International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Efficient D-Matrix construction strategy for Unstructured Text Mining using FDD

Swati S Hinge[1],B. R. Nandwalkar[2]

[1]Department of Computer Engineering, KCT's Late G.N. Sapkal College of Engineering,
swatih110@gmail.com ,
[2]Department of Computer Engineering, KCT's Late G.N. Sapkal College of Engineering,
nandwalkar.bhushan@gmail.com,

Abstract—The term text mining is to perform operation on unstructured (text) information, from the noisy text extract meaningful text data (information), and thus make sure that the information consist in the text available to the various data mining. There is necessity to collect information regarding various symptoms and failure modes to alter the fault dependency matrix which can be helpful to construct correct and efficient fault diagnosis. To decrease the system's downtime the fault detection and diagnosis (FDD) is performed to detect the faults and diagnose the root-causes. In process of engineering Fault detection and diagnosis is an important issue. In proposed system, we initially develop the fault diagnosis ontology of attributes and relationships with it observed in the fault diagnosis database. Next, we construct the algorithms of text mining which will use of the ontology to identify the important artifacts from the unstructured repair verbatim text which is online as well as offline.

Keywords-Text mining, verbatim text, ontology, fault diagnosis, ontology.

I. INTRODUCTION

A system conducts the fault detection and diagnosis (FDD) to reduce the downtime and recognize the faults origin-causes of a system and diagnose that cause [1]. Recent abundance data can generate modern diagnostic systems. The gathered data is distributed across several wide-ranging systems. System cannot easily collect it and aggregated for use. To access the data much of user's specific inquiry is uneventful to the users. This puts a very large burden on the user to amalgamate the data and mine the useful bits relevant to their present requires [2]. So, result in failures of multiple components is more difficult due to complexity. So, there is a require to construct smart diagnostic algorithms that can decide the most likely set of failure causes in a system, given noticed test output result over time [3].

In proposed system, we proposed a text mining technique to map the diagnostic information extracted from the unstructured repair verbatim in a D-matrix [4]. The D-matrix development by using text mining is a demanding job partly due to the unnecessary data noticed in the repair verbatim text data – abbreviated text entries, term disambiguation and incomplete text entries. As shown in fig.1, the process of FD can starts by extracting the error from a required system and which is based on the observed error codes the technicians follow particular diagnosis step can diagnose the faults across with as per the knowledge. The data of various types are assembled throughout fault diagnosis, like as error codes, fetched operating parameters values related with faulty, repair verbatim, system. The collected data can then passed to the particular database and specifically the repair verbatim data gathered over a period of time that should be mined to construct the D-matrix diagnostic models.

The work of this system, can divided into the data-driven and quantitative fault diagnosis classifies, where a text-driven D-matrix construction methodology on online database can propose, where the fault diagnosis ontology is constructed initially by mining the unstructured repair data.

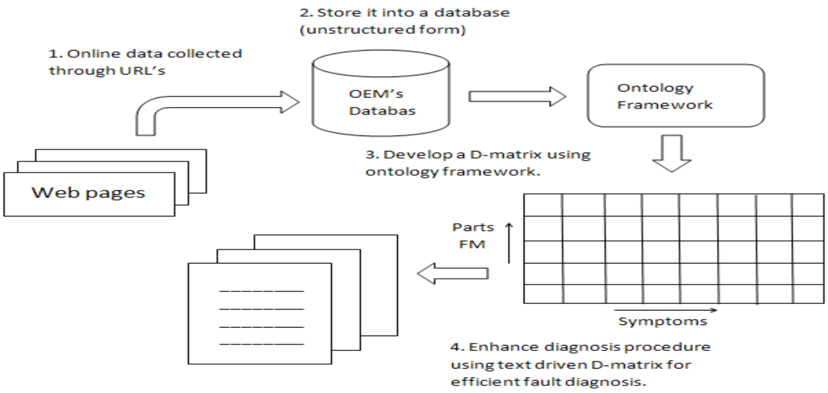
2. Store it into a database (unstructured form)
1. Online data collected through URL's
OEM's Databas
Ontology Framework
Web pages
3. Develop a D-matrix using ontology framework.
Parts
FM
Symptoms
4. Enhance diagnosis procedure using text driven D-matrix for efficient fault diagnosis.

