IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 7, July 2016

Text Mining Methodology

Prajakta R. Pagar[1], Dr. M. U. Kharat[2]

PG Student, Computer Department, MET BKC Adgaon, Nashik, India[1]

HOD, Department of Computer Engineering, MET BKC Adgaon, Nashik, India[2]

Abstract: D-matrix is a standardized diagnostic model which is used to catch the fault system data and its causal relationship at the hierarchical system-level. Construction of D-Matrix by using the data source views the overall data by saving the entire database. Further it parses that data scanning which requires more memory and makes the process time consuming. Is describes construction and updation of D-Matrix by auto mining the unstructured repair verbatim (written in unstructured text) data collected during fault diagnosis using document pre-processing, term extractor and phrase merging. The system composes the fault diagnosis ontology comprising of dependencies and relationships currently saw in the fault diagnosis domain of interest and then text mining algorithms make utilizing of ontology concept to identify the necessary artifacts, like as failure modes, parts, symptoms and conditions from the unstructured repair verbatim content. The proposed scheme is equipment as a prototype tool and accepted by utilizing real time information gathered from the automobile domain of interest.

Keywords: Data Mining, fault analysis, fault diagnosis, information retrieval, text processing.

I. INTRODUCTION

A complex scheme interfaces with its neighbouring to executing some group of assignments by keeping up its achievement inside an respectable range of tolerances. Any deviation of a framework from its worthy execution was dealt with as a faulty component. The Fault Detection and Diagnosis (FDD) performance for distinguish the faulty component and diagnose the fault to root-causes of a scheme. FDD information comes in the unstructured repair verbatim data (also brought up as medicinal records in medical organization, aerospace industry, automotive) that gives affluent authority of diagnostic data. Its comprises of symptom relating to the faulty components, the saw disappointment modes, and repair moves made to exact faults. Unstructured repair verbatim data are added and is contend that there is a pressing urgency to mine this information to, enhance FD. The size of the repair verbatim information reduce an capability of its efficient performance within the method of FD. Text mining [3] is picking up a sincere consideration because of its capacity to automatic discover the ability assets covered in unstructured text. The strategy of text mining to tracing the diagnosis data departure from the unstructured repair verbatim data in a dependency matrix [4]. This model is the recognized analytic relief determined in IEEE Standard [5]. Fault detection process starts by extracting the bugs are present in the scheme and based on the recognized bugs the technicians use diagnosis procedure along with their knowledge to diagnose the faults. fault diagnosis, several information types such as bugs, scanning expenses of operating parameter associating with faulty scheme, repair verbatim are collected. This collected information transferred to the database and this data can be mined to construct dependency matrix (D-Matrix) models. To perform accurate FDD, Such models can be utilized by the technicians and other shareholders. The D-Matrix catches attributes and dependency of system between a one or more faulty mode using a one or more symptoms in a organized fashion. These dependencies among faulty modes (f1, f2, etc.) in parts (p1, p2, etc.) and symptoms (s1, s2, etc.) allow us to state a group of faulty modes causes symptoms. A probability of detection, the originate weights (d11, d12, etc.) is contained at the intersection of a row and a column indicates. In the binary dependency matrix, all the contingency have a expense of 0 or 1, where, 0 denotes no disclosure and 1 denotes complete disclosure of a exact failure mode utilizing a exact symptom. Generally, the D-matrix are constructed by utilizing the history information, engineering information, and sensory information, for sample, [6]. Even so, a practically nothing understanding is given about the disclosure of new symptoms furthermore, faulty condition saw first time and their incorporation in the dependency matrix model. In the methodology the occasional growth of the deficiency finding fault diagnosis ontology support the content mining calculation to develop the perfect D-matrix.

II. RELATED WORK

Dnyanesh G. Rajpathak et.al. have proposed to construct Ontology based text mining method comprising of dependencies and relationships currently saw in the fault diagnosis and then the text mining algorithms makes utilized of this ontology to identify the necessary artifact such as parts, symptoms, faulty modes to construct the D-matrix by auto mining the unstructured repair verbatim information collected during fault diagnosis episode[1].

M. Schuh et.al. have proposed about analyze of information from the on-board Diagnosis by utilizing the data mining process. Onboard diagnosis model collected

Copyright to IJARCCE

DOI 10.17148/IJARCCE.2016.5735

the real life information and integrates onboard ECUs. Onboard diagnosis model is assumed to be still and proper. But in actual world, due to engineering changes and design, newly vehicle structure and vehicle frame is launching. But in this approach to Faults are removed and provided ontology-guided data mining and data transformation, But Discovery is mislaying because result was not in form of matrix [7].

