Understanding causes of low voltage (LV) faults in electricity distribution network using association rule mining and text clustering

Silva, HCE and Saraee, MH

http://dx.doi.org/10.1109/EEEIC.2019.8783949

Understanding Causes of Low Voltage (LV) Faults in Electricity Distribution Network using Association Rule Mining and Text Clustering

Charith Silva
School of Computing, Science and Engineering
University of Salford-Manchester
Greater Manchester, M5 4WT, UK
h.c.e.silva@edu.salford.ac.uk

Mohamad Saraee
School of Computing, Science and Engineering
University of Salford-Manchester
Greater Manchester, M5 4WT, UK
m.saraee@salford.ac.uk

***Abstract*— In-depth understanding of a fault cause in electricity distribution network has always been of paramount importance to Distributed Network Operators (DNO) for a reliable power supply. Faults in the network have direct effect on its stability, availability and maintenance; and so, their quick elimination, prevention and avoidance of fault causes that generated them, is of special interest. Possible opportunity to understanding the causes and correlation of the factors where future faults may arise can significantly help electricity distribution operators who happen to be accountable to detect and repair such problems. Every asset in the distribution network has a different level of reliability and which may vary. Faults identifying in distribution network have rich literature but a very few studies had been done on understanding the factors that contribute to LV Faults using data mining and machine learning techniques. As there are lack of studies on Faults identifying in distribution network with data mining, this study will formulate a starting point. This paper aims to use the association rule mining and clustering techniques to understand the various hidden patterns from the faults database. The uncovered relationships can be represented in the form of Association rules and clusters. The outcomes of this research will hugely beneficial to the engineering departments in DNOs. New knowledge gain from this study will help to priorities investments in new or replacement infrastructure which will ensure that financial and manpower resources are used more efficiently.**

***Keywords*- DNO, Electricity Distribution Network, LV Faults, Data Mining, Association rule, Clustering**

I. INTRODUCTION

Electric power distribution is the delivery system of electricity to places that use it, such as homes and other buildings. The electricity generates from the power station at high voltage and is delivered at medium to low voltage levels. Any Electric power distribution system can be simplified into three main stages, power generation, electricity transmission and distribution. In the UK electricity transmission is carried out by a single company, the National Grid. This monopoly is regulated by Office of Gas and Electricity Markets [1]. Within regions, the distribution of energy is carried out by DNOs [2]. There are 14 of these DNOs in the Great Britain, each of which has a license covering a defined geographical area. They are accountable to the industry regulator, OFGEM. DNOs are responsible for carrying electricity from the high voltage transmission network to industrial, commercial and domestic users and for carrying the power generated directly onto their networks. There are 14 DNOs operating in Great Britain managed by six companies [2].

Power systems are prone to frequent faults [3], which may occur in any of power generating units, transformers, power distribution media such as overheard and underground cables. Specially electricity distribution network components are always vulnerable to frequent failures that may occur in any of the main components or sub components. Faults that generally occur in transmission and distribution networks are short circuit transients caused predominantly by vegetation, animal and weather effects such as tree contact, large birds short circuiting phases, creepage current through path created by rain or moisture and the buildup of contaminants [4]. Weather is the single most influential factor that causes faults on the DNO network. Different weather parameters such as wind, temperature, snow and rainfall all have the potential to cause faults to different types of assets. The National Fault and Interruption Reporting Scheme [6], set up and administered by the Energy Networks Association. Each DNO in Great Britain is required to report all faults which occur on their network, whether the fault results in loss of supply to customers.

Any component malfunction in the power system, causes significant disturbance to the supply or destabilizing the entire system. Detecting faults in electrical power grids is of paramount importance, both from the electricity operator and consumer point of view. Regarding the fault events, the customer satisfaction survey is the key component that determines the quality of customer service delivered by the DNO against a specific fault incident. Following an unplanned fault, the DNOs submit data to OFGEM's independent customer survey organization [5]. The customer survey organization contacts the relevant customers and ask a series of questions associated with the customer's experience during the interruption. The customer scores the DNO out of ten on ease of contact, politeness, accuracy of information and usefulness of information. So, any unplanned outage can harm the DNOs reputation and might receive penalties from the regulators.

