ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

Design Demanding Situations and Misconceptions in Named Entity Reputation

B.Manisha[1] , A.Nivetha[2], V.Gnanasekar[3]

UG, T.J.S Engineering College, Chennai, India [1,2.]

Associate Professor, T.J.S Engineering College, Chennai, India.[3.]

ABSTRACT: Maximal automation of routine IT maintenance procedures is an ultimate goal of IT service management. System monitoring, an effective and reliable means for IT problem detection, generates monitoring ticket. In light of the ticket description, the underlying categories of the IT problem are determined, and the ticket is assigned to the corresponding processing teams for problem resolving. Automatic IT problem category determination acts as a critical part during the routine IT maintenance procedures. In practice, IT problem categories are naturally organized in a hierarchy by specialization. Utilizing the category hierarchy, this paper comes up with a hierarchical multi-label classification method to classify the monitoring tickets. In order to find the most effective classification, a novel contextual hierarchy (CH) loss is introduced in accordance with the problem hierarchy. Consequently, an arising optimization problem is solved by a new greedy algorithm named GLabel. Furthermore, as well as the ticket instance itself, the knowledge from the domain experts, which partially indicates some categories the given ticket may or may not belong to, can also be leveraged to guide the hierarchical multi-label classification. Accordingly, a multi-label inference with the domain expert knowledge is conducted on the basis of the given label hierarchy. The experiment demonstrates the great performance improvement by incorporating the domain knowledge during the hierarchical multi-label classification over the ticket data.

KEYWORDS: System monitoring, Classification of monitoring data, Hierarchical multi-label classification, Domain Knowledge**.**

I. INTRODUCTION

A. Background

CHANGES in the economic environment force companies to constantly evaluate their competitive position in the market and implement innovative approaches to gain com- petitive advantages. Without solid and continuous delivery of IT services, no value-creating activities can be executed. The complexity of IT environments dictates usage of analytical approaches combined with automation to enable fast and efficient delivery of IT services. Incident management, one of the most critical processes in IT Service Management [1], aims at resolving the incident and quickly restoring the provision of services while relying on monitoring or human intervention to detect the malfunction of a component. Thus, it is essential to provide an efficient architecture for the IT routine maintenance.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

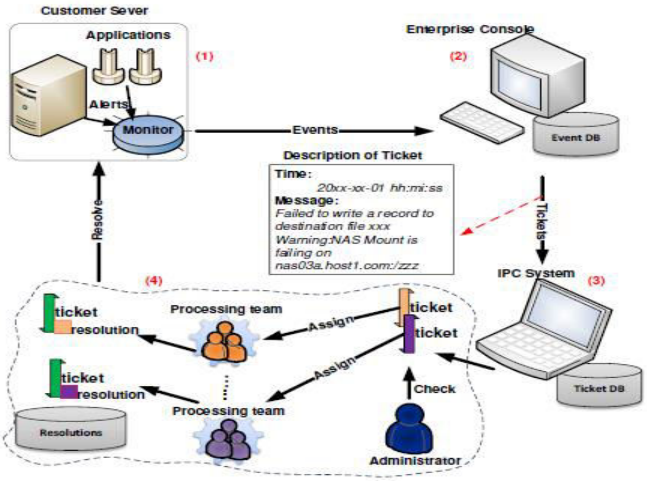

Fig. 1: The overview of the IT routine maintenance procedure.