J. Sheppard, et.al. have proposed the only finite effort are done to determine a dependency matrix by evaluating unstructured repair verbatim diagnosis episodes. D-matrix constructed by using the data sources that views the overall data that is saving all database and firstly parse that data and then after scan overall data so it is takes more database memory and it is very much time consuming for scanning that overall databases [8].

S. Singh, et.al. have proposed data-driven structure for finding anomalies in failure information. The SME's regularly detecting anomalies by physically works and arranging the field faulty information utilizing spreadsheets is very much time consume and hard working flow. Therefore a data-driven architecture is construct which auto detect the unusual activity that leading to fault and saving a expressive expert's times [9].

D. Wang et.al. proposed Ontology-based fault diagnosis for transformers in which a new approaches to transforming failure diagnosis, which follows exchange of data using the Semantic Web. This ontology model collected the serious termed in the FD of power transformer, like as, faulty phenomena, sense of faults [10].

S. Singh et.al. proposed the system-level fault dependency matrices using different data and its sources. The D-matrices is categorize based on their data source and the another symptoms. It is a applicable for Boolean and real value [0, 1] D-matrices [11].

T. Felke proposed the fault diagnosis D-matrix models have been successfully utilized in service technician to detect the dependencies between failure conditions, symptoms, and repair action by analyze the organized service manual information [12].

III. PROPOSED SYSTEM

The system is based on the Ontology text mining method. Construction and updation of dependency matrix by auto mining the unstructured text is captured during fault diagnosis using document pre-processing, term extractor and phrase merging for optimization of time.

Construction the D-Matrix follows,

1. Fault Diagnosis
2. Ontology-based Text Mining

1) Fault Diagnosis Ontology:

Ontology is a mechanism that defines the dependencies and relationships currently observe in the domain of interest.

2) Ontology-Based Text Mining:

In this method describe some few steps viz. term extractor, document annotation, and phrase merge denoted in the ontology development of a D-matrix.

Generally, Ontology-based Text Mining phases works as follows:

i. Document Annotation
ii. Term Extraction
iii. Phrase Merging

i. Document Annotation:

The document pre-processing step supports to clarify out the data that is not related for our search and exact context for the reliable understanding of the information. Originally, the document annotation the sentence boundary detection (SBD), is utilized to part a repair verbatim into the partitioned sentenced , stop words are erased to the non-expressive term, and the lexical matching analyze the right significance of abbreviated. Abbreviation disambiguation helps to find out the echo word data count. Afterwards the terms from the processes repair verbatim are coordinated.

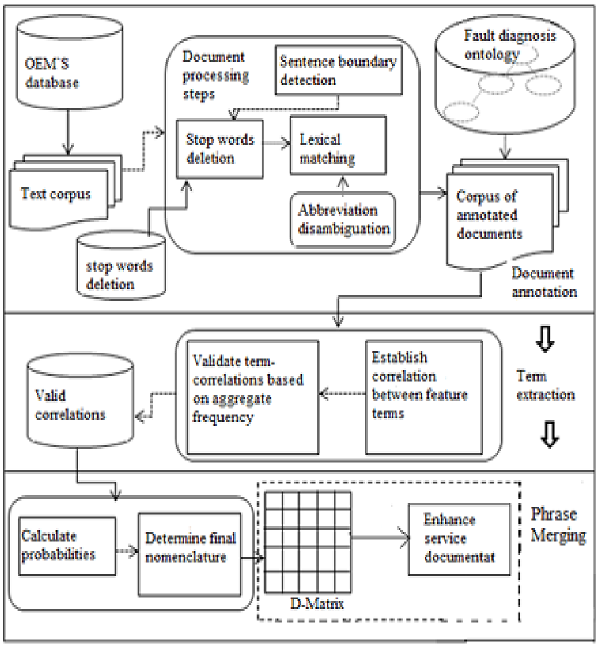

OEM'S database
Text corpus
Document processing steps
Sentence boundary detection
Stop words deletion
Lexical matching
Abbreviation disambiguation
stop words deletion
Fault diagnosis ontology
Corpus of annotated documents
Document annotation
Valid correlations
Validate term-correlations based on aggregate frequency
Establish correlation between feature terms
Term extraction
Calculate probabilities
Determine final nomenclature
D-Matrix
Enhance service documentat
Phrase Merging

