rely on networks to manage their business. Any failure or error which occurs in the network will negatively affect their achievements, productivity and services. Therefore, it is necessary to diagnose and detect the reasons behind network failures and problems in order to fix them and reduce similar occurrences in the future. Network troubleshooting , which begins by diagnosing the problems, is a complex process. The first step is to collect information [1].

Collecting information includes answering this question, "what are the potential errors that can lead to network problems and failures?" Kyas [2], has identified five categories of errors: operator error, mass storage problems, computer hardware problems, software problems, and network problems. Hudyma & Fels [3], added two new categories: failure due to denial of service attacks (Worms, Viruses, Trojan Horses and Malicious software), and failure due to disasters such as fire, flood, earthquakes, outages and the like. Network problems, which include hardware and software problems that are directly related to the network architecture [2],account for more than one-third of information technology (IT) failures.



Network architecture is organized as a series of layers or levels [4], Open Systems Interconnection (OSI) model, and Transmission Control Protocol/Internet Protocol (TCP/IP) model, which separates network functionality into modular layers that provide "a common language for network engineers and is usually used in troubleshooting networks."

Troubleshooting could be an efficient process if it relies on a systematic approach which minimizes confusion and shortened troubleshooting time. It is carried out using the Layered Model [5] as problems are normally described in terms of a specific model layer [1]. Network errors could be distributed into the network layers depending on OSI model or TCP\IP model (physical layer, data-link layer, network layer, transport layer, application layer). And knowing the layer that has problems in

The contents of the thesis is for internal user only

# REFERENCES


[1] J. D. Sloan, *"network management and troubleshooting" in Network Troubleshooting Tools*, 1st ed. USA: O'Reilly, 2001.

[2] O. Kyas, *Network Troubleshooting*. California: Agilent Technologies, 2001.

[3] R. Hudyma and D. I. Fels, "Causes of Failure in IT Telecomunications Networks," in *Proceedings of SCI*, 2004, pp. 35–38.

[4] P. C. Gupta, *Data Communications And Computer Networks*, Eastern ec. New Delhi: Prentice hall of india private limited, 2006.

[5] B. Vachon and R. Graziani, *Accessing the WAN CCNA Exploration Companion Guide*, 1st ed. USA: Cisco Press, 2008.

[6] A. Deveriya, *Network Administrators Survival Guide.*, 1st ed. USA: Cisco Press, 2005.

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, and R. Dobbs, *Big data: The next frontier for innovation, competition, and productivity*, 1st ed. McKinsey Global Institute, 2011.

[8] Q. He and B. Veldkamp, *Classifying unstructured textual data using the Product Score Model: an alternative text mining algorithm*, 1st ed. Enschede, Netherlands: RCEC, Cito/University of Twente, 2012.

[9] T. Qiu, Z. Ge, D. Pei, J. Wang, and J. Xu, "What happened in my network: mining network events from router syslogs," in *Proceedings of the 10th ACM*, 2010, pp. 472–484.

[10] M. Roy, "Empirical Study of Different Classifiers for Sentiment Analysis," in *Data Mining and Knowledge Engineering 6.4*, 2014, pp. 160–164.

[11] T. Kimura, K. Takeshita, T. Toyono, M. Yokota, K. Nishimatsu, and T. Mori, "Network failure detection and diagnosis by analyzing Syslog and SNS data: Applying big data analysis to network operations," *NTT Tech. Rev.*, vol. 11, no. 11, 2013.

[12] K. Fukuda, "On the use of weighted syslog time series for anomaly detection," in *12th IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2011, pp. 393–398.

[13] A. Fox, D. Patterson, and M. I. Jordan, "Invited Applications Paper Detecting Large-Scale System Problems by Mining Console Logs," in *27th International Confer- ence on Machine Learning*, 2010.

[14] S. Hekmat, *Communication networks*. Línea] http://www. pragsoft. com/books/CommNetwork. pdf, 2005.

[15] B. A. Forouzan and S. C. Fegan, *Data communications and networking*, 4th ed. NewYork: The McGraw-Hill Companies, Inc, 2007.

