International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 24 No. 2 Jun. 2016, pp. 285-292
© 2015 Innovative Space of Scientific Research Journals
http://www.ijisr.issr-journals.org/

Cluster Analysis Based Fault Identification Data Mining Models for 3 Phase Power Systems

Tan Yong Sing[1], Syahrel Emran Bin Siraj[1], Raman Raguraman[1], Pratap Nair Marimuthu[1], K. Gowrishankar[1], and K. Nithiyananthan[2]

[1]Faculty of Engineering and Computer Technology, AIMST University,
Bedong, Kedah, Malaysia

[2]Department of Electrical and Electronics Engineering,
Karpagam College of Engineering, Coimbatore, Tamil Nadu, India

ABSTRACT: The main objective of this research work was to develop a Cluster Analysis based fault identification model for the power system. Cluster Analysis based Data Mining Techniques model has been implemented to locate the 3-phase transmission lines fault in IEEE 30 bus power system. Power World version 18 software was used to simulate the IEEE 30 bus power system and the 3-phase transmission lines fault. The bus voltages at fault were collected and import to the Statistical Package for the Social Sciences (SPSS) for determining the bus at fault. Through Cluster Analysis using Squared Euclidean Distance method, fault has been identified at each bus. This meant that the application of Data Mining Techniques yields a huge potential in solving complex problems related to power system, it not only yield an accurate result but also fast computation. The proposed innovative, successful model was able to locate the fault at each bus by bus nominal voltage comparison method.

KEYWORDS: Power system transmission lines faults, Data Mining, Cluster Analysis.

1 INTRODUCTION

Data Mining Technique had gain popularity recently due to its practicality and robustness in solving complicated problems, especially problems involving a huge set of data. Trends and knowledge that exceed simple analysis can be observed and learned through a database by applying Data Mining Techniques which also called as Knowledge Discovery in Data (KDD). Data Mining Techniques had successfully applied to various fields such as business, security, banking, anti-terrorist activities, social networks, medical field and information technology. There was a lot of data mining analysis involved in Data Mining Technique such as Classification, Evolution Analysis, Generalization, Cluster Analysis, Association Analysis, and Outlier Analysis. As stated above, there was a lot application of Data Mining Technique in various fields, however, there were less research on applying Data Mining Technique into Power System field. As the demand of electricity increases, it increases the complexity and the capacity of power system networks, which leads to the huge amount of data was collected every time. Transmission line faults in power system can be categorized as symmetrical fault and unsymmetrical fault. The symmetrical fault will equally affect all the three phases, however, just about 5% of chances it occurred. Upon comparison with asymmetrical fault, three phases were not affected equally and most of the fault occurred was asymmetric. Example of symmetrical fault was Three Phase fault. Asymmetrical fault will not affect all the three phases equally. There were various types of asymmetrical fault in transmission lines such as:

- Single line-to-ground (SLG) fault
- Line-to-line (LL) fault
- Double line-to-ground(DLG) fault

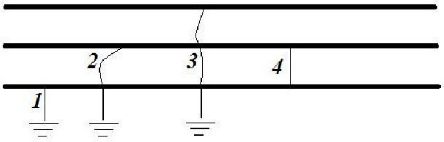


Fig 1: Illustration of transmission line fault

1) SLG Fault
2) DLG Fault
3) Three Phase Fault
4) LL Fault

There were many available techniques available in power system fault detection in transmission lines. Takagi et.al. [1] proposed a impedance-based fault location method by using both pre-fault and pure fault data and minimizing the influence of fault resistance. In 1931 J.Rohrig [2] reported using travelling waves for fault location in both overhead and underground transmission lines. Travelling waves were a technique where the position of the fault can be determined by measuring the time taken for high frequency electromagnetic waves to reach the end of the transmission lines. Later, Amir Tabatabaei et.al. [3] proposed by using Discrete Wavelet Transform in traveling waves to locate the fault of the three phase transmission lines and uses Global Positioning System(GPS) to obtain the actual data. However, there was difficulty in analyzing certain types of signals if the mother wavelet was not chosen properly and it's unable to distinguish between LL and SLG faults in power system due to an overlap in the range of the detail coefficient magnitude. More advance technique, which lead to non-conventional methods were developed to detect fault location in transmission lines fast and accurately. Aravinda Surya. V et.al [4] proposed using ANN to detect LG fault in double circuit transmission lines, the result showed that the maximum absolute error was less than 1% which means it yield an accurate result. Besides that there were more non-conventional techniques such as Fuzzy Logic technique or Genetic Algorithm technique Yagang Zhang et.al. [5] proposed using Data Mining cluster analysis to identify transmission line faults in IEEE 9 bus power system but have not tested in larger bus system, it was only tested for Single Line-To-Ground transmission lines fault, the author directly performs Cluster Analysis without pre process the raw data. The proposed model has been tested for larger bus system IEEE 30 bus system with data preprocessed features to detect 3 Phase fault and using Squared Euclidean Distance (SED) method for calculating the proximity matrix of similarity among each bus and which produce more accurate and distinct results.