The financial penalties during fault outages can be significant and understandably this has, where possible, driven DNOs to avoid interruptions and restore customers quicker when a fault occurs. Initiatives to avoid interruptions can be targeted asset investment where apparatus has reached the end of its life; therefore, removing plant before failure occurs. Reducing interruption times via the introduction of new technology that

improves fault discrimination, fault sectioning and fault re-closure have been instrumental to improving the DNO's interruption performance. Therefore, there is a business need to accurately understand the faults and causes of the faults. In recent years, few researchers have proposed methodologies for fault analysis purely focusing on electric current flow. But this research is trying to introduce a new analysis methodology using data mining techniques. Also in this research, author is trying to establish relationship between environmental features and fault causes.

II. PROBLEM STATEMENT

Both from the DNO and consumer point of view, in-depth understanding of fault cause in electricity distribution network is paramount important. With this new proposed data mining model will help to improve the level of system availability by identifying and reducing the network faults. Also, there is a business need to reduce operational expenditure in engineering departments. So, this model may help to prevent faults before they happens.

III. OBJECTIVE OF THE STUDY

The primary objectives of the study are:

i. To understanding causes of LV Faults in Electricity Distribution Network using Association Rule Mining.
ii. Explore the possibility of enhancing the knowledge gain from Association Rule Mining using Text Clustering.

IV. INDUSTRIAL BENEFITS OF THE STUDY

i. Prioritise investments in new or replacement infrastructure which will ensure that financial and manpower resources are used more efficiently.
ii. Improved level of system availability by reducing the network faults.
iii. Reduction in number of complaints due to less network faults and avoidance of costs and potential fines associated with future network faults from the regulators.

V. RELATED WORK

Zhanjun et al. [16] presents a new method of the distribution network fault diagnose based on data mining. The method synthetical analysises the spatial and temporal characteristics of fault information produced in the distribution network It uses APRIORI algorithm to mine the association rules of fault information and establish strong association rules database of fault attributes. Then the distribution network fault is diagnosed by using the strong association rules database. But this method has some limitation when it applies to large real database like NAFIRS.

VI. NATIONAL FAULT AND INTERRUPTION REPORTING SCHEME

United Kingdom electricity companies, which are private organizations, collect data on faults in the NaFIRS database as part of their regulation criteria set out by the government [2]. This database contains details of all the HV and LV related faults on the electrical distribution system, including date, time, and number of consumers affected, number of minutes lost among others [7]. This Scheme was initially approved by the twenty-seventh Chief Engineers' Conference, held on 14th October 1964, and was subsequently revised several times [6]. NaFIRS is designed to collect information relating to both network performance and equipment performance [1].

According to Ford [6] the main objectives of the Scheme are:

(a) obtain and disseminate information relating to the reliability in service of distribution system equipment;

(b) provide information to permit the study of total distribution system performance, particularly. at times when they fail;

(c) provide information to permit the study of servicing organisations responsible for the operation, control, repair and maintenance of distribution systems and their component parts;

This supply interruptions on distribution network's information is shared nationally and summaries are submitted to Ofgem. Data is available for over thirty years but the quality of the data has improved significantly over the last fifteen years since the introduction of the Ofgem Interruptions Incentive Scheme (IIS). For each interruption, DNO's will capture a large amount of information and up to 100 separate fields will be populated. Using data from the NaFIRS system DNO's can monitor how their networks are performing, identify any trends in faults and respond accordingly.

Since 2010, companies have provided the full dataset to Ofgem who perform their own analysis. Although the data is aggregated at this level, companies actually capture data to a more detailed level, attributing faults to one of 99 different direct causes specified in ENA Engineering Recommendation G43-3 (Instructions for Reporting to the National Fault and Interruption Reporting Scheme). Eleven of these are weather related [8] e.g. : lightning, rain, snow, sleet, blizzard, ice, etc..

