Development of Graph from D-matrix based on Ontological Text Mining Method

Tinal R. Thombare
PG student
Dept of Computer Science and Engg
Nagpur (M.S), India
tinalthombare73@gmail.com

Lalit Dole
Assistant Professor
Dept of Computer Science and Engg
Nagpur (M.S),India
lalit.dole@gmail.com

***Abstract*—** Fault dependency (D-matrix) is a diagnostic model that catches the fault system data and its causal relationship at the hierarchical system-level. It consists of dependencies and relationship between observable failure modes and symptoms associated with a system. Constructing such D-matrix fault detection model is time consuming task. In this paper, a system is proposed that describes an ontology based text mining method for automatically constructing D-matrix by mining hundreds of repair verbatim text data (typically written in unstructured text) collected during the diagnosis episodes. First we construct fault diagnosis ontology and then text mining techniques are applied to identify dependencies among failure modes and observable symptom. D-matrix is represented in graph so that analysis gets easier and faulty parts becomes easily detectable. The proposed method will be implemented as a prototype tool and validated by using real-life data collected from the automobile domain.

Keywords- *Fault diagnosis; fault detection; information retrieval; dependency-matrix; text mining.*

*****

I. INTRODUCTION

A complex system interacts with its surrounding to execute a group of tasks by maintaining their performances within an appropriate vary of tolerances. Any variation of a system from its acceptable performance is treated as a fault. The fault detection and diagnosis (FDD) is performed to observe the faults and diagnose the root-causes to reduce the period of time of a system. Because of ever growing technological sophistication that's embedded within the vehicle systems, for example refined computer code embedded systems [1], diagnostic sensors, internet, etc. the method of FDD becomes a difficult activity within the event of component or system malfunction.

Unsurprisingly, when each diagnosis episode the lessons learnt are maintained in many databases (e.g., the error codes are hold on in on-board computers of aircrafts or automobiles) to observe and diagnose the faults. One usually book unbroken diagnosis information comes within the form of unstructured repair verbatim (also brought up as patient medical records in medical business or service technician records in part, automotive, power plants and producing industries), that supply rich diagnostic info. It consists of symptoms equivalent to the faulty components, the observed failure modes, and therefore the repair actions taken to repair the faults. Hundreds of thousands of such repair verbatim are collected and there is urgent need to mine this information to improve fault diagnosis (FD). However, the overwhelming size of the repair verbatim information restricts a capability of its effective utilization within the method of FD.

Text mining is gaining a heavy attention because of its ability to mechanically discover the data assets buried in unstructured text. During this paper, we tend to propose a text mining method to map the diagnostic info extracted from the unstructured repair verbatim in a very D-matrix [3]. The D-matrix is one amongst the quality diagnostic models laid out in IEEE Standard. This framework catches causal connections between symptoms and failure modes in structured fashion. This framework is called as Dependency or Diagnosis framework (D-matrix). A failure modes contains root cause of a system and symptoms contains a set of fault codes, diagnostic trouble codes, automated tests, technician tests, observed symptoms, etc.

Typically the process of fault diagnosis begins with the extraction of the error codes that are present in the system and based on the observed error codes the technicians follow some diagnostic procedure along with their experience to identify the nature of faults. During FD, different data types are collected, like error codes, diagnostic trouble codes faulty associated with the target system, repair verbatim, and so on. This collected data is then transmitted to the database. This data can be mined to construct diagnosis matrix models. Such models can be used by the technicians and stakeholders to find accuracy of fault detection.

II. LITERATURE WORK

This section mainly concentrates on how fault detection and diagnosis (FDD) is done to detect the faults through D-matrix framework related to the automobile domain. In existing fault models, the knowledge which was embedded in the unstructured repair verbatim (unstructured text) data have improved the performance of fault diagnosis by introducing an approach for constructing D-matrices based on an ontology-based text mining method[1].

In the [11], [12], [13], [4], the limited efforts are done to create a D-matrix by analyzing unstructured repair verbatim. Recently [16] the tool is proposed which discovers the 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. The vehicles launch new symptoms and failure modes. Some of the tools suffer some drawbacks related to the symptoms and fault parts.