Fig.1. Overview of a system

II. LITERATURE SURVEY

In 1999, the Marti A. Hearst in [5] suggest a Text Data Mining (TDM) method It was to locate new information from data, search out patterns across datasets, and differentiating signal from noise. A document gives from retrieval of information system was consisting of the information a user application implies that no new discovery was being made. In LINDI project the main tools for finding early information were of two types: keep going for concerning series of queries and narrated operations with text gathered, and visualization tools and tightly paired statistical for the examination of relations among concepts that co-occur within the retrieved documents.

In 2009 [8], the Satnam Singh, Kihoon Choi, AnuradhaKodali and Krishna Pattipati implemented a solution which could be shows as a two-level framework that contains correlated solution for the DMFD problem. At the high (correlation) level, they upgrade the range multipliers (correlation variables, dual variables) by using the gradient method. The topmost level makes easier correlation between each of the subproblems, and it could be in the vehicle-level diagnostic control unit. At the lower level, they used a dynamic programming method to sort out each of the subproblems. The key pros of their method were that it provided an approximate duality gap, which was determine of suboptimality of the DMFD description. Interestingly, the perfectly-observed DMFD problem guided to a vital set covering problem, which could be approximately calculated via Lagrangian relaxation and Viterbi decoding.

In [1] Dnyanesh G. Rajpathak, Satnam Singh, 2014 they implemented the ontology for fault diagnosis which include all concepts and relationships commonly noticed in the fault diagnosis domain. Next, apply the text mining algorithms that develop and utilize of this ontology to identify the mainly artifacts, like parts, symptoms, failure modes, and their associations from the unstructured repair verbatim data. It was applied to develop a text-driven D-matrix and it was concluded that our method successfully flushed the failure modes and symptoms along with their association from all the related systems to develop the D-matrix. The LDA associations determine that all the failure modes didn't have a similar probability and a high probability was stated to all the symptoms illustrate failure modes agree with to the same system, whereas a tiny probability was allocated to the symptoms arising from the different systems. In fact, it was necessary to find out the symptoms and failure modes coming from all the related systems for accurate FD. They feel that there was including the domain specific relations while constructing the matrices to further increase the performance with a need to supply an addition to the LDA model by.

In the existing technique or system of fault modeling [3], [6], [9], and [10] the restricted work done we identified that by analyzing unstructured repair verbatim data to develop a D-matrix. Only recently [2] the tool is suggest that locates the knowledge by collecting similar sequences from the on-board diagnosis and conservation data by using the ontology-based data mining from the online database. Our system focus the performance of our system when differentiate with the Latent-Dirichlet Allocation (LDA) technique. Traditionally, the D-matrices were implemented by utilizing the sensory data, engineering data, and history data, for example, [3], [6], [7] but a very small insight is supplied about the discovery of currently updated developed symptoms and failure modes observed for the first time and their inclusion in the D-matrix models which will going to work on online as well as offline database.

III. PROPOSED SYSTEM

The proposed system working mainly focuses on following three modules:

1) Document Annotation module.
2) Term Extraction module.
3) Text Mining module.

3.1 Document Annotation: In this module, initially during field FD process the verbatim data points are assemble by retrieving those from the database are documented. In this step, the terms, are annotated like failure mode symptom, parts relevant for the annotated like failure mode symptom, parts relevant for the D-matrix are from the entire repair verbatim by constructing the algorithm for document annotation. Here, by utilizing the sentence boundary detection rules a repair verbatim is firstly split in distinct sentences by and the terms come into view in the same sentence are co-related with each other. The sentence boundary detection (SBD), are used to split a repair verbatim into separate sentences, the stop words are tale out to remove the non-descriptive terms, and the lexical matching recognizes the accurate meaning of abbreviations. Fault diagnosis ontology incarcerates the term a relations discovered in the domain of fault diagnosis. Subsequently the terms from the exercised verbatim are matched using the objects in the fault diagnosis ontology.