Each financial year OFGEM gives all DNO's a CI (Customers Interrupted) & CML (Customer Minutes Lost) budget, this money is paid upfront and whatever is left over they get to keep and re-invest in our network. They have to find a balance though, because too good of a performance results in next year's budget being lower and to poor of a performance means we have to pay the money back [2].

VII. METHODOLOGY

This section outlines the research methodology and approch taken for this study. As outlined by Saunders et al [9] the purposes of a research could be categorised as exploratory, descriptive and explanatory. An exploratory study can be described as valuable means of finding out what is happening to seek new insights[9]. It can be particularly useful in helping to understand a problem, clarify the nature of a problem or define the problems involved. It also enables you to develop propositions and hypotheses for further research, to discover new insights or to reach a greater understanding of an issue. In this study, author is attempting understand most significant factors that contribute to LV faults using Association Rule Mining and explore the possibility of enhancing the knowledge

gain from Association Rule Mining using Term Clustering. Association Rule Mining and Text Clustering techniques were performed in order to achieve the objectives

This study will address one of the main challenge in the Electricity Distribution Industry which is faults in the network. The outcomes of this research will support to the policy formulation engineering departments.

A. Data Mining

Data mining is a process of searching for unknown relationship or information on a large database or data warehouse using intelligence tools such as neural computing, predictive analysis techniques or advanced statistical methods [10]. Data mining uses mathematical and statistical algorithms to discover patterns, segment data and predict probabilities. Unsupervised learning methods such as Association Rules Mining belong to the Knowledge Discovery and Data Mining process (KDD) because they give us the opportunity to discover unknown or hidden patterns and relationships in databases.

B. Association Rule Mining

Association rule is an unsupervised learning technique to discover association of items. Association rule mining has been mainly applied for analysing customer's shopping baskets. This helps retailers identify items that are likely to be bought at the same instance. Association rule identifies combination of items purchase that frequently occur together [11]. Most popular technique for discovering association rules is the Apriori algorithm. Apriori algorithm is used for discovering association rule and finding frequent item set [10]. The Apriori algorithm is credited to Agrawal, Imieliński and Swami who applied it to market basket data to generate association rules [11].

VIII. EXPERIMENTAL EVALUATION

A. Data

Due to the nature of the comercial sensitivity of the data. The datasets contain synthetic data, based on real NAFIRS database. The synthetic datasets have only been developed for research purposes. Produced synthetic data which mimics the real data and preserves the relationships between variables. These data will allow data mining to be carried out more easily and we are confident that the synthetic data will be good enough to produce analysis results very close to those that would be carried out on the real data. The attributes of the data are mix of numerical and categorical in nature. Data transformation has been applied to some attributes to simplify the analysis. Similarly, other attributes have been transformed into appropriate form for better analysis of the data. Table 1: shows the attributes explanation of the synthetic dataset.

TABLE 1: ATTRIBUTES EXPLANATION OF THE DATA SET'S CONTENTS

ATTRIBUTE	TYPE	DESCRIPTION
HOUR	Factor	Hour of fault occurred
WEEKDAY	Factor	Weekday of fault occurred
MONTH	Factor	Month of fault occurred
CAUSE	Factor	Direct Cause
EQUIPMENT	Factor	Equipment Involved
COMPONENTS	Factor	Component Involved
CUSTOMERS	Factor	No. of Customers Interrupted
MINUTES_LOS	Factor	Customer Minutes Lost

Cause (Direct Cause) - Cause of the outage is mainly used to describe the responsibility and reason that cause outage, There are 101 direct cause identified by the NAFIRS.

- Lightning – Lightning striking one of our assets.
- Snow, Sleet and blizzard
- Vermin, Wild Animals and Insects
- Growing Trees – Trees growing through the lines
- Metal Theft

Equipment - (Main Equipment Involved). There are 40.
This is what part of the network has been affected by the fault and is broken down into 4 categories, Overhead, Underground Main, Underground Service & Switchgear/Fusegear/Link-box/Cut-out.