A typical architecture of the IT routine maintenance is illustrated in Fig. 1, where four components are involved. (1) In the case of detection provided by a monitoring agent on a server, alerts are generated and, if the alert persists beyond a predefined delay, the monitor emits an event. (2) Events coming from an entire account IT environment are consoli- dated in an enterprise console, which analyzes the monitoring events and determines whether to create an incident ticket for IT problem reporting. (3) Tickets are collected by IPC (abbr. Incident, P roblem and Change) system and stored in the ticket database [2]. (4) A ticket accumulates the symptom description of an IT problem with a short text message and a time stamp provided. According to the description of a ticket, the system administrators (i.e., sysAdmins) perform the problem category determination and assign the ticket to its corresponding processing teams for problem diagnosis and resolution. The last component gets involved with much labor- intensive effort to resolve each ticket. The efficiency of these transient resources is critical for the provisioning of the services [3]. Many IT Service Providers rely on a partial automation for incident diagnosis and res- olution, with an intertwined operation of the sysAdmins and an automation script. Often the sysAdmins' role is limited to executing a known remediation script, while in some sce- narios the sysAdmin performs a complex root cause analysis. Removing the sysAdmin from the process completely where it is feasible would reduce human error and speed up restoration of service. The move from partially to fully automated problem remediation would elevate service delivery to a new qualitative level where automation is a complete and independent process, and where it is not fragmented due to the need for adapting to human-driven processes. The sysAdmin involvement is required due to the ambiguity of service incident description in a highly variable service delivery environment.

B. Motivation

In order to enhance the efficiency of the routine IT maintenance procedure, our work focuses on the labor-intensive component and tries to reduce human involvement by maximizing the automation of the problem category determination. The *Kilo* algorithm is not only capable of fully exploring both domain knowledge and the data itself, but also provides an effective way for interactive hierarchical multi-label learning. Concretely, based on the current hierarchical multi-label classification result, the experts provide expertise to hint *Kilo* for further refinement. *Kilo* takes the hints from the experts as an input to refine the hierarchical multi-label inference. This iterative process continues until the domain experts are satisfied with the hierarchical multi-label classification result.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

II. RELATED WORK

In this section, we highlight existing literature studies related to our work. In the literature of IT management, [4] In this paper they distinguish control flows from data flows, and introduce a novel dynamic control plane architecture to distribute different control flows among multiple controller instances depending on specific controller load and controller processor utilization or on the data flow service type. They propose "control flow tables"a concept introduced in this paper—that are embedded in OpenFlow flow tables to distribute the control flows among various controller instances.[5] In this paper they define a capacitated controller placement problem (*CCPP*), taking into consideration the load of controllers, and introduce an efficient algorithm to solve the problem. The evaluation shows that the new strategy can significantly reduce the number of required controllers, reduce the load of the maximum-load controller, and reduce the radius stretches compared with the K-center strategy with dynamic controller provision or dynamic scheduling.In this paper they present an elastic-net regularized linear regression (ENLR) framework, and develop two robust linear regression models which possess the following special characteristics. First, our methods exploit two particular strategies to enlarge the margins of different classes by relaxing the strict binary targets into a more feasible variable matrix. Second, a robust elastic-net regularization of singular values is introduced to enhance the compactness and effectiveness of the learned projection matrix. Third, the resulting optimization problem of ENLR has a closed-form solution in each iteration, which can be solved efficiently. Finally, rather than directly exploiting the projection matrix for recognition, our methods employ the transformed features as the new discriminate representations to make final image classification. Compared with the traditional linear regression model and some of its variants,
our method is much more accurate in image classification. Extensive experiments conducted on publicly available datasets well demonstrate that the proposed framework can outperform the state-of-the-art methods.[6] In this paper they presents POCO, a framework for Pareto-based Optimal COntroller placement that provides operators with Pareto optimal placements with respect to different performance metrics. In its default configuration, POCO performs an exhaustive evaluation of all possible placements.While this is practically feasible for small and medium sized networks, realistic time and resource constraints call for an alternative in the context of large scale networks or dynamicsnetworks whose properties change over time. For these scenarios, the POCO toolset is extended by a heuristic approach that is less accurate, but yields faster computation times. An evaluation of this heuristic is performed on a collection of real world network topologies from the Internet Topology Zoo. Utilizing a measure for quantifying the error introduced by the heuristic approach allows an analysis of the resulting trade-off between time and accuracy. Additionally, the proposed methods can be extended to solve similar virtual functions placement problems which appear in the context of Network Functions Virtualization (NFV).[7] In this paper they opens the investigation by focusing on two specific questions: given a wide-area network topology, how to partition it into several small SDN domains, and where should the controller go in each SDN domain they propose a novel approach to efficiently and accurately evaluate SDN controller placement problem for WAN. This approach uses the Spectral Clustering placement algorithm to partition a large network into several small SDN domains. After presenting our metrics of the SDN controller placement problem, we evaluate our approach using the Internet2 topology and other available WAN topologies. For testing purposes we developed a new framework to test the SDN controller in WAN. The results show its effectiveness for the SDN controller placement problem. Here a large number of analytical methodologies based on data mining and machine learning have been explored to optimize the routine IT maintenance and deal with problem detection, determination, diagnosis and resolution ([15], [16], [17], [18], [19], [20]).