3.2 Term Extraction: As an input module 2 take tuples form module 1 which are constructed by using the document annotation algorithm to inhabit a D-matrix. In term extraction having annotated the terms, for the development of a D-matrix critical terms are needed, i.e., symptoms and failure modes are extracted by using the term extractor algorithm. At the initial stage, relevant symptom-failure mode pairs and casual relation between them is detected to conclude that only the exact pairs of tuples are extracted. The annotated terms are extracted tuples are implement first by identifying the all the failure modes, and symptom. Next, the tuples are integrating using parts as the common tuple member and the tuples are developed. Typically, from the corpus several tuples are building, but only the relevant tuples must be maintained at the time of developing a D-matrix correlating to a particular system. At the end of this step, tuples are developed.The tuples clustered normalized frequency is calculated and the tuples with their frequency above a particular threshold are kept as the reasonable tuples.

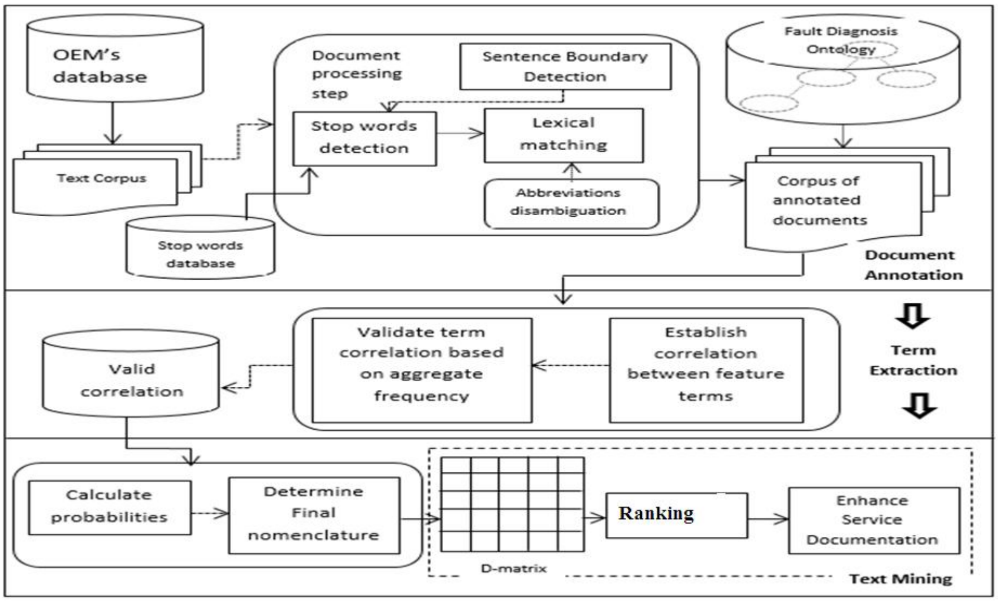
OEM's database
Text Corpus
Stop words database
Document processing step
Sentence Boundary Detection
Stop words detection
Lexical matching
Abbreviations disambiguation
Fault Diagnosis Ontology
Corpus of annotated documents
Document Annotation
Term Extraction
Valid correlation
Validate term correlation based on aggregate frequency
Establish correlation between feature terms
Calculate probabilities
Determine Final nomenclature
D-matrix
Ranking
Enhance Service Documentation
Text Mining

Fig.2 Architecture of online text mining using D-matrix from unstructured data.

3.3Text Mining: Next, the text mining algorithm is used to avoid failure mode phrase's ambivalent references. The failure mode phrases that are written by using an inconsistent vocabulary, are combined into a consistent and single failure mode phrase, maintain the homogeneity. The contextual data co-occurring along with the phrases, i.e., parts, symptoms, failure mode, and actions is utilized to estimate the conditional probabilities and the phrases with their probability score above the particular threshold can be merged. Finally, the newly develop D-matrix is audited by subject matter experts (SMEs) to concluded the discovery of new symptoms and failure modes for in-time FDD.