- Overhead Main Insulated Conductors
- Overhead Service (Metered) Insulated Conductors
- Underground Main Districable
- Switchgear/Fusegear
- Un-Metered Service Underground

Component - The next section is Component and is directly linked to the MEI. It describes the equipment that has faulted. Dependant upon your MEI depends upon the outcome of your component as there are multiple options for the same input.

- Conductor – the cable
- Insulator – this part of the circuit
- Jointed Termination Compression
- Heat Shrink termination Pole Mounted
- Main Contacts – LV board Jaws

Customer – No of customers effected

- No_customer_involved
- Only_one_customer_involved
- Between_2_and_10
- Between_11_and_25
- Between_26_and_50
- Between_51_and_100
- Between_101_and_250
- Between_251_and_500
- Between_501_and_1000

Minutes Lost

- No_minutes_lost
- Between_1_and_15
- Between_16_and_30
- Between_31_and_60
- Between_61_and_120
- Between_121_and_360
- Between_361_and_720
- Between_721_and_1440
- More_than_1_day

B. Data Cleansing & Quality Assurance

Key criterion for the success of the data mining or machine learning project is the cleanliness and cohesiveness of the data. Real world data are generally incomplete, inconsistent and noisy. Therefore, data cleansing is paramount to the data mining project, in this stage work can be done on missing values, smooth noisy data, identify outliers and resolve inconsistencies. Machine learning algorithms can be used to handle missing data. But as we use synthetic dataset, this step is not applicable. But in real fault dataset can have missing and inaccurate data.

C. Predictive modelling using Association Rules

i. Apriori algorithm

The Apriori is the best-known algorithm to mine association rules. Apriori algorithm is used to discover association rules. Given a set of retail transactions, the algorithm attempts to find subsets that are common to at least a minimum number of the item set. Apriori uses an iterative approach known as a level-wise search where k-itemsets are used to explore (k + 1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to collect the count for each item and accumulating those items that satisfy minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found [15]

Two important measures to evaluate the rule are Support and confidence. Support denotes the probability the set of items "(A => B)" how frequently occurs in total transaction. Confidence measure indicates how often the generated rule is found to be true. Users set a minimum support threshold and minimum confidence threshold and check if the rule satisfies both minimum thresholds.

Support (A ⇒B) = P (A∪B)

Confidence (A=>B) = p (B/A) = support (A∪B)/ support (A)

The key importance of the rules is "if an itemset is large if and only if all the subsets also large".

Lift defines the strength of association between the Left-hand side rule and right-hand side rule. If lift results larger the value then the rule is stronger . Lift ratio is calculated as below-

Lift (A=>B) = support (A∪B)/ support (A) * support (B).

ii. Results and discussion

The synthetic dataset consists of more than 50 000 faults each described by 8 attributes. First experiment would be find out all the associated factors that contributes to the no minutes lost faults. The first set of rules were obtained for minutes_lost=no_minutes_lost in RHS with the support set to 0.01 and confidence set to 0.8. In total, these settings generated 50 rules.

	2 ITEMS	3 ITEMS	4 ITEMS	5 ITEMS
NO OF RULES	1	30	17	2

As dataset contain over 50,000 faults and some faults are rare events, the minimum support threshold has to be reduced to identify less common faults. So, second set of rules were obtained with the support set to 0.001 and confidence set to 0.8. In total, these settings generated 972 rules. As Table 1 shows, the number and complexity of rules increase enormously.

	2 ITEMS	3 ITEMS	4 ITEMS	5 ITEMS	6 ITEMS
NO OF RULES	1	84	484	348	55

Table 3 show the number of rules generated with different confidence thresholds.

A scatter plot can be used to visualise the generated association rules. Support and Confidence can be used as X and Y axes. In addition, third measure Lift is used by color (grey levels) of the points.