2 CLUSTER ANALYSIS BASED DATA MINING

Cluster Analysis was a method of grouping objects such as observations or events found based on the information in the data describing the relationships between the objects. The product of Cluster Analysis will contain multiple groups of objects where each object in a same group will be closely related or similar and objects that were different from others will not be clustered into the same group. Distinct clusters of groups can be produced if there were large difference and great homogeneity in the group.

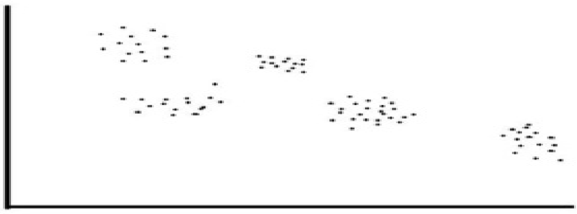


Fig 2: Illustration set of data

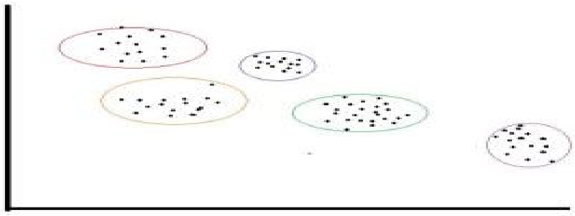


Fig 3: Illustration of 3 types of clusters in a set of data

As shown in Fig 2 there were raw data which carry no knowledge or no distinct relationships among other data. After Cluster Analysis was applied into the set of data, as shown in Fig 3 there were five different types of clusters or groups, thus from here crucial information can be extracted which simple observation unable to extract it. Cluster Analysis has long been used in a wide variety of fields such as pattern recognition, retrieval of information, psychology, biology, statistics, machine learning and data mining. Generally there were two types of clusters in Cluster Analysis which are Partitional (un nested) set of clusters and Hierarchical (nested) set of clusters [6]. Data objects are randomly divided into non-overlapping clusters such that all data objects were in exactly one subset, this is called Partitional clustering. Let K be the desired amount of clusters, Partitional clustering techniques will find all K cluster at once in conditions that K must less than the number of data set. As shown in Fig 2 and Fig 3 the illustrations of partitional clustering where Fig 2 was the original points and Fig 3 was the partitional clustering. Hierarchical clustering produces a nested sequence of partitions which contain only one, all inclusive clusters at the top of the hierarchy and finally form a singleton clusters of individual points at the lowest level of hierarchy as shown in Fig 4. The outputs of hierarchical techniques can be easily observed and analyzed by producing a dendrogram shown in Fig 5. Dendrogram able to represent the degree of difference between each element by analyzing the horizontal lines in the dendrogram, the element 4 in the Fig 5 had the longest horizontal lines which represent as outlier.

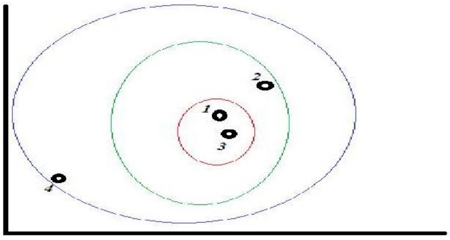


Fig 4: Illustration of Hierarchical clustering

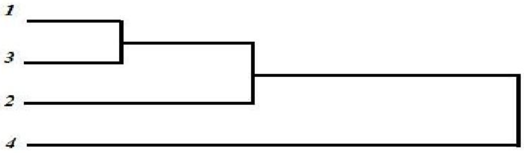


Fig 5: Illustration of dendrogram

Generally there were two available methods involve in Hierarchical clustering, which are divisive (Top-down approaches) and agglomerative (Bottom-up approaches). Divisive hierarchical clustering also called as the flat algorithm which the algorithm starts with all data set in a single cluster and then form hierarchy clusters by splitting the least similar clusters. Agglomerative hierarchical clustering is the reverse approach of divisive approach. Agglomerative algorithm starts with individual data, data with highest similarity between each other will form a cluster in a subsequent hierarchical stage [6]. Cluster Analysis with agglomerative algorithm is a powerful tool to locate transmission line faults, especially 3 Phase fault in power system, due to the, ability of Cluster Analysis able to identify the bus at fault fast and accurately. There were very few research has been done where this analysis was used in power system application which might produce a new powerful solution for the power system.

