DISTRIBUTION COMMUNICATION SUBSTRUCTURE FAULT ANALYSIS BY DATA MINING

Bahman Jamshidi Eini
Sharif university of technology, Alborz PPDC – Iran
b_jamshidieini@ie.sharif.edu

Mofid Ahmadi Asl
Alborz PPDC– Iran
mofid_ahmadi@gmail.com

Reza Vasigh
Alborz PPDC– Iran
rezavasigh@gmail.com

Hassan Shafiei
Alborz PPDC– Iran
hasan.shafei@gmail.com

ABSTRACT

The communication between control centers, contact centers, field technicians and customers have been changed in previous decade and traditional analogue communication devices have been being replaced by integrated digital systems. Nowadays communication and computer network substructures are vital to distribution companies because communication system interruption can disrupt majority of services. Therefore, it is necessary to control any situation which can threat communication system. Alborz province power distribution company uses data mining methods like classification and clustering in addition to quantitative statistics for analyzing historical fault data. The result of this study is useful for evaluating the effect of investment on reducing system disruption. In this paper, after introducing methods which were used in this study, the concise result of data mining was presented.

INTRODUCTION

Customer trend toward using multimedia tools such as World Wide Web and Email motivates distribution companies to upgrade their traditional call-centers to interactive multimedia contact-centers. Computer networks and communication substructures are used in distribution companies not only for making connection between dispatching control-centers, high voltage substations, remote controlled switches and other network components, but also between customers and contact-centers. In addition, distribution companies use contact-center for real-time outage alarm and managerial purposes. These contact centers consist of local and wide area networks (LAN and WAN), network switches, databases, servers, Interactive voice response (IVR), automatic call distributor, Internet protocol (IP) based phones and peripheral interfaces like Fax server, SMS server and Email Server. Although high technology devices used in communication and computer networks benefit distribution companies, their vulnerability can reduce network reliability and customer satisfaction. These treats can be categorized into two types: external threats like storms and other meteorological events and internal threats, including software and hardware problems. According to significance of communication and computer network for contact-centers and dispatching control-rooms, evaluating vulnerability of these networks is recommended to distribution companies. Therefore, emergency service department of Alborz province power distribution company has started a data mining research project for interpreting relation between recorded faults. After understanding this relation, it is possible to evaluate the networks reliability and their susceptibility to frequent threats.

In the first stage of this project, about eight hundred recorded communication and data network faults were clustered by partitioning and hierarchical methods. Hierarchical clustering method constructs the clusters by recursively partitioning the data. Each event initially represents a cluster. Then the similar clusters are merged together until a few manageable clusters remain. K-Means is a partitioning method which relocates an event from one cluster to another one for finding a clustering structure with minimum Euclidian distance. In last stage, apriori association rule learning was used for discovering relations between events and their sources. Association rules generated by data mining procedure represent important relations not only between faults and their sources, but also between two faults. Although these rules derived from Alborz province events is not valid for other distribution companies, the procedure used in this project can be applied to other event database, including communication and electrical network faults.

DATA MINING TOOLS

There are many tools which can be used for examining data and Clustering and association rule learning are famous unsupervised methods selected for current project.

Supervised and unsupervised methods

There are two main classes of tools which are used in data mining [1]:

I-Supervised methods

II- Unsupervised methods

Supervised methods are used for finding the relationship between input attributes and target attributes. Model is a discovered relationship which describes hidden interconnection in the data set and can be used for predicting the value of the target attribute, if the values of the input attributes are known. There are two main supervised models. Classifiers mapping input space into pre-defined classes and regressor (regression model) which maps the input space into a real-value domain [2]. The target of unsupervised data mining can be data reduction,

clustering or extracting relationship between variables. There is no difference between input and target variables in unsupervised methods. Supervised methods require sufficient number of well defined target variables but unsupervised tools like clustering and association rule learning can be used if the target variables are unknown or a few records are available.

Association rule learning

Association rule learning is one of the most important methods for finding rules representing relation between variables. Usefulness of a rule can be measured with a "support" which measures how many events match both sides of the implication in the association rule. Certainty of a rule can be measured with "confidence" which measures how often an event that matches the left side of the implication in the association rule also matches for the right side. Generation of Association rule consists of two steps: In first step, all object sets repeated more than minimum support are found and in next step, objects set which their confidence is more than predefined limit are used to make rules. In current study Apriori is used for rule generation.