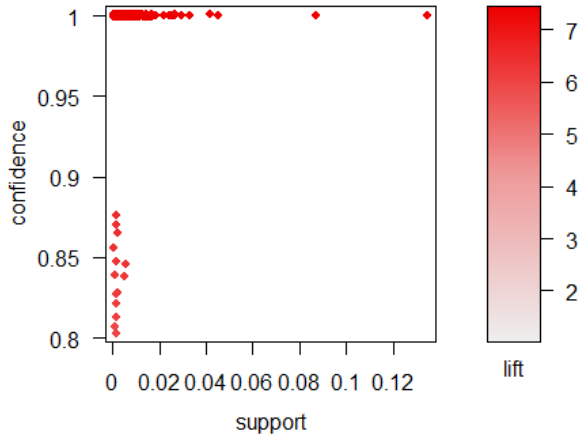

FIGURE 1: VISUALIZATION OF ASSOCIATION RULES ON SCATTRE PLOT

The above plot shows that rules with high lift have low support, but there 972 rules on the scatterplot distort the graph and the interpretation. Let's increase the Confident to reduce the number of rules. With a minimum confidence of 90%, the algorithms produce 958 rules. So, number of rules can not be reduce by increasing confident level(Table 1).

CONFIDENT	2 ITEMS	3 ITEMS	4 ITEMS	5 ITEMS
95%	1	80	477	345
90%	1	80	477	345
85%	1	81	479	346
80%	1	84	484	348
75%	1	88	487	348
70%	2	97	494	348

As each rule has to be analysed and eventually confirmed by an industry expert; So, clearly there are too many rules to be considered. An attempts to reduce the number of rules have failed, we have therefore proposed a new way by combining association rules mining and text clustering. Before explaining this new approach, we present variable clustering methods in the next section.

D. Enhance the results using Text Clustering

Text Clustering

Document clustering is considered as a fundamental operation utilized in the automatic topic extraction, organization of unsupervised document with information retrieval. Text clustering is an unsupervised process forming its basis solely on finding the similarity relationship between documents with the output as a set of clusters [12]. Frequent Term Based Text Clustering proposed by Beil et al. [13] is desired to solve the problems of applying conventional clustering methods on text datasets, such as not suitable for high dimensionality and large size of database, not be able to give cluster descriptions. The concept of the frequent term set is based on the frequent item

set of the transaction data set [14]. Figure 2 shows steps to enhance the results using Text Clustering

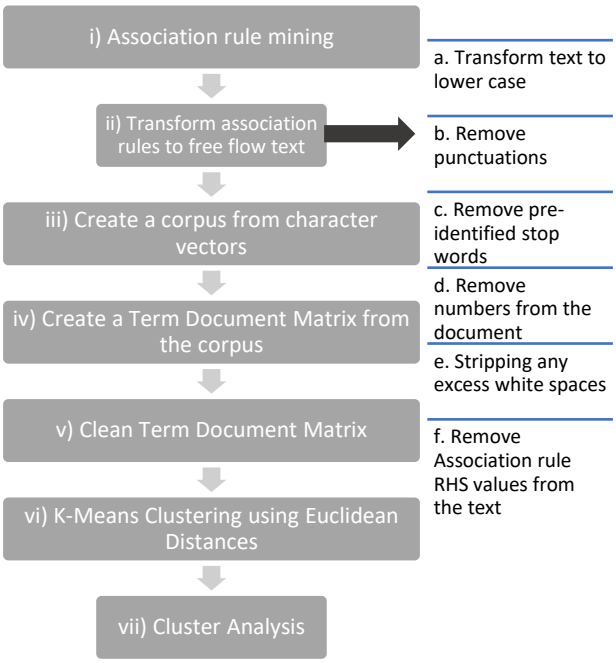

FIGURE 2: STEPS TO ENHANCE THE RESULTS USING TEXT CLUSTERING

i) Association rule mining

Previously generated same 972 rules will be used with text clustering. This set of rules were obtained with the support set to 0.001 and confidence set to 0.8.

ii) Transform association rules to free flow text

The cleaning process of the dataset was carried out through various steps. Text document has a collection of sentences, this step divides the whole statement into words by removing spaces, commas, numbers etc. Usually, stemming also part of this section. Stemming is the process of converting the word to their stem. As, in this method author tries to isolate association rules, stemming has been ignored from the text transformation process.