To develop a D-matrix framework, a method is proposed that analyze the unstructured repair verbatim data by using ontological text mining methods associated with the multiple systems in parallel. Previously D-matrix

frameworks were constructed manually. So this method overcomes the problems in the real life industry of having to construct model manually.

Traditionally, the D-matrices are constructed by using the knowledge buried in the field failure data. The data includes historical data, engineering data, and sensory data, error codes [11], [12], [13], [4], [14], [15], but the authors have not provided any perception for new symptoms and failure modes which are observed for the first time and their insertion in the D-matrix models. The periodic or prior knowledge is necessary for constructing D-matrix fault diagnosis model to make it more accurate.

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 [2] 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. Further in [6] the researcher worked on 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.

The fault diagnosis D-matrix models have been used successfully in aerospace industry [9], [10] to identify the dependencies among failure modes, symptoms, and repair claims by analyzing the structured service manual data.

III. PROPOSED WORK

Our methodology consists of ontological text mining method. The fault diagnosis ontology is formalized by using the ontology development methodology. It captures the terms and the relations observed in the automobile fault diagnosis domain. The concepts system, subsystem, and part formalize the main parts that are under focus during fault diagnosis. The concept FailureMode forms the system level engineering faults observed during the root-cause investigation, Further, the concept attributes e.g., hasCause (fault cause) capture the domain specific data with the internal structure of the ontology and a minimal set of attributes are used to formalize the concepts.

Mainly our methodology contains two main module namely document pre-processing steps and extraction of relevant terms with probability calculations involved in ontological text mining method. Figure 1. shows the flow diagram of the proposed work.

A. *Document preprocessing step:*

Due to the many kinds of noises observed in our knowledge the task of identifying the main building blocks of D-matrix, like fault components, symptoms, and failure modes becomes a non-trivial exercise. The document preprocessing helps to remove the data that is irrelevant for our analysis and it provides a selected context [26] for the consistent and shared interpretation of the data.

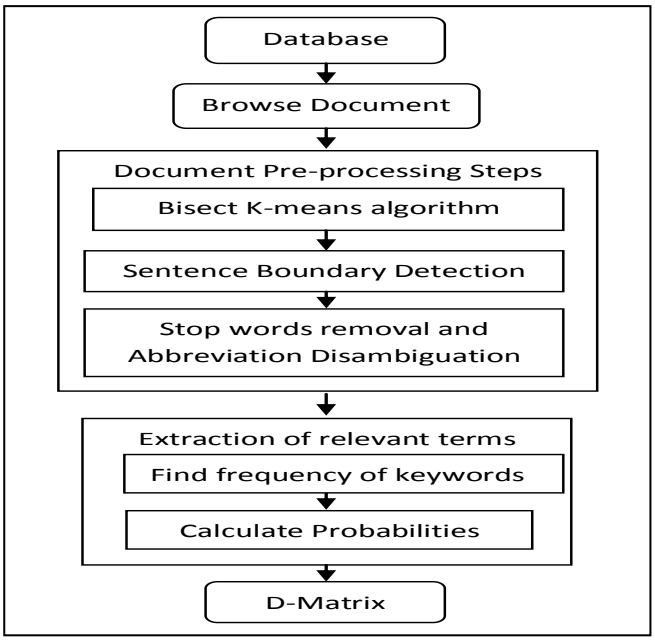

Fig. 1: Flow diagram of Proposed work

Initially, the preprocessing steps consist of bisect k-means clustering algorithm. This algorithm has following steps—the sentence boundary detection (SBD) splits a repair verbatim into separate sentences, the stop words are deleted to get rid of the non-descriptive terms, and also the lexical matching identifies the correct that means of abbreviations. The abbreviation disambiguation helps to find out the repeated word data count. Afterwards the terms from the processed repair verbatim are matched.