Fig. Text-driven D-matrix development methodology from unstructured text data

ii. Term Extractor:

This method we expounded the terms, the critical terms required for the building of a dependency matrix i.e. failure modes and symptoms are extricated by utilizing the term extraction algorithms. Originally, the originative connection among the applicable symptoms-faulty modes sets is distinguished to verify that just the right matches are extricated.

iii. Phrase Merging:

In this method is utilized to decline uncertain reference of the failure condition phrases, where the failure condition phrase that are accounting by utilizing a conflicting

DOI 10.17148/IJARCCE.2016.5735

vernacular. The logical information co-happing with the expressions, namely, symptoms, parts, faulty condition and actions is utilized to gauge the equivalent probabilities and the expressions with their likelihood score over the particular threshold are combining

Conclusively, the new dependency matrix (D-Matrix) is calculated using Fault Diagnosis and Text Mining methods. Which identifies analysis of newly symptoms and faulty parts of the particular domain

Mathematical Model

Mathematical Modeling for document annotation, term extraction, phrase merging involved in the ontology based text mining construction of a D-Matrix, is as follows,

S=is a System, D=set of datasets, A=Document Annotation, T= Term Extraction, M=Phrase Merging.

System S is represented as S = {D, F, DA, TE, PM}

A]. Database

D= {d1, d2, d3, ...,dn}

Where, D is the set of documents and d1, d2, d3, ...,dn are the number of documents.

B]. Fault diagnosis ontology

F={G, I}

G= {g1, g2, g3, ... gn }

Where G is represent as a set of concepts and attributes and g1, g2, g3, …., gn is a number of concepts and attributes.

I= {i1, i2, i3, ..., in }

Where I is the set of instances and i1, i2, i3,..., in represent as a number of instances.

C]. Document Annotation

DA= {P, C}

Where P= {S, T}

Where, P is represent as a set of preprocessing and

S= {s1, s2, s3, ..., sn}

Where S is represent as a set of Stop Words and s1, s2, s3, ...sn number of stop words and

T= {e1, e2, e3, ..., en}

Where T is represent as a set of steaming words and t1, t2, t3, ... tn is a number of steaming word.

C= {c1,c2 ,..,cn}

Where C is the set of corpus annotated document and c1, c2ˆacn represent as a number
of corpus annotated document.

D]. Term Extraction

TE= {U, V}

Here T is a set term extraction and

U= {u1, u2, u3, ...un }

Where U is represent as a set of feature terms and u1, u2, u3, ...,un is a number of feature terms.

V= {v1, v2, v3, ...vn }

Where V is represent as a set of valid correlations and v1, v2, v3,....vn is a number of valid correlations.

E]. Phrase Merging

PM = {Y, Z}

Where Y is represent as a set of Phrase Merging

Y={y1, y2, y3, ..,yn }

Where Y is represent as a set of context information and y1, y2, y3, ...yn is a number of context information.

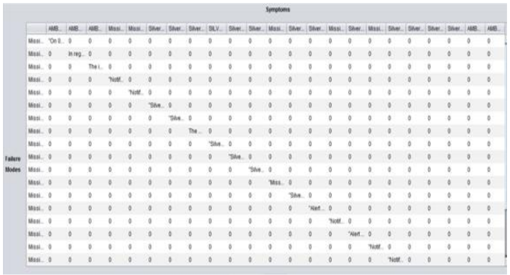
Z={z1,z2, z3, ...zn }

Where Z is represent as a set of merged phrases and z1, z2, z3, ...zn is a number of merged phrases.

IV. RESULT ANALYSIS

The previous method using the data sources that views the overall data that is saving all database and firstly parse that data after that scan overall data so it is takes more data base memory and it is very much time consuming. In my proposed system mainly works on ontology text mining method which will perform auto mining construction and updation the dependency matrix (D-Matrix) for optimization of time. Also it will improve the accuracy of Fault Detection and Diagnosis

In the following table shows the time required for existing system to calculate D-Matrix of few tuples like missing child, utility, mass transit disruption etc. is labor intensive and time consuming than proposed system.