a. Transform text to lower case
b. Remove punctuations
c. Remove pre-identified stop words
d. Remove numbers from the document
e. Stripping any excess white spaces
f. Remove Association rule RHS values from the text eg: minutes_lost=no_minutes_lost

iii) Create a corpus from character vectors

A corpus is a collection of texts used for linguistic analyses, usually stored in an electronic database so that the data can be accessed easily. Corpus can be created from various available sources, in this study corpus has been created using character vector consisting of one document (One association rule) per element. There are 972 documents in the corpus.

iv) Create a Term Document Matrix from the corpus

A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. Create a term-document matrix from a corpus is a important part of the text mining process.

v) Clean Term Document Matrix

Go through each row of the matrix and determine if all the value are zero , if so remove the row from the matrix.

vi) K-Means Clustering using Euclidean Distances.

K-Means clustering algorithm values of k is given and k-means algorithm is implemented in 4 steps:

a) Partition objects into k nonempty subsets.
b) Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
c) Assign each object to the cluster with the nearest seed point. Go back to Step 2, stop when no more new assignment.

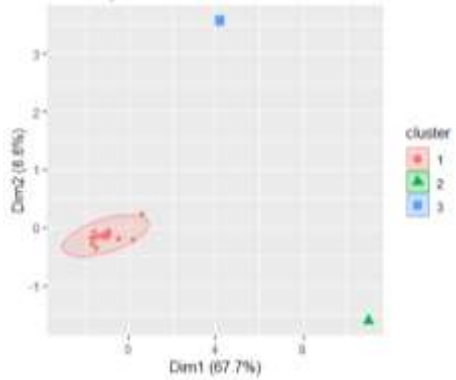

FIGURE 3: CLUSTER PLOT

Given a fixed number of k clusters, assign observations to those clusters so that the means across clusters are as different from each other as possible. In this study k=3 has been used.

1-cause=by_private_developers 1-cause=corrosion
1-cause=deterioration_due_to_ageing
1-equipment=underground_main_plcs
1-equipment=underground_service_(metered)
1-hour=10 1-hour=11
1-hour=14 1-weekday=friday
1-weekday=monday 1-weekday=thursday
1-weekday=tuesday 1-weekday=wednesday
2-customers=no_customer_involved
3-components= underground_cable__other_than_joints &_terminations

vii) Cluster Analysis

Cluster analysis is an exploratory analysis that tries to identify structures within the data. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are meaningful. Text clustering has been used to group similar documents and make meaningful groups. In this study, text clustering has been used to group factors that contain in the assassination rules.
According to the Association Rule mining and text cluster analysis, most of the none_minutes lost LV faults happened due to underground cable other than joints &_terminations and those faults had not impact any customers.

viii) Results validation
Set of rules were obtained for minutes_lost= more_than_1_day in RHS with the support set to 0.001 and confidence set to 0.8. In total, these settings generated 1384 rules. As Table 1 shows, the number and complexity of rules increase enormously.

	2 ITEMS	**3 ITEMS**	**4 ITEMS**	**5 ITEMS**	**6 ITEMS**
NO OF RULES	3	161	715	424	81

According to the Association Rule mining and text cluster analysis, most of the LV faults that caused more than I day outage affected more than 50 customers but less tha 100 customers. Cluster plot looks like below.