[16] P. Simoneau, *The OSI Model: understanding the seven layers of computer networks*. www.globalknowledge.com: Global Knowledge Training LLC, 2006.

[17] S. R. Wilkins, *Designing for Cisco internetwork solutions (Desgn) foundation learning guide*, 3rd ed. Indianapolis: Cisco Press, 2012.

[18] V. Karman, "Understanding downtime, A Vision solutions white paper," California, 2006.

[19] S. Pertet and P. Narasimhan, "Causes of failure in web applications," in *Research Showcase in CMU*, 2005, p. 48.

[20] K. Takeshita, M. Yokota, and K. Nishimatsu, "Early network failure detection system by analyzing twitter data," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 279–286.

[21] M. a. Mohamed, O. G. Altrafi, and M. O. Ismail, "Relational vs. NoSQL databases: A survey," *Int. J. Comput. Inf. Technol. (IJCIT )*, vol. 03, no. 03, pp. 598–601, 2014.

[22] S. Pospiech, S. Mielke, R. Mertens, K. Jagannath, and M. Stadler, "Exploration and analysis of undocumented processes using heterogeneous and unstructured musiness data," in *IEEE International Conference on Semantic Computing Exploration*, 2014, pp. 191–198.

[23] I. Neeman and B. H. Lovering, "Executing structured queries on text records of unstructured data," in *U.S. Patent No. 20,150,149,496*, 2015.

[24] S. Reissmann, D. Frisch, C. Pape, and S. Rieger, "Correlation and consolidation of distributed logging data in enterprise clouds," *Int. J. Adv. Internet Technol.*, vol. 7, no. 1, pp. 39 – 51, 2014.

[25] S. Geetha and G. Anandha Mala, "Effectual extraction of data relations from unstructured data," in *Th ird International Conference on Sustainable Energy and Intelligent System,VCTW, Tiruchengode, Tamilnadu, India*, 2012.

[26] A. Bacchelli, N. Bettenburg, and L. Guerrouj, "Workshop on mining unstructured data (MUD) ... Because 'Mining unstructured data is Like fishing in muddy waters'!," in *19th Working Conference on Reverse Engineering. IEEE*, 2012, pp. 5–6.

[27] F. S. Gharehchopogh, "Approach and review of user oriented interactive data mining," in *4th International Conference on Application of Information and Communication Technologies. IEEE*, 2010, pp. 1–4.

[28] F. S. Gharehchopogh, "Approach and developing data mining method for spatial applications," in *International Conference on Intelligent Sysyems and Data Processing (ICISD)*, 2011, pp. 342–344.

[29] L. Huo, Y. Fang, and H. Hu, "Dynamic service replica on distributed data mining grid," in *International Conference on Computer Science and Software Engineering*, 2008, pp. 390–393.

[30] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in *5th International Conference on Application of Information and Communication Technologies (AICT)*, 2011, pp. 1–4.

[31] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, 2009.

[32] H. H. Malik and V. S. Bhardwaj, "Automatic training data cleaning for text classification," in *11th IEEE International Conference on Data Mining Workshops*, 2011, pp. 442–449.

[33] K. Nithya, P. C. D. Kalaivaani, and R. Thangarajan, "An enhanced data mining model for text classification," in *nternational Conference on Computing, Communication and Applications (ICCCA)*, 2012, pp. 1–4.

[34] J.-C. Lamirel and P. Cuxac, "Improving textual data classification and discrimination using an ad-hoc metric: Application to a famous text discrimination challenge," in *4th IEEE International Symposium Concepts and Tools for knowledge Management (ISKO-Maghreb)*, 2014.

[35] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, 2010.

[36] L. Dan, L. Lihua, and Z. Zhaoxin, "Research of text categorization on WEKA," in *3rd IEEE International Conference on Intelligent System Design and Engineering Applications (ISDEA)*, 2013, pp. 1129–1131.

[37] G. Wei, X. Gao, and S. Wu, "study of text classification methods for data sets with huge features," in *2nd international conference on inustrial and information systems*, 2010, pp.

433–436.