III. ALGORITHM

ttLabel(**H**) **H** is the label hierarchy, with σ available
1: define L as a set, and initialize $L = 0$
2: define U as a set, and initialize $U =$ **H** 0
3: **while** TRUE **do**
4:**if** all the nodes in **H** are labeled **then**
5:**return** L
6:**end if**
7:find the node i with maximum $\sigma(i)$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

8:**if** $\sigma(i) < 0$ **then**
9:**return** L
10:**end if**
11:**if** all the parents of i are labeled **then**
12:put i into L, and remove it from U
13: **else**
14:merge i with its parent as a super node i
15:$\sigma(i)$ =average σ values of the two nodes
16:put the i^* into U
17:**end if**
18:**end while**

Since the labeling procedure for each node may involve a merging procedure, the time complexity is no worse than $O(N\ log(N))$, the same as HIROM. However, as shown in the experimentation section below, *ttLabel* performs more efficiently than HIROM while not requiring knowledge of the maximum number of labels.

IV. EXPERIMENTATION

A. Setup

We perform the experiment over the ticket data set generated by monitoring of the IT environments of a large IT service provider. The number of tickets in the experiment amounts to about 23,000 in total. The experiment is executed 10 times and we use the mean value of the corresponding metric to evaluate the performance of our proposed method. At each time, 3000 tickets are sampled randomly from the whole ticket data set to build the testing data set, while the rest of the tickets are used to build the training data set. The class labels come from the predefined catalog information for problems occurring during maintenance procedures. The whole catalog information of problems is organized in a hierarchy, where each node refers to a class label. The catalog contains class labels; hence there are 98 nodes in the hierarchy. In addition, the tickets are associated with 3 labels on average and the height of the hierarchy is 3 as well.

Then, natural language processing techniques are applied to remove the stop words and build Part-Of-Speech tags for the words in the rephrased text. The nouns, adjectives and verbs in the text are extracted for each ticket. Second, we compute the TF-IDF [50] scores of all words extracted from the text of tickets. And the words with the top 900 TF-IDF score are kept as the features for the tickets. Third, the feature vector of each ticket has 900 components, where value of each feature is the frequency of the feature word occurring in the text of the ticket.Based on the features and labels of the tickets, we build a binary classifier for each node in the hierarchy with the SVM algorithm by using library libSVM. The training data for each node i are the tickets with a positive parent label. To speed up evaluation of the 98 SVM classifiers, we parallelize the process of training classifiers, using the fact that all the classifiers are independent.The experiments are mainly conducted by comparing the proposed the *ttLabel* algorithm with state of the art algorithms such as CSSA and HIROM.