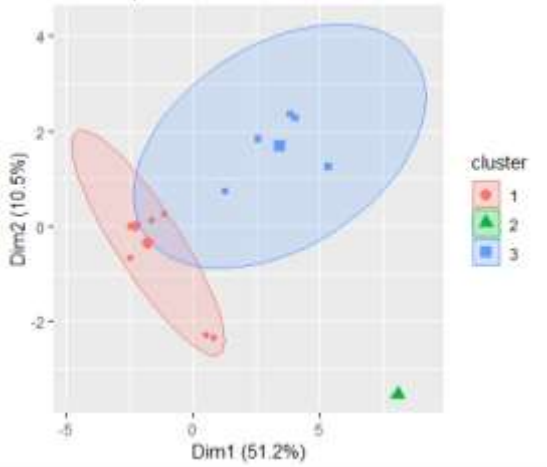

FIGURE 4: CLUSTER PLOT

IX. CONCLUSION

There is a business need to reduce operational expenditure in engineering departments. The intention of the research study was to build a model which can analyse and understand the potential causes for LV network faults, which will help the DNOs to save time and money as well as it also helps to prioritise investments in new or replacement infrastructure which will ensure that financial and manpower resources are used more efficiently. This research has proposed a new method that analyse the historical fault data and trying to understand the behaviour of the fault with other factors. This method may allow saving power distribution system equipment which can be destroyed by an upcoming major fault. Still there are many areas where the data mining can be improved using advanced algorithms.

REFERENCES

[1] Medium term network performance monitoring (Rep.). (2000, November). Retrieved January 16, 2019, from OFGEM website: https://www.ofgem.gov.uk/ofgem-publications/79303/report-medium-term-network-performance-07-12pdf

[2] RIIO-ED1 Annual Report 2016-17 (Rep.). (2017, December 19). Retrieved March 10, 2019, from OFGEM website: https://www.ofgem.gov.uk/system/files/docs/2017/12/riio-ed1_annual_report_2016-17.pdf

[3] Filomena, A. D., Resener, M., Salim, R. H., & Bretas, A. S. (2011). Distribution systems fault analysis considering fault resistance estimation. International Journal of Electrical Power & Energy Systems, 33(7), 1326-1335. doi:10.1016/j.ijepes.2011.06.010

[4] Lout, K. (2015). Development of a fault location method based on fault induced transients in distribution networks with wind farm connections (Unpublished master's thesis). University of Bath.

[5] Climate Change Adaptation Report (Rep.). (2015, July). Retrieved February 6, 2019, from Scottish and Southern Energy Power Distribution website: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/478927/clim-adrep-sse-power-distribution-2015.pdf

[6] Ford, D. (1972). The British Electricity Boards National Fault and Interruption Reporting Scheme-Objectives, Development and Operating Experience. IEEE Transactions on Power Apparatus and Systems, PAS-91(5), 2179-2188. doi:10.1109/tpas.1972.293201

[7] Dunn, S., Wilkinson, S., Alderson, D., Fowler, H., & Galasso, C. (2018). Fragility Curves for Assessing the Resilience of Electricity Networks Constructed from an Extensive Fault Database. Natural Hazards Review, 19(1), 04017019. doi:10.1061/(asce)nh.1527-6996.0000267

[8] Climate Change Adaptation Reporting Power Second Round (Rep.). (2015). Retrieved December 14, 2019, from ENERGY NETWORKS ASSOCIATION

[9] Saunders, M. N., Lewis, P., & Thornhill, A. (2019). Research methods for business students. New York: Pearson.

[10] Turban, E., Delen, D., & Sharda, R. (2018). Business intelligence, analytics, and data science: A managerial perspective. Harlow ; Munich: Pearson Prentice Hall.

[11] Olson, d. L. (2019). Descriptive data mining. S.l.: springer verlag, singapor.

[12] Reddy, G., Rajinikanth, T., & Rao, A. (2014). A frequent term based text clustering approach using novel similarity measure. 2014 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2014.6779374

[13] Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 02. doi:10.1145/775047.775110

[14] He, Q., Li, T., Zhuang, F., & Shi, Z. (2010). Frequent term based peer-to-peer text clustering. 2010 Third International Symposium on Knowledge Acquisition and Modeling. doi:10.1109/kam.2010.5646177

[15] Han, Jiawei, and Micheline Kamber. Data Mining: Concepts and Techniques. Elsevier, 2012.

[16] Zhanjun, G., Zhengliang, P., Nuo, G., & Bin, C. (2014). A distribution network fault data analysis method based on association rule mining. 2014 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). doi:10.1109/appeec.2014.70