A typical repair verbatim (*Vehicle*^{P} *ck.*^{A} *for hard start*^{S} *after codes P0451*^{S}*...P0452*^{S} *to found PCM has internal−Short*^{FM}*. Replaced* ^{A} *PCM*^{P} *and0.3olh is claimed*^{A}) consists of multiple components ^{P}, symptoms ^{S}, failure modes ^{FM}, and actions ^{A} and it's necessary to spot the right associations between them for constructing a D-matrix. We have a tendency to take the perspective that the terms that are showing in an exceedingly same sentence represent high relatedness in comparison with those that are written in separate sentences. In sentence boundary, each repair verbatim is split into completely different sentences by verifying its starting and end. The task of identifying the sentence boundary is called as the sentence boundary detection (SBD) in natural language processing. Many approaches have been projected to identify the SBD and a few crucial instances include Satz system [27], Alembic [28], mxTerninator [29], and Punkt [30]. The in-house SBD heuristics are developed to see the sentence boundary by mistreatment a amount as a boundary delimiter and such periods are considered that are used to specify the sentence endings by successfully ignoring different instances during which they're used to specify the abbreviations.

Having split every repair verbatim, the non-descriptive stop words (for example, a, an, the), that don't seem to be member of the critical component, symptom, or failure mode phrases area unit deleted to reduce the noise. In our data, some abbreviations, like Abbri, have different meanings based on the context in which they are used, for example, TPS: Tire Pressure Sensor or Tank Pressure Sensor and it is difficult to identify their correct meaning before constructing the D-

ISSN: 2321-8169
2765 - 2768

matrices with the terms. In literature survey, different approaches have been proposed to handle the abbreviation disambiguation problem, for example, [31]–[33] either by using the Naive Bayes, or decision trees, or Support Vector Machines algorithms.

B. *Extraction of relevant terms:*

After the preprocessing steps, the critical terms which are useful for constructing D-matrix, i.e., symptoms and failure modes are extracted by using the extraction of relevant terms. Initially, the relationship between the relevant symptom-failure mode pairs that seems to be causal is identified to make sure that only the correct pairs are extracted. The existing approaches [34] for frequent item sets mining ignores the order during which the term phrases are recorded in documents, but we should maintain such ordering to grasp however the fault identification is performed. The frequently occurring terms i.e. fault parts are considered as keywords/constant in fields of symptoms, failure modes and repair action. The number of such keywords in symptoms, failure modes and repair action is calculated for finding frequency. The frequency is found on the basis of the maximum occurrence of the fault part in the repair data.

The contextual information i.e., parts, symptoms, failure mode and actions are used to estimate the conditional probabilities. As the D-matrix catch component and system level dependencies between a single and multiple failure modes with a single and multiple symptoms (a set of fault codes, observed symptoms, etc.) in a structured way. Using bayes theorem, 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. The causal weights (d11, d12, etc.) are contained at the intersection of a row and a column indicates a probability of detection. 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. After formation of D-matrix, each can be represented as graph.

IV. RESULTS AND ANALYSIS

. The D-Matrix framework is created using proposed methodology. The real-life data is collected from the automobile car system. A text driven D-Matrix is created using the symptoms shown in column and failure mode in rows. 0 and 1 represents the probability of detection of faults. The components of car are taken that have faults occurred during fault diagnosis and detection. Table 1. shows the text driven D-Matrix of the car system components containing some critical fields.

Figure 2 shows the representation of D-matrix into Graph which consist of car components. This graph represents the fault parts and their frequency of occurrences. Due to limited size, it is not possible to show every field in D-matrix therefore all the fields are shown in graph. Figure 3 shows the comparison of the previous technique with the proposed technique in terms of fault detection. The fault detection (FD) is given as the percent of faults detected by the symptoms by observing the failure modes associated with a system. It was used to evaluate the fault coverage. Less than 100% fault detection indicates that there are faults, which cannot be detected by this system. Previous technique is shown in blue and proposed technique is shown in yellow. It is observed that more than 80% faults are detected using proposed system only less than 20% does not gives accuracy of the results.

TABLE I. TEXT DRIVEN D-MATRIX OF CAR SYSTEM COMPONENTS

Symptoms → Part_failure mode ↓	Gear slipping	Loud noise	Rattling noise when starting acceleratin g or braking	Dra ggi ng clut ch	Grin ding or shak ing
Backing plate_air spring fails	0	0	1	0	1
Anti-lock braking_gear slipping	0	1	0	1	1
Bumper1_increased engine operating temperature	0	0	0	0	0
Headlight moto_uneven engine idling	1	0	0	0	0
Controlsystem_ lack of response	0	1	1		1