[38] S. L. Bang, J. D. Yang, and H. J. Yang, “Hierarchical document categorization with k-NN and concept-based thesauri,” in *11th International Conference of String Processing and Information Retrieval (SPIRE). Padova. Italy*, 2006, pp. 387–406.

[39] J. W. Kim, B. H. Lee, M. J. Shaw, H. L. Chang, and M. Nelson, “Application of decision-tree induction techniques to personalized advertisements on internet storefronts.,” *Int. J. Electron. Commer.*, vol. 5, no. 3, pp. 45–62, 2001.

[40] M. R. Murty, J. V. R. Murthy, P. Reddy, and S. C. . Satapathy, “A Survey of cross-domain text categorization techniques,” in *1st IEEE International Conference on Recent Advances in Information Technology (RAIT)*, 2012.

[41] J. He, A.-H. Tan, and C.-L. Tan, “A comparative study on chinese text categorization methods,” 2000.

[42] T.-Y. Wang and H.-M. Chiang, “One-against-one fuzzy support vector machine text categorization classifier,” in *IEEE IEEM*, 2008, pp. 1519–1523.

[43] Y. Liu and Y. F. Zheng, “One-against-all multi-class SVM classification using reliability measures,” in *International Joint Conference on Neural Networks*, 2005, pp. 849–854.

[44] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines.,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–25, 2002.

[45] P. Pawar and S. H. Gawande, “A Comparative study on different types of approaches to text classification,” in *3rd International Conference on Machine Learning and Computing (ICMLC)*, 2011, pp. 423–426.

[46] P. Cunningham and S. J. Delany, “K -Nearest Neighbour classifiers,” 2007.

[47] Y. Liao and V. R. Vemuri, “Use of k-nearest neighbor classifier for intrusion detection,” *Comput. Secur.*, vol. 21, no. 5, pp. 439–448, 2002.

[48] S. Oleiwi, “Enhanced ntology-based text classification algorithm for structurally organized documents suha sahib oleiwi doctor of philosophy permission to use,” 2015.

[49] T. G. Dietterich, “Ensemble methods in machine learning.,” 2000.

[50] G. Kaur and A. Chhabra, “Improved J48 classification algorithm for the prediction of diabetes,” *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, 2014.

[51] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 5, pp. 1–35, 1999.

[52] D. . Liparas, Y. . HaCohen-Kerner, A. . Moumtzidou, S. . Vrochidis, and I. . Kompatsiaris, “News articles classification using random forests and weighted multimodal features,” in *In Multidisciplinary Information Retrieval, Springer International Publishing*, 2014, pp. 63–75.

[53] charu c. Aggarwal and cheng xiang Zhai, *Mining text data. Chapter six of the book XII, 524*. 2012.

[54] N. K. Korada, N. S. P. Kumar, and Y. V. N. H. Deekshitulu, “Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm using Maize ExpertSystem,” *Int. J. Inf. Sci. Tech.*, vol. 2, no. 3, pp. 63–75, 2012.

[55] E. Frank and R. R. Bouckaert, “Naive bayes for text classification with unbalanced classes,” in *PKDD’06 Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*, 2006, pp. 503–510.

[56] P. Achananuparp, Xiaohua Zhou, Xiaohua Hu, and Xiaodan Zhang, “Semantic representation in text classification using topic signature mapping,” in *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1034–

1040.

[57] B. Harish, D. Guru, and S. Manjunath, "Representation and classification of text documents: A brief review," *IJCA, Spec. Issue Recent Trends Image Process. Pattern Recognit.*, no. 2, pp. 110–119, 2010.

[58] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," *Int. J. Comput. Appl.*, vol. 85, no. 17, pp. 42–45, 2014.

[59] K. Celik and T. Gungor, "A comprehensive analysis of using semantic information in text categorization," in *Paper presented in IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 2013, pp. 1–5.

[60] M. Radovanović and M. Ivanovi, "Text mining: Approaches and applications," *Novi Sad J. Math*, vol. 38, no. 3, pp. 227–234, 2008.

[61] Y. Chen, "New Feature Selection Methods Based on Context Similarity for Text Categorization," in *11th International Conference on Fuzzy Systems and Knowledge Discovery New*, 2014, pp. 598–604.