Clustering

Clustering is an unsupervised tool for decomposing a set of objects into subgroups or clusters based on similarity. The objects in same cluster are as similar as possible, but objects of different clusters are as different as possible. Clustering is a powerful tool for discovering hidden structure in a set of unordered data. In addition, clustering is used for data reduction. It is possible to use manageable number of homogeneous groups instead of numerous single objects [3]. Because clustering is the grouping analogous objects, a criterion which can determine the level of similarity, is needed. Distance measure and similarity measure are two criteria for estimating this relation [2].

Clustering methods can be divided into two main groups: hierarchical and partitioning methods. Hierarchical methods can be sub-divided as agglomerative and divisive. If each object initially represents a cluster of its own and then clusters are successively merged, the technique is called agglomerative hierarchical clustering. In contrast, if all objects initially belong to one cluster and then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters, the method is divisive hierarchical clustering. In both approaches, process continues until the desired cluster structure is obtained. The result of the hierarchical clustering is called dendrogram, which represents the nested grouping of objects and similarity levels at which groupings change. For obtaining preferred cluster structure, dendrogram is cut at a desired similarity level [2]. There are many criteria which can be used for determining how the distance between two clusters is defined. In current research ward's method was applied. This method which used the sum of squared deviations from points to centroids as linkage criterion, have a tendency to produce clusters with similar numbers of items, but it is susceptible to outliers.

Partitioning methods are based on relocating object by moving them from one cluster to another for minimizing error criterion. K-means algorithm used in current study is the simplest partitioning algorithm. Although linear complexity of this method is a reason of its attractiveness, this algorithm is sensitive to noisy data and outliers too. In addition, the number of clusters is needed in advance, which is hard to find when prior knowledge is not available [2].

OTHER ASPECTS OF COMMUNICATION AND COMPUTER NETWORK FAULTS

Communication system vulnerability is not only depends on faults frequency, but also on time needed for system restoration, domain of disruption caused by faults and redundancy.

If we want to go into more detail, there are redundant units in call-centers, if one of them is interrupted, redundant one can be used so whole system works without considerable interruption, but there are not redundant communication links in small agents in rural areas, so any link failure can lead to disruption. If the fault is not momentary and redundant unit is not available, system restoration may be prolonged. Domain of disruption is another aspect of faults. Some faults like IP phone problems disrupt only one person operation, but some faults like E1 or optical fibre cables damages can threat total system reliability.

PROBLEM DEFINITION

In current research, 790 faults which happened in Alborz and west Tehran province was analysed for finding meaningful relation between faults and their probable causes. Weather condition, place of fault occurrence, operator expertise and network traffic was these possible causes. We categorized all fault into 19 clusters including software problems, hardware problems and link problems. This study was based on the idea that if considerable relations between a fault and possible causes exist, it is possible to reduce occurrence rate of this fault by controlling its causes.

RESULT AND DISCUSSION

It was not possible to insert entire result in this paper so we showed important part of result in diagrams and charts. In 2012, emergency service center, which includes one central contact-center, two local call-centers and 19 agencies placed in urban and rural areas, encountered 790 minor and moderate faults. Although none of these fault have not lead to severe condition, we decided to analyze them for determining the relation between faults and their sources and evaluating the system vulnerability to them. The study showed that about 70% of fault affected agencies operation and only 30% of them had an impact on call-centers or contact centers (Figure 1).In addition, majority of faults were minor which can be overcome by replacing defected

components or eliminating software bugs.

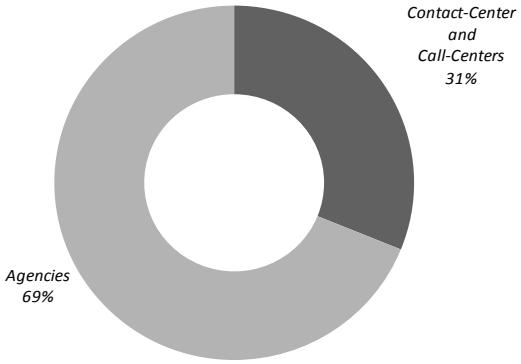

Figure 1: Fault percentage according to location

Contact-center and call-center related faults happened in all three centers, but the fault rate in contact-center was more than call-centers (Figure 2), because contact-center has more servers and more operators work in it. In addition, IP phones used in contact-center are rather old. The fault rate in west call-center was less than east one because west part of province is less populated than east part and the calls answered in west call-center were less than other centers.

