MACHINE LEARNING BASED AUTOMATIC TROUBLE TICKET ASSIGNMENT FOR EFFICIENT RESOLUTION IN SUBMARINE NETWORK OPERATIONS

Madanagopal Ramachandran, Arun Balaji V N (NMSWorks), Shirshendu Bhattacharya (Google), Krishna M. Sivalingam (Indian Institute of Technology Madras)
Email: madan@nmsworks.co.in

NMSWorks Software Private Limited
#C3, 6th Floor, IIT Madras Research Park, Taramani, Chennai - 600 113, Tamil Nadu, INDIA.

Abstract: Given the critical role played by submarine networks in modern hyper-scalar and cloud operator networks; rapid and accurate diagnosis of problem and expedient resolution are very important for network health. One approach to inject greater efficiency into the process is automation of ticket assignment to Network Operation Center (NOC) personnel based on their area of expertise and past history of ticket handling. We present a new approach using machine learning for automatic trouble ticket categorization and assignments for specific resolution actions. This approach significantly reduces the time taken for the resolution action to commence, while maintaining accuracy comparable to manual categorization assignment.
The proposed approach uses intelligence from mining of past data on ticket handling and machine learning techniques to assign new tickets to the most appropriate network operator. The technique uses topic and correlation mining to map the keywords in ticket resolution comments and assigns it to the most appropriate network operator. Proposed technique's performance is evaluated in terms of number of correctly assigned tickets to the concerned network operator with respect to the total number of tickets handled. It is found that the proposed technique performs 100% accurate assignment for synthetic ticket data when the ticket keywords are mutually disjoint amongst network operators and performs close to 50% accurate assignment for real service provider ticket data over a period of 50 days.

1. INTRODUCTION

As the demand for data is increasing, the submarine networks are continuously expanding to meet the data requirements. Failures that occur in the network have to get resolved in the minimum possible time to meet the guarantees given to the customers. The mean time to repair (MTTR) a failure in the network will be a critical parameter that decides how the network is managed.

Submarine Network Service Providers use advanced Fault Management Systems (FMS) that manage faults from the network and compute the root cause and impact for the events received. Trouble Tickets System is another Operation Support System (OSS) application that is used to manage tickets which are either automatically created for root cause problem or manually created based on the complaint by customer.
Analysing the tickets and assigning to the concerned network operator is typically a manual activity which requires troubleshooting of the root problem and handing over the problem for resolution. The messages and comments entered in the tickets as part of their lifecycle typically do not follow standard conventions. Due to the manual process of assigning a network operator, the time taken for the ticket handling and resolution tend to be high. Automation of the ticket assignment to network operators based on their area of

expertise and past history of ticket handling would help in minimizing the delays due to manual process.

In this work, automation of ticket assignment to network operators based on mining of the ticket description and the operator who resolved the ticket is proposed. The proposed method learns from the past ticket handling by the network operators and assigns new ticket to the most suitable operator automatically.

A combination of topic and correlation mining is used to identify keywords or topics in the ticket description which is then correlated to the user or operator best suited to resolve the ticket. This approach is based on unsupervised learning which does not require explicit labelling of ticket data to resolving user.

The training is carried out on synthetic ticket data containing 1000 tickets where the ticket keywords are mutually disjoint amongst network operators and on real service provider ticket data containing 3500 tickets obtained for 50 days. For the synthetic dataset, 100% match is found for correlation between ticket data and network operator or user. For the real service provider dataset, close to 50% match is found since there is lack of demarcation across which user resolves particular type of tickets.

2. RELATED WORK

The problem of automatic ticket assignment is similar to the problem of bug triaging in open source projects where the suitable developer for resolving a bug is recommended. In [1], the bug triaging problem is formulated using machine learning as a classification task where the terms in the bug report instance are features and it is labelled with the developer who fixed that bug.

The problem of classifying e-mails received in a helpdesk by system administrators by grouping the related text messages into clusters is addressed in [2]. K-means clustering is used to group the word vectors generated from the input message after removing unimportant messages.

The problem of classifying user e-mail queries in real time by using text mining techniques is addressed in [3]. A method for pseudo-randomly responding to the user queries is also proposed.

In [4], the problem of identifying the issues or problems from the trouble ticket description data is described. Information retrieval techniques along with domain knowledge are used to extract the key issues or problems from the ticket description and it is shown that the proposed approach is able to extract most of the key issues.

In IT infrastructure, a database of customer problems and solutions is created based on past inputs and a cognitive IT system that extracts knowledge about different problems from the database is proposed in [5]. This enables identification of root cause for the problems using which pro-active remedial measures are possible.

In [6], the problem of identification of server names from unstructured ticket text so that it can be used for linking various information sources is addressed. A machine learning method namely Conditional Random Field (CRF) is used to identify the server names with high accuracy.

All the related work use some sort of text mining and machine learning techniques to get insight from the textual content like ticket data. Few of the mentioned work use supervised learning techniques where explicit labelling of data in training is required before they can give results for the test data. In this work, unsupervised learning in the form of topic mining and correlation mining is used for automatic ticket assignment.