B. Hamming loss

The ttLabel algorithm can obtain optimal $\hat{y}$ with minimum Hamming loss by setting the parameters for Hamming loss, since Hamming loss is a special case of CH-loss. Given $w1 = w2 = w3 = w4 = 1$ for ttLabel, $\alpha = \beta = 1$ for HIROM and $Ci = 1$ for both of them, empirical results ,the ttLabel algorithm can automatically find the optimal $\hat{y}$ with minimum Hamming loss, while both CSSA and HIROM require the number of class labels and the maximum number of class labels, respectively, to get the optimal $\hat{y}$. Particularly, by increasing the value for setting the maximum number of class labels, HIROM incurs smaller Hamming loss. The HIROM algorithm does not reach the optimal $\hat{y}$ with minimum Hamming loss until a sufficiently large value is used for specifying the maximum number of class labels. Moreover, CSSA may incur more Hamming loss as the number of class labels increases and it is difficult to identify the correct number of class labels for each ticket in practice During the Hamming loss minimization, we track the varying performance in terms of precision, recall and F-Measure score.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

C. HMC-Loss

To simplify the empirical study for HMC-Loss, we set $w_1 = w_2 = w_3 = w_4 = 1$ for *ttLabel*, and $\alpha = \beta = 1$ for HIROM. *ttLabel* algorithm automatically obtains the lowest HMC-Loss, which can be achieved by the HIROM algorithm after choosing the proper value for the maximum number of class labels.

D. Experiment with prior knowledge

Setup: In order to demonstrate the effectiveness of the proposed *Kilo* algorithm for knowledge guided hierarchical multi-label classification, we perform the experiments over the same real ticket data set described in section VI-A. The only difference lies in the additional knowledge construction. Each ticket instance in the test data set are associated with a knowledge vector **k**, where the i^{th} component are exposed with its ground true value randomly according to a predefined prior knowledge ratio γ [0, 1]. The prior knowledge ratio γ is the probability that true labels are observed as either *positive* or *negative*, while $1\ \gamma$ denotes the probability that the label can not be observed before classification.The performances in terms of different metrics in comparing the *ttLabel* algorithm without prior knowledge and the one with prior knowledge incorporated by *Kilo* algorithm.

1) ***Hamming loss with changing prior knowledge ratio:*** Similar to previous configuration, we obtain Hamming loss by setting the CH-loss parameters $w_1 = w_2 = w_3 = w_4 = 1$, $C_i = 1$. As illustrated in Fig. 8a, Hamming loss drops as the prior knowledge ratio increases, and reaches to zero as prior knowledge ratio achieves one.Illustrate the performance in terms of Precision, Recall and Precision, respectively. It shows that the knowledge guided algorithm gets better performance as prior knowledge ratio increases.

2) ***HMC-Loss:*** The HMC-Loss is obtained by the same settings as provided in section VI-C. As expected, HMC- Loss decreases as more prior knowledge is provided.

3) H-Loss: The H-Loss is obtained by the same param- eter settings as provided in section VI-D. By increasing the

4) H-Loss: The H-Loss is obtained by the same param- eter settings as provided in section VI-D. By increasing the prior knowledge ratio, H-Loss caused by knowledge guided algorithm decreases.

5) ***Parent-Child Error:*** The Parent-Child Error is obtained by the same parameter settings as presented in section VI-E. By increasing the prior knowledge ratio, Parent-Child errors caused by knowledge guided algorithm decreases.

6) ***CH-Loss:*** The CH-Loss is obtained by the same pa- rameter settings as given in section VI-E. By increasing the prior knowledge ratio, CH-Loss caused by knowledge guided algorithm decreases.

V.CONCLUSION AND FUTURE WORK

In this paper, we employ hierarchical multi-label classifi- cation over ticket data to facilitate the problem diagnosis, determination and an automated action, such as auto-resolution or auto-check for enriching or resolving the ticket in the complex IT environments. CH-loss is proposed by considering the contextual misclassification information to support different scenarios in IT environments. In terms of CH-loss, an optimal prediction rule is developed based on Bayes decision theory. This paper comes up with a greedy algorithm *ttLabel* by extending the HIROM algorithm to label the predicting ticket without knowing the number or the maximum number of class labels related to the ticket. Additionally, taking the real scenario in practice into account, the *Kilo* algorithm is proposed to utilize the knowledge from the domain expert during routine IT maintenance procedure to effectively improve the IT problem category determination.