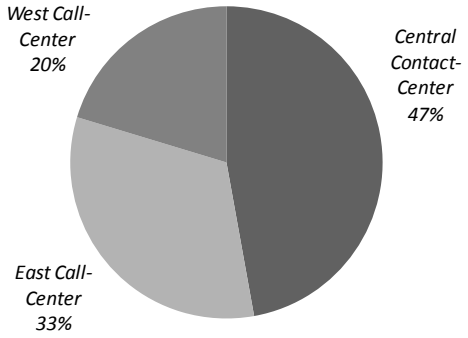

Figure 2: Fault of contact-center and call-center

There is meaningful difference between different agency faults rate too. Although the most vulnerable agencies are ones located in mountainous area, agencies which are in urban locality are not immune to faults, because IP phones and computers can be damaged after long-term usage. The place of agency has two effects on its fault rate. Firstly, sending expert technician to isolated agencies in mountainous region is time consuming, therefore preventive maintenances in these areas are less practical than urban ones. Secondly, the weather and distance have effects on radio communication (Figure 3).
The faults categorized into three groups (Figure 4):

1- Server problems including E1 ports, central servers, peripheral servers and their softwares.

2- Agent problems which are related to Operators' errors and IP phones and computers used by operators.

3- LAN and WAN problems

Details data of these groups were shown in figures 5 to 7.

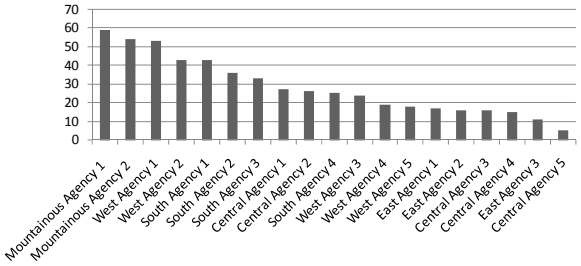

Figure 3: Faults of Agencies

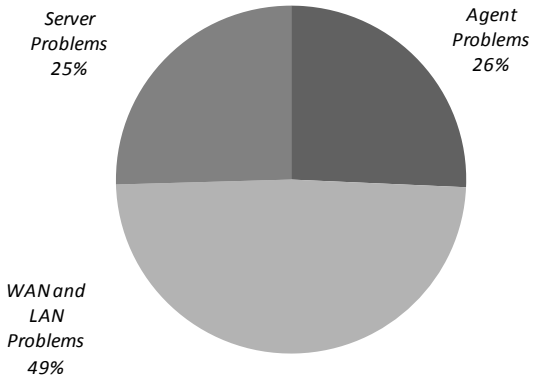

Figure 4: Fault percentage according to type

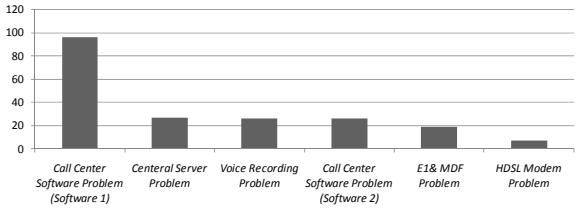

Figure 5: Servers faults

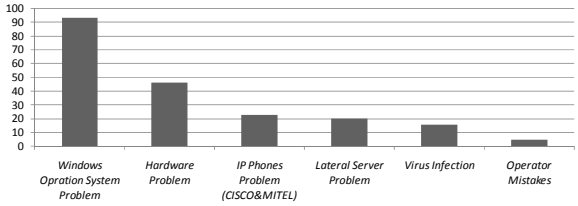

Figure 6: Agent faults

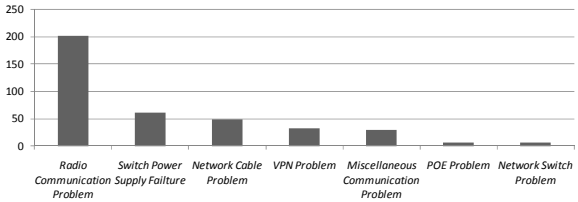

Figure 7: WAN and LAN faults

Figure 8 showed part of dendrogram which was used for fault clustering and table 1 represented some of relations detected by association rule generation. These rules were extracted from 400 faults.