3 CLUSTER ANALYSIS BASED POWER SYSTEM DATA MINING MODEL FOR 3 PHASE FAULT DETECTION

As the complexity of the power system, increase the information regarding of the power system network also increases. Traditional method of transmission line fault identification has become more costly and ineffective as power system network keeps growing to sustain a healthy supply and demand relationship. At present, more powerful and effective methods are required to replace the traditional methods in detecting transmission line fault. Data mining technique was one of the windows of opportunity for effective transmission lines fault detection in power system, making benefits from the ever increasing raw data collected such as bus voltages, bus currents, transmission line impedances and many others. Upon all the collected data and information, new hidden knowledge regarding the behavior of the power system can discover and with this knowledge detecting transmission lines fault will become simpler, faster and more accurate. In this paper, the proposed agglomerative hierarchical clustering technique with Squared Euclidean Distance (SED) method and preprocessed bus phase voltage data to determine the bus in the IEEE30 bus system was at fault. The Cluster Analysis model for proposed methodology can be viewed in Fig 6.

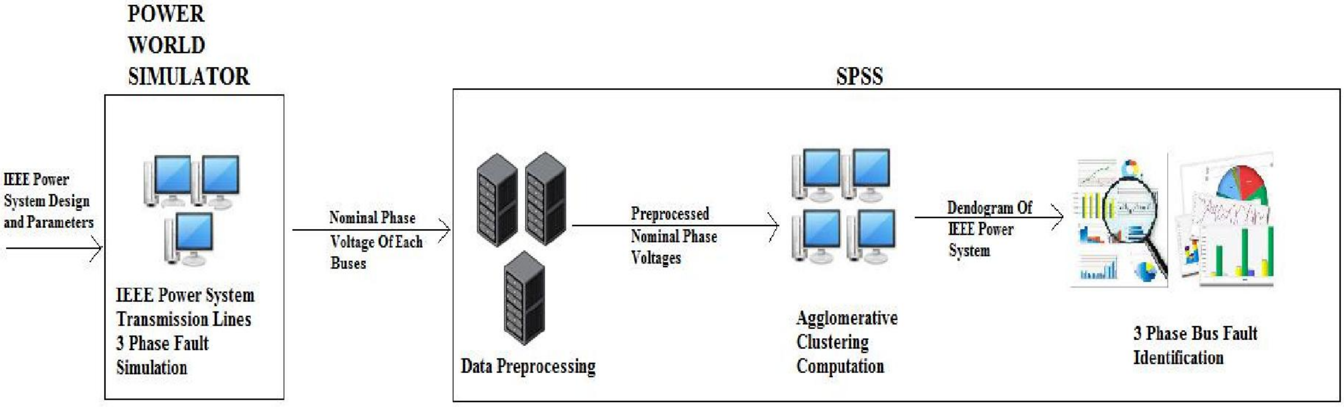
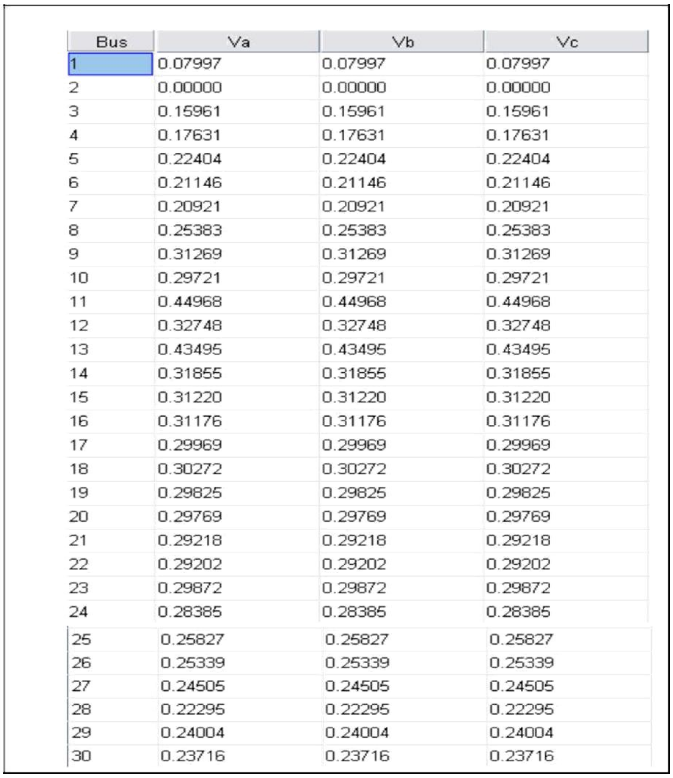


Fig 6: Cluster Based Fault Identification Model

The whole proposed design structure can be broken down into four main subdivisions which will be discussed in the following sections.