3. PROPOSED APPROACH

In order to enable automatic assignment of tickets to network operators, the ticket description has to be mined for keywords that distinguish a given category of problem. The mining of keywords is performed in this work with topic mining. The next step is to correlate the topics with the network operators who are most suited to resolve the problems based on the identified topics. Such correlation is done using a text correlation mining algorithm.

a. Topic Mining

The most popular approach for topic mining is the Latent Dirichlet Allocation (LDA) [7]. In LDA, a generative probabilistic model for discrete collection of text content is described. This method identifies the set of topics that capture the key information from the given set of documents. It provides top-N topics for each document. The algorithm also outputs the probability of each topic in a given document. This model is an evolution from the earlier work on topic mining namely Probabilistic Latent Semantic Analysis (PLSA) [8].

b. Correlation Mining

Once the topics are identified, the next step is to map them with the network operator who is most suitable for resolving the problem. Correlation mining addresses this requirement. One of the techniques proposed for doing text correlation is the NICoMiner [9]. This work presents new properties for null-invariant measures and adopts pruning techniques for doing the correlation. Based on the new properties, an algorithm that ranks correlated patterns in the given text content is proposed.

c. Ticket Assignment Process

The automatic ticket assignment process is carried out in two steps namely training and testing or validation. The total number of tickets is divided into two sets namely training set and test set based on the usually followed 80-20 division rule where 80% of the tickets are used for training and the remaining 20% are used for testing or validation.

d. Training Step

The first step in the training process is to use LDA to find the top M topics from the set of tickets such that each topic contains W keywords. The topics and keywords are identified from the ticket description after removing stop words. The keywords obtained from the LDA result are then combined into a set of unique keywords named as keywords set.

The next step in the training process is to find correlation using NICoMiner between ticket resolving user and the keywords in the tickets. For this step, the ticket data in the training set is divided into multiple sets where each set corresponds to a ticket resolving user along with the corresponding ticket description. This corresponds to a transaction database which is generated user wise. Each transaction in this database is generated for each ticket in the set by performing intersection of the set of the words in the ticket description with the keywords identified from the topics. The user name is added to each transaction in the database. For each transaction database, NICoMiner algorithm is executed which gives Top K correlated patterns between the user and the topic keywords from the ticket data. The correlation measure used in the NICoMiner algorithm namely cosine is also obtained for each correlated pattern.

e. Testing or Validation Step

The testing or validation process is used to check whether the assignment of tickets to the correlated user in the test data matches with the actual user who resolved those tickets. It is executed on the test data. For each item in the test data, the ticket description is split into tokens or words which is then matched with the user-wise correlated patterns. The following measure is to compute the score for the matching:

$$score = (\alpha * C / C_{max}) + ((1 - \alpha) * cosine)$$

where C represents the number of words where there is a match between the tokens or words in the ticket description and the keywords in the correlated pattern, C_{max} represents the size of the total keywords set, cosine represents the cosine values from NICoMiner for the correlated pattern and α represents a value between 0 and 1. This score gives a weighted sum of normalized values of correlated match count and cosine. For each test data, the scores obtained are sorted in descending order and the Top N patterns are selected. The user corresponding to the Top N patterns are compared with the actual user who resolved that ticket. If there is a match, then the match count is incremented by one. Finally, the percentage of match count is found which indicates the possible correct assignment of ticket to the ticket resolving user.

4. PERFORMANCE RESULTS

The performance of the proposed approach is evaluated and results obtained are provided in this section. The evaluation is performed on two types of ticket dataset: a synthetic dataset and real dataset from a network service provider.

The synthetic dataset is simulated to contain 1000 tickets. Each ticket is generated with a ticket description that contains 50 random words from the English dictionary and between 3 to 8 random technical words from the service provider network problem description. Each ticket description is associated with a resolving user name generated randomly among 5 possible ticket resolvers. The ticket description for the 5 resolving users are generated such that the technical words are mutually disjoint.

The synthetic dataset is split into training and test data in the 80-20 ratio. The topic mining is performed to extract the keywords from the training data and it is found to successfully find the technical words from the ticket description. Correlation mining is executed using the approach proposed in the previous section such that for resolving user, Top 20 correlations between the user name and the topic keywords in the ticket description are found. During this step, a support threshold (defined in [9]) of 5.0 is used.

During the testing step, for each ticket in the test dataset, a match between the topic keywords in the ticket and the correlated keywords for each user is performed and the resulting scores are ranked. For the maximum score, the user name is compared with the actual user who resolved that ticket. It is found that the user name match is 100% for the synthetic data when the technical words in the ticket description is mutually disjoint among the resolving users. This validates the correctness of the proposed approach.

For the real dataset from the network service provider, the same evaluation is performed. The real dataset contains 3500 tickets collected over a one month period. This dataset is also split into training and test data using the 80-20 split ratio. Topic mining is used to extract the keywords from the ticket description and the correlation mining is executed to find Top 20 correlations for each ticket resolving user. During the correlation step, a support threshold of 5.0 is used.