IV. ALGORITHM

4.1 Data Mining:

Input: parts and failure mode (FM) pairs.
Output: structured data (tabular form).
1.Two FM are extracted by term extraction algorithm. Pairs are randomly selected and corpus of repair verbatim is filter into a subset.
2. Attributes of parts and FM mode will be collected in parallel manner.
3. The high degree of similarity association is observed with mode.
4. Common attributes are share sample space i.e. filter the corpus.
5. Average cumulative sum of the estimated probability is above threshold get merged with each other.
6. D - Matrix will generate according to FM, symptoms, parts.
7. High number of attributes get high priority.

V.RESULTS

For online and offline system patient's dataset is used. user should enter the any symptoms according to the symptoms disease names, causes and diagnosis can be viewed to the user. Following graph shows the comparison of proposed system D-matrix and existing system D-matrix.

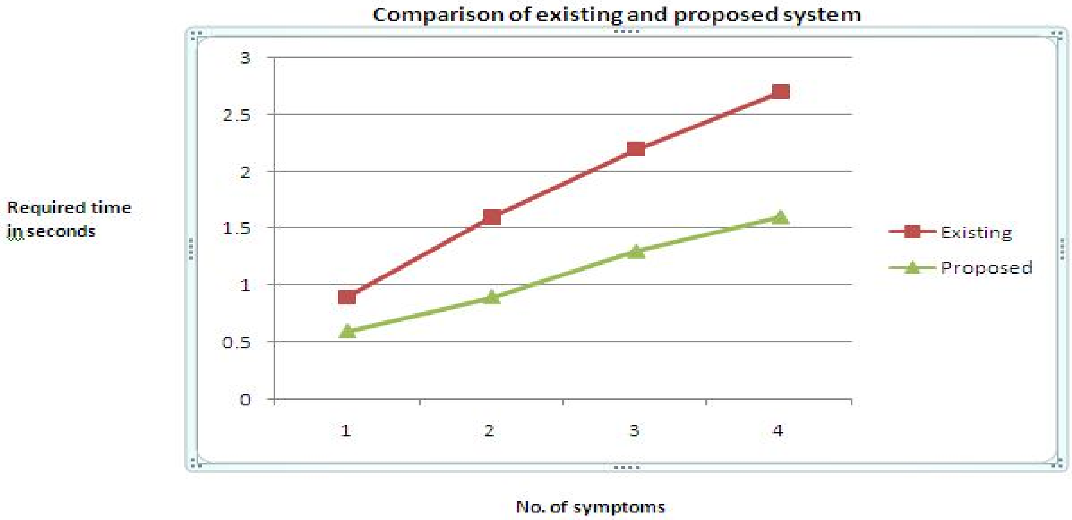

Fig.3 comparison of D-matrices

CONCLUSION

To implement fault diagnosis system even unable to grasp all the associations between failure modes and symptoms concluded into restricted support. These cons will filter these barriers where natural language processing algorithms were proposed to automatically implement the D-matrices from the unstructured repair verbatim. In propose system work, we compared the text-driven D-matrix for online database as well as on offline system, where the processing time will be reduce and efficiency will increase. In the text-driven D-matrix process we will find more fault detection, more fault isolation. Finally, the computation performance of the text-driven D-matrix when compared with online system substantiation better fault detection and fault isolation rate while presenting lower error rate on the comparison of online database and offline database.

REFERENECS

[1] Dnyanesh G. Rajpathak and Satnam Singh "An Ontology-Based Text Mining Method Develop D-Matrix from Unstructured Text" IEEE Trans. Syst., Man Cybern. A, Syst. Humans, vol. 44, no. 7,pp. 966-977,2014.

[2]]M. Schuh, J. W. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, "A Visualization tool for knowledgediscovery in maintenance event sequences," *IEEE* Aerosp. Electron. Syst. Mag., vol. 28, no. 7, pp. 30–39,Jul. 2013.

[3] 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.

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

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

[6] 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.

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

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

[9] ISO, "10303-11:2004, Industrial automation systems and integration Product data representation and exchange—Part 11: Description methods: The EXPRESS language reference manual," 2004.

[10] 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.1–2,2011.