REFERENCES

[1] "ITIL," http://www.itil-officialsite.com/home/home.aspx, 2016.

[2] L. Tang, T. Li, L. Shwartz, F. Pinel, and G. Y. Grabarnik, "An integrated framework for optimizing automatic monitoring systems in large IT infrastructures," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013, pp. 1249–1257.

[3] Y. Jiang, C.-S. Perng, T. Li, and R. Chang, "Intelligent cloud capacity management," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*. IEEE, 2012, pp. 502–505.

[4] WIKI, 2016, https://en.wikipedia.org/wiki/Network-attached storage.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 4, April 2018

[5] " Dynamic Management of Control Plane Performance in Software Defined Networks" Burak Görkemli, A. Murat Parlakışık, Murat Tekalp.
[6] "On the Capacitated Controller Placement Problem in Software Defined Networks "Guang Yao, Jun Bi, *Member, IEEE*, Yuliang Li and Luyi Guo..
[7] "Discriminative Elastic-Net Regularized Linear Regression" Zheng Zhang, Zhihui Lai, Yong Xu, Senior Member, IEEE,Ling Shao, Senior Member, IEEE,Jian Wu, Guosen Xie 2017.
[8] O. Dekel, J. Keshet, and Y. Singer, "Large margin hierarchical classifi- cation," in *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, 2004, p. 27.
[9] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Incremental algorithms for hierarchical classification," *The Journal of Machine Learning Research*, vol. 7, pp. 31–54, 2006.
[10] N. Cesa-Bianchi and G. Valentini, "Hierarchical cost-sensitive algo- rithms for genome-wide gene function prediction," 2010.
[11] W. Bi and J. T. Kwok, "Multi-label classification on tree-and dag- structured hierarchies," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 17–24.
[12] L. Tang, T. Li, L. Shwartz, and G. Grabarnik, "Recommending Resolu- tions for Problems Identified by Monitoring," pp. 134–142, 2013.
[13] L. Cai and T. Hofmann, "Exploiting known taxonomies in learning overlapping concepts." in *IJCAI*, vol. 7, 2007, pp. 708–713.
[14] C. Zeng, T. Li, L. Shwartz, and G. Y. Grabarnik, "Hierarchical multi- label classification over ticket data using contextual loss," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*. IEEE, 2014, pp. 1–8.
[15] C. Bozdogan and N. Zincir-Heywood, "Data mining for supporting it management," in *2012 IEEE Network Operations and Management Symposium*. IEEE, 2012, pp. 1378–1385.
[16] C. Bozdogan, A. N. Zincir-Heywood, and Y. Gokcen, "Automatic optimization for a clustering based approach to support it management," in *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*. IEEE, 2013, pp. 1233–1236.
[17] S. Sozuer, C. Etemoglu, and E. Zeydan, "A new approach for clustering alarm sequences in mobile operators," in *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*. IEEE, 2016, pp. 1055–1060.
[18] H. M. Tran and J. Schönwälder, "Discariałdistributed case-based rea- soning system for fault management," *IEEE Transactions on Network and Service Management*, vol. 12, no. 4, pp. 540–553, 2015.
[19] C. Zeng, L. Tang, W. Zhou, T. Li, L. Shwartz, G. Grabarnik *et al.*, "An integrated framework for mining temporal logs from fluctuating events," *IEEE Transactions on Services Computing*, 2016.
[20] C. Zeng, L. Tang, T. Li, L. Shwartz, and G. Y. Grabarnik, "Mining temporal lag from fluctuating events for correlation and root cause analysis," in *10th International Conference on Network and Service Management (CNSM) and Workshop*. IEEE, 2014, pp. 19–27.