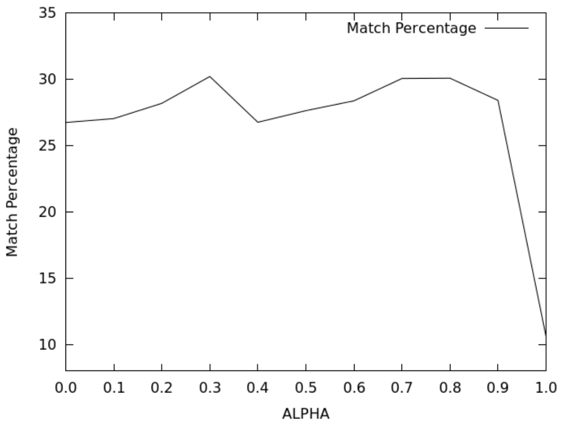

Fig. 1: Match for Top 3 Correlation

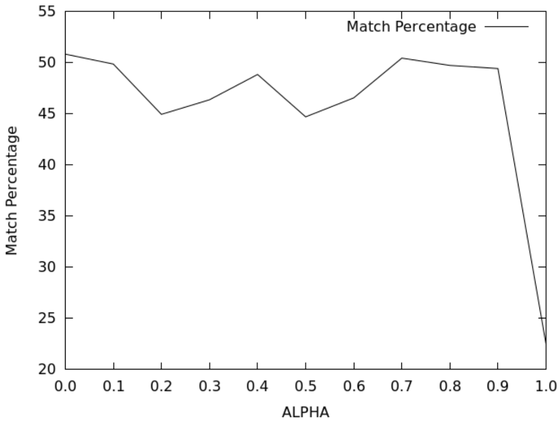

Fig. 2: Match for Top 5 Correlation

During the testing step, for each ticket data, the top correlation obtained was used to compare the corresponding user with the actual user who resolved the ticket. Since in real networks, the mutually disjoint data for the keywords is not possible due to the fact that multiple users would be capable of resolving a particular problem. Therefore, Top 3 and Top 5 scores are taken and compared for the match between the corresponding users and the actual resolving user. The match percentage results obtained for Top 3 and Top 5 scores for different values of α ranging from 0 to 1 in increments of 0.1 are shown in Fig. 1 and Fig. 2.

From the results obtained, it is found that the match percentage for Top 3 and Top 5 correlations are around 30% and 50% respectively. This is because the service provider network considered is small and due to lack of demarcation across the problem type and the users who resolve them, the user names from the output of the proposed approach do not match the actual user who resolved the ticket. If user names from the Top 8 correlations are considered, then the match percentage increased to 75%.

5. CONCLUSIONS

In this work, the problem of automatic assignment of tickets to the user best suitable to resolve the tickets is considered. A machine learning based approach which uses a combination of topic mining and correlation mining is proposed. The proposed approach is based on unsupervised learning which obviates the need for explicit labelling of ticket data to the resolving user.

During the evaluation of the proposed approach, it is observed that when the keywords are correlated with the users in a mutually disjoint way, the match percentage is 100%. For the evaluation of the real ticket data from a network service provider, the match percentage is around 50% when Top 5 correlations are considered. This is due to the smaller service provider network considered where a problem type is resolved by multiple users.

As part of future work, it is planned to implement automatic ticket assignment such that it adapts dynamically to explicit feedback from the assigned network operator if the ticket assignment is sub-optimal. Also, it is planned to use other metrics like mutual information for computation of correlation between ticket keywords and network operators or users.

ACKNOWLEDGEMENT
The authors would like to thank Dr. V. R. Sitaraman for his valuable inputs and guidance.

6. REFERENCES

[1] M. Alenezi, K. Magel, S. Banitaan, “Efficient Bug Triaging Using Text Mining”, JSW Vol. 8, No. 9, 2013, pp. 2185-2190.
[2] N. E. Eide, A. N. Blaafadt, B. H. R. Johansen, F. E. Sandnes, "DIGIMIMIR: A Tool for Rapid Situation Analysis of Helpdesk and Support E-mail", In LISA, 2004, pp. 21-32.
[3] A. N. Blaafladt, B. R. Johansen, N. E. Eide, F. E. Sandnes, “A text mining approach to helpdesk and e-mail support”, In NIK, 2004.
[4] V. Shimpi, M. Natu, V. Sadaphal, V. Kulkarni, “Problem identification by mining trouble tickets”, In 20th International

Conference on Management of Data, 2014, pp. 76-86.
[5] S. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta, G. Sridhara, “Automatic problem extraction and analysis from unstructured text in IT tickets”, IBM Journal of Research and Development, Vol. 61, No. 1, 2017, pp. 4-41.
[6] E.E. Jan, J. Ni, N. Ge, N. Ayachitula, X. Zhang, “A statistical machine learning approach for ticket mining in IT service delivery”, In IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), pp. 541-546.
[7] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent dirichlet allocation”, Journal of machine Learning research, 2003, pp.993-1022.
[8] T. Hofmann, “Probabilistic latent semantic analysis”, In Fifteenth conference on Uncertainty in artificial intelligence, 1999, pp. 289-296.
[9] S. Kim, M. Barsky, J Han. "Efficient mining of top correlated patterns based on null-invariant measures", In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 177-192.