ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A Survey on Improved Text Mining Method to Construct Unstructured text for Development of D-Matrix

Padmini Magdum, Prof. V. S. Nandedkar

M.E Student, Dept. of Computer Engineering, P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

Dept. of Computer Engineering, P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

ABSTRACT: In text mining, the process of unstructured text information, accurate numeric indices from the text and easily make the information contained in the text accessible to the different data mining approach. The text mining method developing a D-matrix by data mining a large number of update or repaire verbatim(i.e. written in unstructured text) gathered during the analysis. The fault analytic model to catch the different system level fault diagnostic information consisting of condition between recognizable symptoms and failed modes connected with a system. The system construct a D-matrix for the fault detection and diagnosis using the graph comparison algorithm and next utilize the content mining calculations that easily make utilization of this ontology based text mining to distinguish or identity the fundamental of artifacts for example, parts, symptoms, failures modes, and their conditions from the unstructured repaire verbatim text. The proposed system implements as comprehensive graph model generation for each generated D-matrix depending upon data given by the similarity graph. A comprehensive D-matrix development using text mining method which store unstructured information obtained during fault recognizing and fault solving.

KEYWORDS: Data mining, Information retrieval, Unstructured data, D-matrix, ontology, Fault analysis, Fault diagnosis, Graph Comparison Algorithm.

I. INTRODUCTION

A Dependency matrix (D-matrix) is a systematic and proper way to captured system-level fault diagnostic information. The D-matrix is derived from a dependency systematic modeling framework structure to capture the relationships between failure modes and data [1]. The proper and systematic diagnostic framework is known as a Dependency matrix or D-matrix. D-matrix are established from different types of sources like that historical field failure data, documents of service, engineering drawing, and Failure Modes, different Effects and Criticality Analysis (FMECA) data. Here, we review the existing research work on establishing D-matrices from different data sources and data formats. The D-matrices depends on their data source and the imperfectness of symptoms for both boolean and real-valued D-matrices. An industrial perspective is announced to explaning the advantages and disadvantages of different types of D-matrices along with the challenges faced while establishing and applying them for vehicle health management. The relationships among failure data and symptoms should be interpreted as causal i.e., failure modes "cause" symptoms. Also there could be causal relationships among single failure modes and single symptoms. There are relationship between multiple failure and multiple symptoms i.e., many failure modes may be causing a single symptom and vice versa. In that situations determine the root cause becomes a complex problem [2]. However, using the D-matrix, one can develop intelligent diagnostic reasoners to easily solve the root cause problem.

Ontology learning systems based on words. Before keywords were identified from the text. The identified words are typically single-word terms and they are considered as the concepts. Then, by combining these keywords possible multi-word terms are formed. As a result, the multi-word terms generated were not natural and only single-word terms were formed from most of the extracted concepts. So while processing documents using the NLP component most noun terms were found in the text was multi-word term. As it was also shown in text that 85% of the terms were multi-word terms, so traditional systems focusing on single-word term extraction will thus miss many concepts. The D-matrix is mostly used in the complex systems such as aircraft, spacecraft, nuclear power plant, etc. There are commercial tools which employ D matrix based models for system-level diagnosis [3]. The diagnostic inference model as a standard

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

based on the D-matrix in as provided a theoretical analysis of the Dmatrix and proved that a boolean D-matrix with single fault assumption is linearly separable [3]. There are several benefits of using a D-matrix-based framework for on-board and off-board fault diagnosis. The D-matrix is a framework to capture causal dependencies (*dij*) among failure modes (*fi*) and symptoms (*sj*). The dependencies could be represented as a boolean *i.e.*, 1 (related) or 0 (not related) or a realnumber in [0,1] which indicates the probability of detecting a specific failure mode. The off board implementation employs a complete (full-order) D-matrix and relies on data collection either remotely via telematics or directly by a technician in the service bay and is then transferred to a remote center for the processing. The remote decision support center processes the vehicle health data and fault codes via an advanced reasoner for the root cause analysis. The off-board D-matrix involves data rich decision making and planning. Due to the memory constraints, the on-board D-matrix is a reduced-order and compressed version. It is installed either on a diagnostic electronic control unit (DECU) or existing ECU which collects the vehicle information such as fault codes and sensor values and processes the same through a reasoner to detect faults and estimate severity of faults. If the fault's severity is high, the customer may need to be notified of need for quick repair at a nearby dealer via a telematics service. Figure 1 illustrates the off-board and on-board benefits of the D-matrix along with the need to integrate the D-matrix development into the vehicle development process. Another advantage of the D-matrix-based structure is that the D-matrix could be utilized to access a given system from a testability and diagnosability point of view [4] during the design stage *i.e.*, built-in diagnostics [8]. There are several ways to established a D-matrix such as manual or automated text processing of service procedures, modeling using engineering documents and domain knowledge and data mining of field failure data (case-based approach). Ontology of text mining is the study of the nature of being, becoming, reality and existence, and also of the basic categories of being and their relations.