Table 1: association rules extracted from first

Antecedent	Consequent	Support %	Confidence %	Rule Support %
Miscellaneous communication problem	Mountainous agency 1	5.88	91.67	5.39
Wireless network Problem	South agency 1	24.27	14.14	3.43
Miscellaneous software problem	Central contact center	9.31	34.21	3.19
Central contact center	E1 and MDF Problem	15.69	17.19	2.70
Mountainous agency 2	Wireless network problem	7.84	34.38	2.70
Wireless network problem	Mountainous agency 2	24.27	11.11	2.70
Wireless network problem	South agency 3	24.27	10.10	2.45

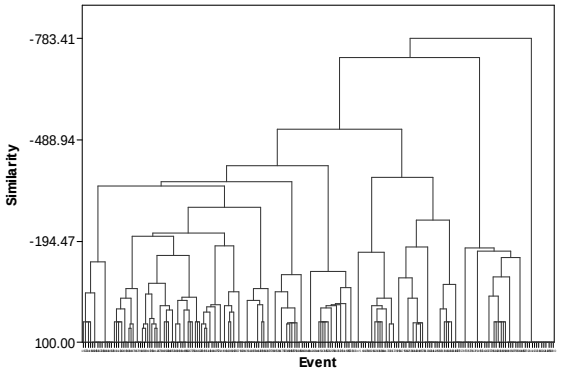

Figure 8: Part of faults dendrogram

The effect of each fault on system accessibility is not always equal to other faults. For example, IP phone problems only affect one operator answering and call queue is distributed to other operators, but if radio communication link between central point and one of distant agency is failed, the risk of emergency services disruption is significant. In addition, some faults can be eliminated after few minutes, but software debugging usually need days. Figure 9 represented the effect of each type of fault on system accessibility.

After analyzing extracted information and field study, many subjects were revealed:

1- Some faults like call-center software errors and voice recording software errors were interconnected.

2- Call-center softwares were the main source of faults in servers because these softwares were developed gradually for a few local distribution companies in Iran and were not bought from renowned international company. Therefore, there were many bugs in softwares which were found and eliminated.

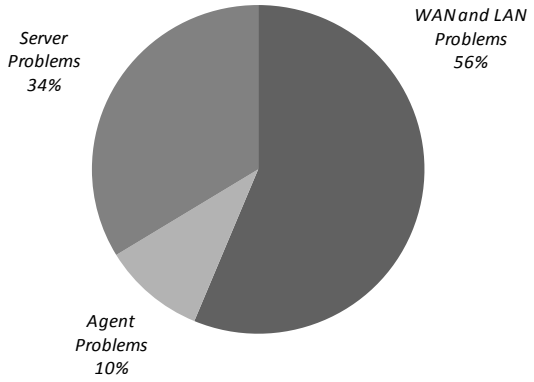

Figure 9: Effect of faults on system accessibility

3- After field study, it was revealed that some of system operators misused their computers. Installing games, changing software setting without permission and using system without care were major threats which reduced components life and increase fault rate.

4- Field study showed that although preventive maintenance plans have been being used for medium voltage and low voltage distribution network reliability enhancement, this concept was not applied for supporting facilities like communication networks.

5- Power supplies used for radio communication were not reliable enough and the defects in these components interrupted radio communication frequently.

6- When primary call-center was designed in previous years, servers and other components are selected for basic usage, but after converting call-center to contact-center, upgrading of these items were neglected.

CONCLUSION

In current project data mining tools were used for evaluating threat which can disrupt emergency system operation in Alborz province distribution company. Although its result is only valid for this case study, this approach is useful for other purposes including electricity distribution network fault analysis and customer clustering.

REFERENCES

[1]: Trevor Hastie, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, 2008

[2]: Oded Maimon and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, 2010

[3]: Valente de Oliveira J. and Pedrycz W., *Advances in Fuzzy Clustering and its Applications*, John Wiley & Sons, 2007