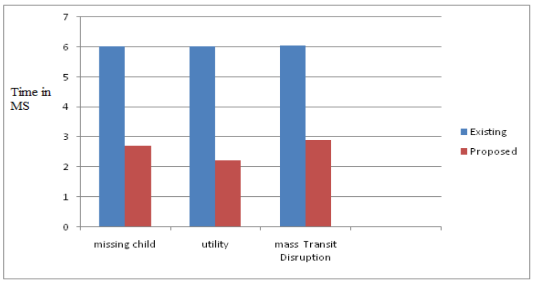
Symptoms

Failure Modes

	AMB...	AMB...	AMB...	Missi...	Missi...	Silver...	Silver...	Silver...	SILV...	Silver...	Silver...	Missi...	Silver...	Silver...	Missi...	Silver...	Missi...	Silver...	Silver...	Silver...	Silver...	AMB...	AMB...
Missi...	"On t...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	In reg...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	The i...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	"Notif...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	"Notif...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	The ...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	"Miss...	0	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	"Silve...	0	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	0	"Alert...	0	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	"Notif...	0	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	"Alert...	0	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	"Notif...	0	0	0	0	0	0
Missi...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	"Notif...	0	0	0	0	0

r. no	Parts failure modes	Existing system	Proposed system
1	Missing Child	6.031	2.75
2	Utility	6.032	2.27
3	Mass Transit Disruption	6.047	2.97

In following fig. After analyzing the generated graph of the resultant D-Matrix shows the efficiency and correctness of proposed system is better than existing system.



CONCLUSION

Ontology based text mining method consist of concepts and relationships commonly observed in the fault diagnosis and then the text mining algorithms makes use of this ontology to identify the necessary artifact such as

DOI 10.17148/IJARCCE.2016.5735

parts, symptoms, failure modes to construct the D-matrix by automatically mining the unstructured repair verbatim data collected during fault diagnosis.

ACKNOWLEDGMENT

The author wish to thank MET's institute of engineering Bhujbal Knowledge city Nashik, HOD of computer department, guide and parents for supporting and motivating for this work because without their blessing this wasn't be possible softly.

REFERENCES

[1] Dnyanesh G. Rajpathak, "An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text," IEEE Transactions on systems, man, and cybernetics: systems, vol. 44, no. 7, July 2014.

[2] O. Benedittini, T. S. Baines, H. W. Lightfoot, and R. M. Greenbush, "State-of-the-art in integrated vehicle health management," J. Aer.Eng., vol. 223, no. 2, pp. 157– 170, 2009.

[3] T. Hearst, "Untangling text data mining," in Proc. 37th Annu.Meeting Assoc. Comput.Linguist, 1999, pp. 3– 10.

[4] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. Kavuri, "A review of process fault detection and diagnosis Part I: Quantitative model based methods," Comput. Chem. Eng., vol. 27, no. 3, pp. 293–311, 2003.

[5] IEEE standard for artificial intelligence exchange and service tie to all test environments (AI-ESTATE), IEEE Std. 1232–2002, 2002.

[6] E. Miguelanez, K. E. Brown, R. Lewis, C. Roberts, and D. M. Lane, "Fault diagnosis of a train door system based on semantic knowledge representation railway condition monitoring," in Proc. 4th IET Int. Conf., 2008, pp. 1–6.

[7] M. Schuh, J. W. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, "A Visualization tool for knowledge discovery in maintenance event sequences", IEEE Aerosp.Electron. Syst. Mag., vol. 28, no. 7, pp. 3039, Jul. 2013.

[8] J.Sheppard, M. Kaufman, and T. Wilmering, "Model based standards for diagnostic and maintenance information integration," in Proc. IEEEAUTOTESTCON Conf., pp. 304–310, 2012.

[9] S. Singh, H. S. Subramania, and C. Pinion, "Data-driven framework for detecting anomalies in field failure Data," in Proc. IEEE Aerosp.Conf., pp. 1–14, 2011

[10] D. Wang, W. H. Tang, and Q. H. Wu, "Ontology-based fault diagnosis for power transformers," in Proc. IEEE Power Energy Soc. Gen.Meeting, pp. 1–8, 2010.

[11] S. Singh, S. W. Holland, and P. Bandyopadhyay, "Trends in the development of system-level fault dependency matrices," in Proc. IEEE Aerosp.Conf., pp. 1–9, 2010.

[12] T. Felke, "Application of model-based diagnostic technology on the Boeing 777 airplane," in Proc. 13th AIAA/IEEE DASC, pp. 1–5, 1994.