IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Text Mining Methodology

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

Student, Computer Department, MET BKC College, Nashik, India[1]

HOD, Computer Department, MET BKC College, Nashik, India[2]

Abstract: D-matrix is a systematic diagnostic model which is used to catch the fault system data and its causal relationship at the hierarchical system-level. Developing a D-matrix from first standards and updating it using the domain information is a labor intensive and time consuming task. Further, in-time augmentation of D-matrix through the discovery of new symptoms and failure modes observed for the first time is a challenging task**.** Is describes construction and updation of D-Matrix by automatically mining the unstructured repair verbatim (written in unstructured text) data collected during fault diagnosis using document annotation, term extraction and phrase merging. The system construct the fault diagnosis ontology consisting of concepts and relationships commonly observed in the fault diagnosis domain. Next, employ the text mining algorithms make use of ontology concept to identify the necessary artifacts, such as parts, symptoms, failure modes, and their dependencies from the unstructured repair verbatim text. The method is implemented as a prototype tool and validated by using real-life data collected from the automobile domain.

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

I. INTRODUCTION

A complex system collaborates with its surrounding to execute a set of assignments by keeping up its performance inside an acceptable range of tolerances. Any deviation of a framework from its worthy execution is dealt with as a fault. The Fault Detection and Diagnosis (FDD) is performed to distinguish the faults and diagnose the root-causes to minimize the downtime of a system. Due to ever growing technological complexity that is inserted in the vehicle system, internet, diagnostic sensors etc [2] the process of fault detection and diagnosis becomes a challenging activity in the occasion of component or system malfunction.

After every diagnosis episode the lessons learnt are kept up in a few databases (e.g., the error codes are put away in on-board PCs of aircraft or automobiles) to detect and diagnose the faults. FDD data comes as unstructured repair verbatim (additionally alluded to as patient medicinal records in medical industry, service technician record in aerospace, automotive etc) that gives a rich source of diagnostic data. It comprises of indications relating to the faulty parts, the observed disappointment modes, and the repair moves made to fix faults. Unstructured repair verbatim data are collected and we contend that there is a pressing need to mine this information to, enhance fault diagnosis (FD). However, the overwhelming size of the repair verbatim data restricts an ability of its effective utilization in the process of FD. Text mining [3] is picking up a serious attention because of its capacity to automatically discover the knowledge assets covered in unstructured text. Text mining strategy to map the diagnosis information separated from the unstructured repair verbatim in a D-matrix [4]. The D-matrix is one of the standard analytic models determined in IEEE Standard 1232 [5]. The development of a D-matrix by utilizing text mining is a challenging task partly because of the noises observed in the repair verbatim text information. The abbreviated text entries: it is used to record the terms and it is essential to disambiguate their meaning. Incomplete text entries: the incomplete repair information makes it hard to determine the exact learning from the data. Term disambiguation: the same term is composed by using inconsistent vocabulary, e.g. FTPS Inop and FTPS Internal Short.

The process of FD starts by extracting the error codes from a target system and based on the observed error codes the technicians follow specific diagnosis procedure along with their experience to diagnose the faults. During fault diagnosis, several data types such as error codes, scanned values of operating parameters associated with faulty component/system, repair verbatim, and so on are collected.

The collected data is then transferred to the OEM database and particularly the repair verbatim data collected over a period of time can be mined to develop the D- matrix diagnostic models. To perform accurate FDD, Such models can be used by the field technicians and other stakeholders. The D-Matrix captures component and system level dependencies between a single or multiple failure modes with a single or multiple symptoms in a structured fashion.

These dependencies among failure modes (f1, f2, etc.) in parts (p1, p2, etc.) and symptoms (s1, s2, etc.) allow us to state a set of failure modes causing symptoms. A probability of detection, the causal weights (d11, d12, etc.) is contained at the intersection of a row and a column indicates. In the binary D-matrix, all the probabilities have a value of either 0 or 1, where 0 indicates no detection and 1 indicates complete detection of a specific failure mode using a specific symptom. The values between 0 and 1

DOI 10.17148/IJARCCE.2016.5459

indicate the level of strength of detecting a failure mode by using a symptom.

Generally, the D-matrix built by utilizing the history information, engineering information, and sensory information, for sample, [6]. However a practically nothing understanding is given about the disclosure of new symptoms furthermore, failure modes observed first time and their incorporation in the D-matrix models. In the methodology the occasional growth of the deficiency finding fault diagnosis ontology helps the content mining calculation to construct the correct D-matrix.

II. RELATED WORK

Dnyanesh G. Rajpathak et.al. have proposed to construct the D-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis. In real-life, manual construction of D-matrix diagnostic model corresponding to the complex systems is not practical as it would involve significant effort to integrate the knowledge from SMEs and represent it in a D-matrix. This approach overcame these limitations where natural language processing algorithms were proposed to automatically develop the D-matrices from the unstructured repair verbatim. They compared the testability and diagnosability metrics of the historical data-driven D-matrix and the text-driven D-matrix. They have also proposed naive Bayes probability model for developing abbreviated terms by considering context. Their methodology for D-matrix construction consists of three building blocks document annotation, term extraction, and phrase merging [1]. M. Schuh et.al. have proposed about discovery of knowledge from the on-board Diagnosis by using the ontology-based data mining. The onboard diagnosis collects the real time data and integrates onboard ECUs. This model is assumed to be static and complete. But in real world, due to engineering changes and design, new vehicle structure and vehicle architecture is launching. But in this approach to Faults are removed and provided ontology-guided data mining and data transformation, But Discovery is loss because result is not in form of matrix [7].