3.1.1 PHASE FAULT SIMULATION IN POWER WORLD SIMULATOR

After collections bus system data of the IEEE 30 power system, the IEEE 30 bus system was designed and the 3 phase transmission lines fault was simulated by Power World Simulator v.18. The one-line diagrams of each individual bus was first located in the new design window shown in Fig 7 with maximum (1.05 p.u.) and minimum (0.95 p.u.) voltage magnitude were assigned to each bus. Then each bus was connected to the transmission line. After all the transmission lines had been configured the generator and load were connected. The phase voltages of each bus produced by the simulation as shown in Fig 8 was collected, and save in spreadsheet format for data preprocessing.



Bus	Va	Vb	Vc
1	0.07997	0.07997	0.07997
2	0.00000	0.00000	0.00000
3	0.15961	0.15961	0.15961
4	0.17631	0.17631	0.17631
5	0.22404	0.22404	0.22404
6	0.21146	0.21146	0.21146
7	0.20921	0.20921	0.20921
8	0.25383	0.25383	0.25383
9	0.31269	0.31269	0.31269
10	0.29721	0.29721	0.29721
11	0.44968	0.44968	0.44968
12	0.32748	0.32748	0.32748
13	0.43495	0.43495	0.43495
14	0.31855	0.31855	0.31855
15	0.31220	0.31220	0.31220
16	0.31176	0.31176	0.31176
17	0.29969	0.29969	0.29969
18	0.30272	0.30272	0.30272
19	0.29825	0.29825	0.29825
20	0.29769	0.29769	0.29769
21	0.29218	0.29218	0.29218
22	0.29202	0.29202	0.29202
23	0.29872	0.29872	0.29872
24	0.28385	0.28385	0.28385
25	0.25827	0.25827	0.25827
26	0.25339	0.25339	0.25339
27	0.24505	0.24505	0.24505
28	0.22295	0.22295	0.22295
29	0.24004	0.24004	0.24004
30	0.23716	0.23716	0.23716

Fig 7: One-line representation of IEEE 30 Bus System Design

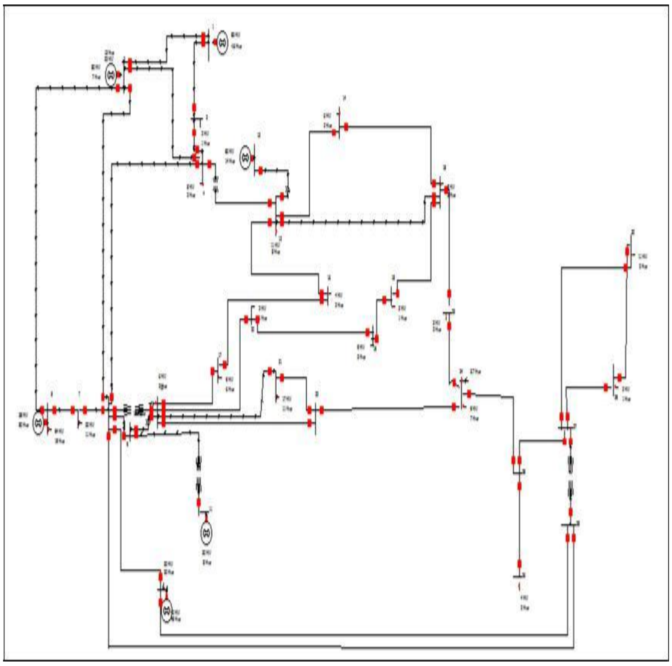
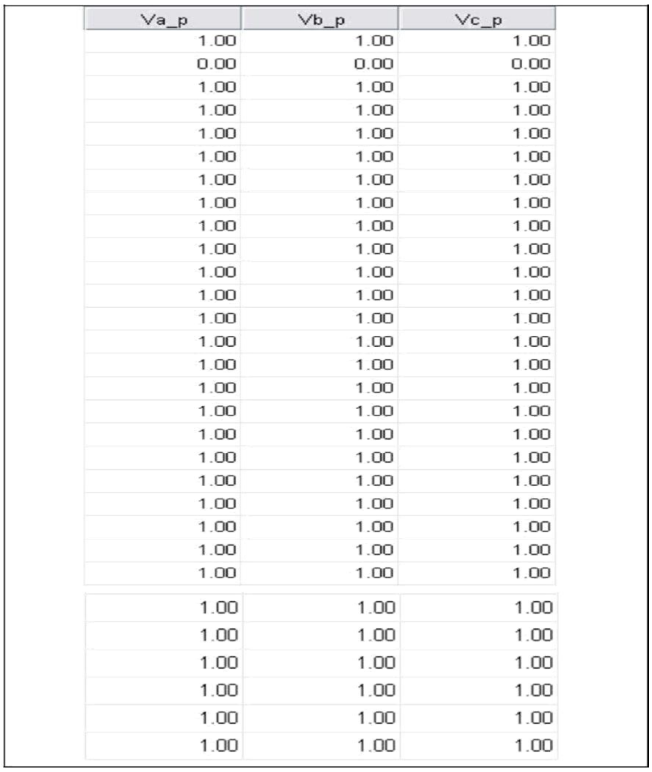