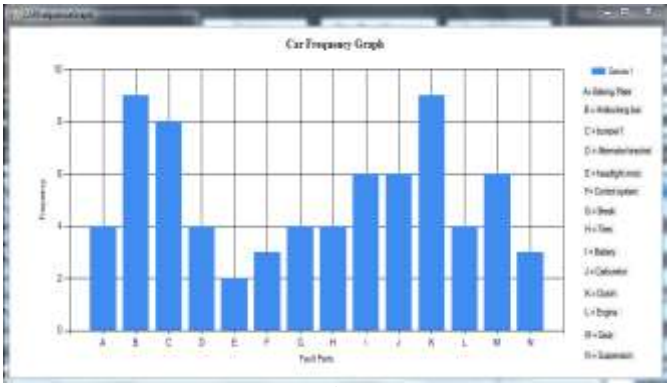

Fig. 2. Representation of D-matrix into graph

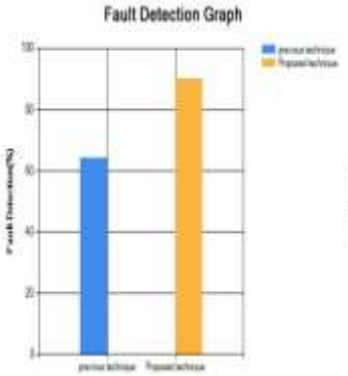

Fig 3: Fault detection graph

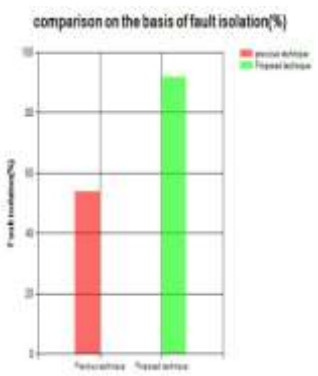

Fig 4: Fault isolation graph

Figure 4 shows the comparison of the previous technique with the proposed technique in terms of fault isolation. The fault isolation (FI) is defined as the probability that the symptoms and failure mode do not have any

dependency associated with a system. In other words a fault can be treated isolatable if it has 0. Some of the components have low fault isolation rate (around 50%) and does not give accuracy of finding it in the previous technique. In some instances, it was not necessary to isolate to a single fault 100% of the time. Previous technique is shown in red and proposed technique is shown in green. It is observed that accuracy of finding fault isolation was more in proposed technique compared to previous technique.

V. CONCLUSIONS

In this paper an ontology-based text mining methodology is implemented to construct the D-matrix by automatically mining the unstructured repair verbatim text data to structured text collected during fault diagnosis. This framework helps the service technician to detect the faults related to complex system and diagnose it. This dependency model (D-Matrix) contains symptoms, failure modes and their causal relationship. Using probabilistic approach, fault detection and isolation has become easier. Development of a graph from D-matrix model gives better visualisation and analysis. It helps in real world industry to identify the necessary facts.

REFERENCES

[1] Dnyanesh G. Rajpathak, Satnam Singh, "An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text", IEEE trasactions on system, man and cybernetics : systems, vol. 44, no.7, pp. 966-977, Jul. 2014.

[2] 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.

[3] R. Chougule and S. Chakrabarty, "Application of ontology guided search for improved equipment diagnosis in a vehicle assembly plant", in Proc. IEEE CASE, pp. 90–95, 2009.

[4] S. Strasser, J. Sheppard, M. Schuh, R. Angryk, and C. Izurieta, "Graph based ontology-guided data mining for d-matrix model maturation," in Proc. IEEE Aerosp. Conf., pp. 1–12, 2011.

[5] 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.

[6] 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.

[7] T. J. Felke and J. F. Stone, "Method and Apparatus for Developing Fault Codes for Complex Systems Based on Historical Data", US Patent 003318 A1, Jan. 2004.

[8] S.P. Eagleton and T. Felke, "Method and Apparatus using Historical data to Associate Deferral Procedures and D-matrix", US Patent 6,725,137 B2, Apr. 2004.

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

[10] G. Ramohalli, "The Honeywell on-board diagnostic and maintenance system for the Boeing 777", in Proc. IEEE/AIAA DASC, pp. 485–490, 1992.

[11] 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., pp. 1–6, 2008.