J. Sheppard, et.al. have proposed Model based standards for diagnostic and maintenance information integration in which the limited efforts are done to create a D-matrix by analyzing unstructured repair verbatim data. D-matrix constructed by using the data sources that views the overall data that is saving all database and firstly parse that data and after that scan overall data so it is takes more data base memory and its very much time consuming for parsing and scanning that overall databases [8]. S. Singh, et.al. have proposed data-driven framework for detecting anomalies in field failure data. The subject matter expert generally detects the anomalies by manually working and sorting the field failure data using spreadsheets which is time consuming and labor-intensive process. Therefore a data-driven framework is develop which automatically detect the unusual activity that leads to fault and saves a significant expert's time. This framework is developed so that they could completely work on analyzing anomalies and taking proper actions [9]. D. Wang et.al. proposed Ontology-based fault diagnosis for power transformers in which a new approach to transformer fault diagnosis, which is based on the idea of exchanging information with formal, explicit and machine accessible descriptions of meaning using the Semantic Web. This ontology model captured the crucial terms in the fault diagnosis of power transformer, e.g., fault phenomenon, fault reason, along with the relations among them [10]. S. Singh et.al. proposed trends in the development of system-level fault dependency matrices developing D-matrices from dissimilar information format and data sources. The D-matrices is classified based on their data source and the imperfectness of symptoms. They have Considered for both Boolean-value and real-valued [0, 1] D-matrices [11]. T. Felke [12] proposed the fault diagnosis D-matrix models have been used successfully in aerospace industry to identify the dependencies among failure modes, symptoms, and repair claims by analyzing the structured service manual data.

III. SYSTEM OVERVIEW

An ontology based text mining method for automatically constructing and updating D-matrix by mining hundreds of thousands of repair verbatim (typically written in unstructured text) collected during the diagnosis episodes. To construct the D-matrix, following steps has to be executed,

1. Fault Diagnosis Ontology
2. Ontology-based Text Mining

(i) **Fault Diagnosis Ontology:** Ontology is a mechanism that describes the concepts and also the relationships that hold between those concepts observed in the domain of vehicle fault diagnosis. Ontology system for the fault diagnosis of automobile systems, it is necessary to analyse numerous concepts and relationships exhibited.

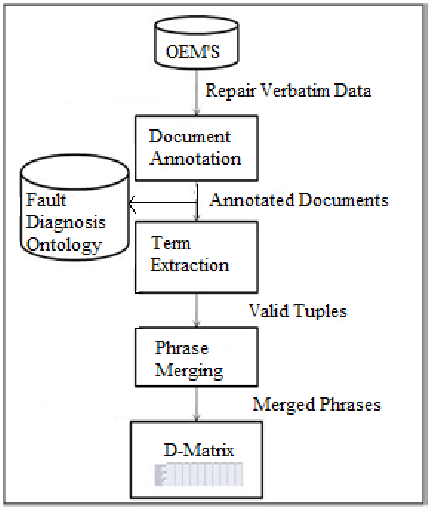

Fig.1. Text-driven D-matrix development methodology from unstructured text

(ii)**Ontology-Based Text Mining:** In this method describe some few steps viz. term extractor, document annotation,

DOI 10.17148/IJARCCE.2016.5459

and phrase merging involved in the ontology based text mining construction of a D-matrix.

(a)**Document Annotation:**The document annotation helps to filter out the information that is not related for analysis and it provides a specific background for the reliable understanding of the data. Initially, the pre-processing steps—the sentence boundary detection (SBD), are utilized to part a repair verbatim into partitioned sentences, the stop words are erased to the non-expressive terms, and the lexical matching identifies the right significance of abbreviations. The abbreviation disambiguation helps to find out the repeated word data count. Afterwards the terms from the processed repair verbatim are matched.

(b)**Term Extractor:** In this method we expounded the terms, the critical terms required for the development of a D-matrix i.e. symptoms and failure modes are extracted by utilizing the term extractor algorithms. Initially, the causal connection between the relevant symptom-failure mode sets is distinguished to verify that just the right matches are extracted.

(c)**Phrase Merging:** The phrase merging is used to avoid ambiguous references of the failure mode phrases, where the failure mode phrases that are written by using an inconsistent vocabulary, e.g., Tank Pressure Sensor-Short, or FTP Sensor-Internal Short, or Fuel Tank Pressure Sensor-Internal Short Observed are merged into a single, consistent failure mode phrase, e.g., Fuel Tank Pressure Sensor-Internal Short to maintain the homogeneity. The contextual information co-occurring with the phrases, i.e., parts, symptoms, failure mode, and actions is used to estimate the conditional probabilities and the phrases with their probability score above the specific threshold are merged.

IV.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 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.

DOI 10.17148/IJARCCE.2016.5459