Fig 8: Simulated 3 Phase fault at bus 2 in IEEE 30 bus power system

3.2 DATA PREPROCESSING

The raw data (Va,Vb,Vc) collected needed to be preprocessed in order to obtain the best results in term of accuracy and speed in data mining technique. All the data were transformed into “0” or “1” values before Cluster Analysis was applied as shown in Fig 9. This was because the SED method for measuring the similarity of each bus worked best at values “0” and “1”.



Va_p	Vb_p	Vc_p
1.00	1.00	1.00
0.00	0.00	0.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00
1.00	1.00	1.00

Fig 9: Result after data preprocessed

As shown in Fig 9, all the above phase voltages become values “1” except for all the three phase voltages in the second bus which having the phase voltages “0”. Data preprocessed was essential as it not only making sense of raw data, but also leveraging the efficiency of data mining techniques. Raw data contain a lot of noises and disturbances which may affect the final output.

3.3 CLUSTERING OF POWER SYSTEM DATA

After the data had been preprocessed, it was ready to mine information using Agglomerative Clustering techniques to the preprocessed bus voltage data. The proximity matrix was generated using Squared Euclidean distance method can be calculated as follows:

$$Sq.Euclid(Bus_j, Bus_k) = \sum_{j=1}^{30}\sum_{k=1}^{30}\sum_{i=1}^{3}(V_{ij} - V_{ik})^2$$

where:
i represent the phase voltages at each bus
j and k represent the bus numbers

Hence, starting from each individual bus and then the bus, which has the highest similarity between each other based on the proximity matrix was cluster together. The bus, which having least similarity in the proximity matrix will be the last and the furthest to cluster into the final group. SED is the distance between two elements which was the sum of the squared differences between the values for each element. SED equation can generally be written as follows:

$$Sq.Euclid(x, y) = \sum_i (x_i - y_i)^2$$

3.4 SQUARED EUCLIDEAN DISTANCE BASED ALGORITHM IN 3 PHASE FAULT DETECTION[8]

The final step was to evaluate the dendrogram which result from Cluster Analysis with agglomerative algorithm and the result was compared with the fault simulated in the Power World Simulator. The algorithm to perform agglomerative clustering for 3 Phase fault identification in power system using SPSS was as follow:

1. *Compute the proximity matrix of the preprocessed phase voltages of each bus using the SED equation in (4).*
2. *Let each bus point forms an individual cluster.*
3. *Each cluster was compared and similarity of each cluster was computed*
4. *After similarity of each cluster was computed, merge two clusters which had the highest similarities to form a new cluster*
5. *The proximiy matrix was then updated as in step (1)*
6. *The algorithm iterates step (1) to step (5) until a single cluster was obtained.*

4 RESULTS

The proposed methodology architecture was entirely carried out in Windows 7 workstations with SPSS version 16.0 and Power World Simulator version 18. The result obtained from the proposed methodology was very promising as the Cluster Analysis with SED method and data preprocessing. The dendrogram of Cluster Analysis with Squared Euclidean distance method as shown in Fig 10 successfully identify the Bus 2 was at fault distinctly. Besides that, the proposed method was effective where there were a total of 3 only clusters for identifying the outlier bus. However, if for the cluster analysis with Squared Euclidean distance method without data preprocessing having 5 clusters for identifying the outlier bus. This show that suitable data preprocessing technique was required for power system data.