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

[13] M. Schuh, J. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, "Ontology-guided knowledge discovery of event sequences in maintenance data", inProc. IEEE AUTOTESTCON Conf., pp. 279–285, 2011.

[14] S. Deb, S. K. Pattipati, V. Raghavan, M. Shakeri, and R. Shrestha, "Multi-signal flow graphs: A novel approach for system testability analysis and fault diagnosis", IEEE Aerosp. Electron. Syst.., vol. 10,no. 5, pp. 14–25, May 1995.

[15] S. Singh, A. Kodali, K. Choi, K. R. Pattipati, S. M. Namburu, S.C. Sean, D. V. Prokhorov, and L. Qiao, "Dynamic multiple fault diagnosis: Mathematical formulations and solution techniques", IEEE Trans. Syst., Man Cybern. A, Syst. Humans, vol. 39, no. 1, pp. 160–176,Jan. 2009.

[16] 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. 30–39, Jul. 2013.

[17] P. M. Frank and J. W¨ unnenberg, "Robust fault diagnosis using unknown input observer schemes", in Proc. Fault Diagnosis Dynamical Syst.: Theory Appl., pp. 47–98, 1989.

[18] N. Viswanadham and R. Srichander, "Fault detection using unknown input observers", Control-Theory Ad. Tech., vol. 3, pp. 91–101, 1987

[19] P. M. Frank, "Fault detection in dynamic systems using analytical and knowledge-based redundancy—a survey and some new results", Automatica, vol. 26, no. 3, pp. 459–474, 1990.

[20] A. S. Willsky, "A survey of design methods for fault detection in dynamic systems", Automatica, vol. 12, no. 6, pp. 601–611, 1976.

[21] V. Venkatasubramanian and S. H. Rich, "An object-oriented two-tier architecture for integrating compiled and deep-level knowledge for process diagnosis", Comput. Chem. Eng., vol. 12, no. 9–10, pp. 903–921, 1988.

[22] C. Charniak and D. McDermott, Introduction to Artificial Intelligence. Reading, MA, USA: Addison Wesley, 1985.

[23] V. R. Benjamins, "Problem-solving methods for diagnosis and their role in knowledge acquisition," Int. J. Expert Syst.: Res. Appl., vol. 8, no. 2, pp. 93–120, 1995.

[24] M. Iri, K. Aoki, E. O'Shima, and H. Matsuyama, "An algorithm for diagnosis of systems failures in the chemical process", Comput. Chem. Eng., vol. 3, nos. 1–4, pp. 489–493, 1979.

[25] T. Umeda, T. Kuriyama, E. Oshima, and H. Matsuyama, "A graphical approach to cause and effect analysis of chemical processing systems", Chem. Eng. Sci., vol. 35, no. 12, pp. 2379–2388, 1980.

[26] M. Agosti and N. Ferro, "Annotations as context for searching documents", inProc. Int. Conf. Concept. Library Inf. Sci.—Context: Nature, Impact Role, LNCS, pp. 155–170, 2005.

[27] D. D. Palmer and M. A. Hearst, "Adaptive multilingual sentence boundary disambiguation", Comput. Linguist., vol. 23, no.2, pp. 241–318, 1997.

[28] J. B. Aberdeen, D. Day, L. Hirschman, P. Robinson, and M. Vilain, "MITRE: Description of the alembic system used for MUC-6", inProc. Conf. Message Understand., pp. 141–155, 1995.

[29] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries", in Proc. 5th Conf. ANLP, pp. 16–19, 1997.

[30] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection", Comput. Linguist.,vol. 32, no. 4, pp. 485–525, 2006.

[31] H. Liu, Y. Lussier, and C. Friedman, "Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method", J. Bio. Inf., vol. 34, no. 4, pp. 249–261, 2001.

[32] H. Yu, W. Kim, V. Hatzivassiloglou, and J. Wilbur, "A large scale, corpus-based approach for automatically disambiguating biomedical abbreviations", ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 380–404, 2006.

[33] M. Joshi, T. Pedersen, and R. Maclin, "A comparative study of support vector machines applied to the word sense disambiguation problem for the medical domain", in Proc. 2nd IICAI, pp. 3449–3468, 2005.

[34] M. F. Porter, "An algorithm for suffix stripping", Program, vol. 14, no. 3,pp. 130–137, 1980.