[62] Y. Xu and L. Chen, "Term-frequency based feature selection methods for text categorization," in *4th IEEE International Conference on Genetic and Evolutionary Computing*, 2010, pp. 280–283.

[63] F. Yigit and omer kaan Bayakan, "A new feature selection method for text categorization based on information gain and particle swarm optimization," in *3rd IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2014, pp. 523–529.

[64] A. Shadvar and A. Erfanian, "Mutual information-based fisher discriminant analysis for feature extraction and recognition with applications to medical diagnosis.," in *32nd Annual International Conference of the IEEE EMBS*, 2010, pp. 5811–5814.

[65] S. Arani and S. Mozaffari, "Genetic-based feature selection for spam detection," in *21st IEEE Iranian Conference on Electrical Engineering (ICEE)*, 2013.

[66] H. Chen, "Partially supervised learning for radical opinion identification in hate group Web forums," in *IEEE International Conference of Intelligence and Security Informatics (ISI)*, 2012, pp. 96–101.

[67] M. Maleki, "Utilizing category relevancy factor for Ttxt categorization," in *IEEE 2nd International Conference on Software Engineering and Data Mining (SEDM)*, 2010, pp. 334–339.

[68] X. Gang and X. Jiancang, "Performance analysis of chinese webpage categorizing algorithm Based on support vector machines ( SVM )," in *IEEE Fifth International Conference on Information Assurance and Security Performance*, 2009, pp. 231–235.

[69] F. Xia, T. Jicun, and L. Zhihui, "A text categorization method based on local document frequency," in *IEEE Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, pp. 468–471.

[70] N. Yusof and C. J. Hui, "Determination of Bloom ' s Cognitive Level of Question Items using Artificial Neural Network," in *10th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2010, pp. 866–870.

[71] F. Harrag, E. El-qawasmah, A. M. S. Al-salman, and S. Arabia, "Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm," in *IEEE First International Conference on Integrated Intelligent Computing Comparing*, 2010, pp. 6–11.

[72] K. Shiomoto, "Technologies for traffic and network management data applications of big data analytics technologies for traffic and network management data gaining useful insights from big data of traffic and network management," *NTT Tech. Rev.*, vol. 11, no. 11, pp. 1–6, 2013.

[73] K. Yamanishi and Y. Maruyama, "Dynamic syslog mining for network failure monitoring," in *11th ACM SIGKDD International Conference on Knowledge Discovery in dData Mining*, 2005, pp. 499–508.

[74] C. Lim, N. Singh, and S. Yajnik, "A log mining approach to failure analysis of enterprise telephony systems," in *International Conference on Dependable Systems & Networks: Anchorage, Alaska*, 2008, pp. 398–403.

[75] E. Martinez, E. Fallon, S. Fallon, and M. Wang, "ADAMANT - an anomaly detection algorithm for mAintenance and network troubleshooting," in *1st IFIP/IEEE IM Workshop on Cognitive Network & Service Management (CogMan)*, 2015, pp. 1292–1297.

[76] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, "Detecting large-scale system problems by mining console logs," *22nd ACM SIGOPS Symp. Oper. Syst. Princ. SOSP*, vol. 10, no. 7, p. 117, 2009.

[77] M. Liu and J. Yang, "An improvement of TFIDF weighting in text categorization," in *International Conference on Computer Technology and Science (ICCTS)*, 2012, vol. 47, pp. 44–47.

[78] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Am. Stat. Assoc. Am. Soc. Qual.*, vol. 49, no. 3, pp. 291–304, 2007.

[79] V. Maurya, P. Pandey, and L. S. Maurya, "Effective information retrieval system," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 4, pp. 787–792, 2013.

[80] F. Ag, S. Rakshit, and C. V. R. Nagar, "Feature selection using bag-Of-visual-words representation," in *2nd IEEE International Advance Computing Conference (IACC)*, 2010, pp. 151–156.

[81] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, *WEKA Manual for Version 3-6-13*. University of Waikato, Hamilton, New Zealand, 2015.