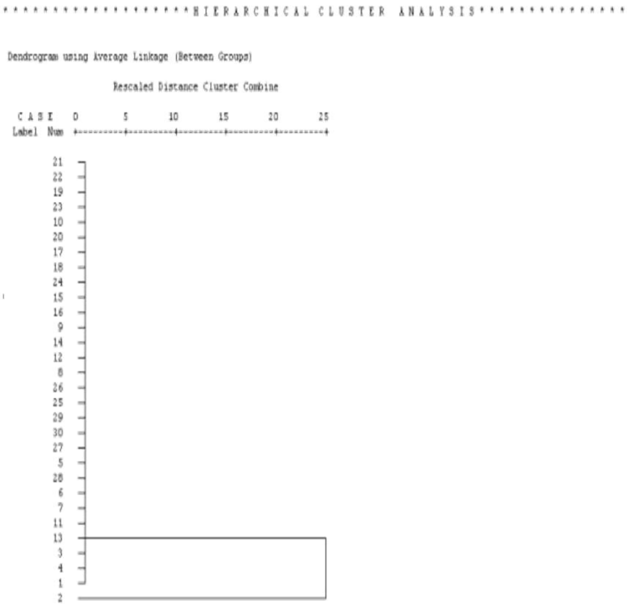


Fig 10: Dendrogram of Agglomerative clustering with SED

The scaled distance for Bus 2 for Fig 10 was at maximum scaled of 25 which means it's definitely an outlier bus which was abnormal from other buses. The ratio of the scale distance of each shown in the dendrogram was exactly same as the ratio of the actual distance between each bus. Thus, it can be concluded that with data preprocessing and SED clearly able to identify the Bus 2 at 3 Phase fault with the bus nominal voltages as the original data.

5 CONCLUSION

An effective Custer Analysis based Power System model has been developed to identify symmetrical transmission line's fault. The proposed method by using data mining technique such as Agglomerative clustering successfully identify the Bus 2 at 3 Phase fault in an IEEE 30 bus power system. This showed that there were huge potential by using DMT in power system, although very less research has been done. Therefore, in today's world of ICT there are a lot of information being presented in order to optimize the efficiency in solving complex problems in power system conventional methods will not perform as well as non-conventional methods. DMT was one of the non-conventional methods that are gaining popularity in analysis that involving large data in the database.

REFERENCES

[1] T.Takagi, Y. Yamakoshi, R. Kondow, M. Yamaura, T. Matsushima, *"Development of a New Type Fault Locator Using the One-Terminal Voltage and Current Data."* IEEE Power Eng Rev. PER-2(8), 59-60 (1982).

[2] J. Rohrig, "*Location of Faulty Places by Measuring with Cathode Ray Oscilloscope*". Electicitat Zeitschrift, 19th Feb, 1931, pp 241-242.

[3] Amir Tabatabaei, Mohammad Rez Mosavi, Abdolreza Rahmati, "*Fault Location Techniques in Power System based on Traveling Wave using Wavelet Analysis and GPS Timing",* Electrical Review, ISSN 0033-2097, 2012.

[4] Aravinda Surya. V, Ebha Koley, Anamika Yadav and A.S. Thoke, *"Artificial Neural Network Based Fault Locator for Single Line to Ground Fault in Double Circuit Transmission Line",* DOI: 10.7763/IPEDR, vol. 75, 2014.

[5] Yagang ZHANG, Jing MA, Jinfang ZHANG, Zengping WANG, *"Application of Data Mining Theory in Electrical Engineering"*, SciRes, DOI:10.4236/eng.2009.13025.

[6] Technical paper *"An Introduction to Cluster Analysis for Data Mining*", 2000

[7] Nithiyananthan K. and Ramachandran V.,'RMI Based Multi Area Power System Load Flow Monitoring', Iranian Journal of Electrical and Computer Engineering, Tehran, IRAN, Vol. 3, No.1, pp. 28-30.

[8] Tan Yong Sing, Syahrel, Emran bin Siraj, Raman Raguraman, Pratap Nair Marimuthu, K. Nithiyananthan,*'Local Outlier Factor Based Data Mining model for Three phase Transmission Lines Faults Identification*', International Journal of Computer Applications, USA, 2015, Vol130, No2.17-23.