II. LITERATURE SURVEY

In this paper different methods are discussed. Those methods have some advantages and disadvantages.

An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text by Prof. Dnyanesh G. Rajpathaket. al [1] have proposed to determine 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. They compared the testability and diagnosability metrics of the historical data-driven D-matrix and the text-driven D-matrix. They have also proposed naïve 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]. The performance of the text-driven D-matrix when compared with LDA demonstrated improved fault detection and fault isolation rate while exhibiting lower error rate.

Ontology-based Knowledge Discovery from Unstructured Text by prof. JantimaPolpinij [2] has introducing Knowledge discovery database KDD process for finding knowledge from unstructured textual data is major problem in area of knowledge discovery in database. To solve this problem they have present a unified methodology is called ontology based knowledge discovery in unstructured text (ON-KDT) methodology, to discover knowledge from unstructured text. The plan should be as detailed as possible that have step-by-step to perform during project including initial selection of relevant techniques and tools. ON-KDT process model deals with two main manual tasks. The first one is to determine the application domain. It includes defining of the problem and the goals of end-user. And second is to collect an initial data and proceed with activities in order to make more understanding target data and identify data quality problems.

Graph based ontology-guided data mining for d matrix model maturation by prof. Shane Strasser [5] have proposed the alarm dependency maturation algorithm would be able to determine the missing link in the faulty TFPG model. In addition, the subtraction of a link was only for a specific test case. More experimentation is needed in which a large variety of links is added or deleted to fully test whether our algorithm can find all missing or added links.

Trends in the development of system-level fault dependency matrices by Prof. S. Singh, S. W. Holland, and P. Bandyopadhyay [6] proposed the D-matrix-based fault modeling for system-level diagnosis which is widespread in complex systems such as aircraft, spacecraft, etc. D-matrix provides only theoretical assessments of the D-matrix. However, they share little on approaches to develop the D-matrix in the first place. Here, discuss several ways to develop the D-matrix from different data sources and disparate data formats. Based on their data sources, we termed

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

three types of D-matrices: **E**ngineering **D**-matri**x** (EDx), **D**ocuments **D**-matri**x** (DDx), and **H**istorical data **D**-matri**x** (HDx). We offer an industrial perspective on developing these matrices.

III. PROPOSED SYSTEM

A. *Proposed Work:*

The proposed system implements as comprehensive graph model generation for each generated D-matrix depending upon data given by the similarity graph. A comprehensive D-matrix development using text mining method which store unstructured information obtained during fault recognizing and fault solving.

B. System Architecture:

As shown in the below fig. the system will work as follows:

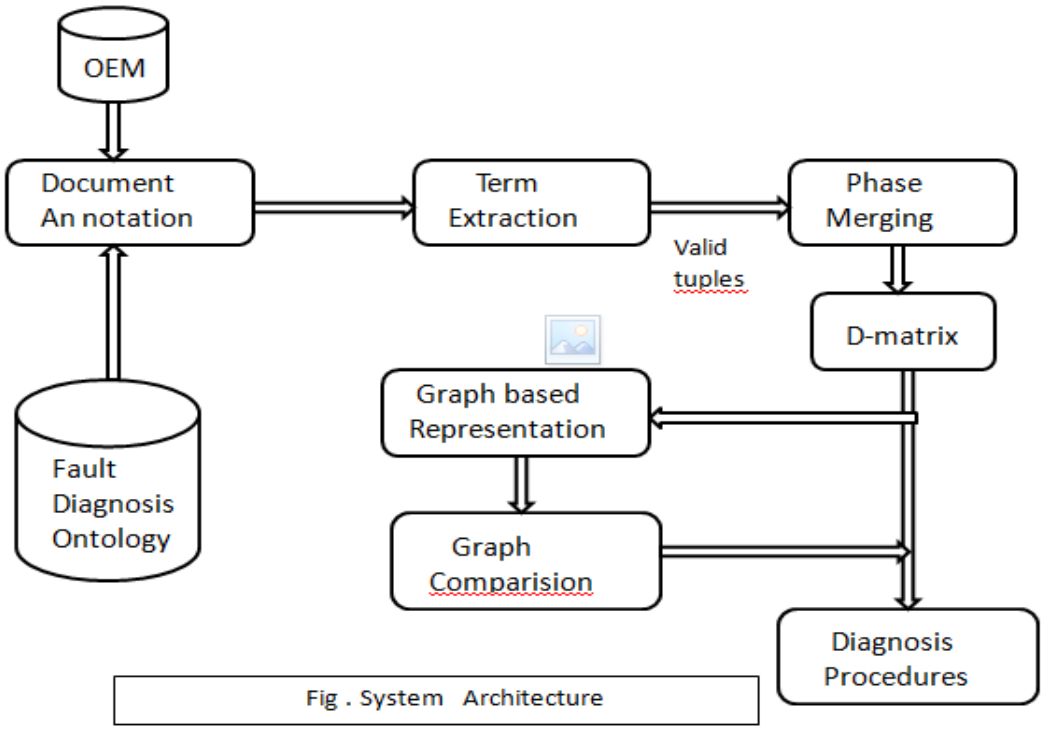

Fig . System Architecture

The system architecture as shown in fig. the system architecture [1] creates the D-Matrix for one dataset. It provides accurate d-matrix and graph is generated from the matrix. So that, every time, the new d-matrix is created for the dataset. Even if the different datasets contains some similar data, the new D-Matrix is developed for each dataset.
Our proposed system provides contribution to the existing system. The system architecture consists of document annotation, term extraction, term extraction, phase merging, D-matrix, diagnosis procedures, fault diagnosis ontology and for comparison for graph representation. The proposed system implements as comprehensive graph model generation for each generated D-matrix depending upon data given by the similarity graph. A comprehensive D-matrix development using text mining method which store unstructured information obtained during fault recognizing and fault solving.

C. Algorithm:

System S is represented as S = {DB, DA, TE, PM, U}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A]. Database
DB= {d1, d2, d3, ...,dn}
Where, D is the set of documents annotation and d1, d2, d3, ...,dn are the number of documents.

B]. Fault diagnosis ontology
FD={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.

C]. Document Annotation
DA= {P, C}
Where P= {W, E}
Where, P is represent as a set of preprocessing and
W= {w1, w2, w3, ...,wn}
Where W is represent as a set of Stop Words and w1, w2, w3, ...wn number of stop words and
E= {e1, e2, e3, ...,en}
Where E is represent as a set of steaming words and e1, e2, e3, ...en is a number of steaming word.
C= {c1,c2,..,cn}
Where C is the set of corpus annotated document and c1, c2, c3,……..,cn represent as a number of corpus annotated document.

D]. Term Extraction
TE= {F, V}
Here TE is a set term extraction and
Where F is represent as a set of feature terms and f1, f2, f3, ...,fn 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 = {X, R}
Where PM is represent as a set of Phrase Merging
X={x1, x2, x3, ..,xn }
Where X is represent as a set of context information and x1, x2, x3, ...xn is a number of context information.
R={r1, r2, r3, ...rn }
Where R is represent as a set of merged phrases and r1, r2, r3, ...rn is a number of merged phrases.

F]. Graph Comparison Algorithm
U= {u1, u2, u3, ...un }
Where U is the set of D-Matrix graphs and u1, u2, u3, ...,on represent as a number of D-Matrix graphs.
S= {s1,s2,s3….sn}
Where S is the set of similarities in the D-matrix graphs and s1,s2,s3,….,sn is the number of similarities in D-matrix graphs.

IV. CONCLUSION

Ontology based text mining to develop D-Matrix where regular natural language transforming algorithm were proposed to consequently create the D-matrix from the unstructured repair verbatim and develop d-matrices from different datasets and then represent each d-matrix in graph and by using graph merging algorithm, and combine common patterns of generated graph into new graph. This newly generated graph is then used to construct D-matrix which is comprehensive and combination of two d-matrixes. In future, our aim is to reuse this newly generated D-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

matrix whenever the relevant dataset will come to generate the D-matrix. So, this will save the time and useful in many manners.

REFERENCES

[1] Dnyanesh G. Rajpathak, et.al, "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] JantimaPolpinij , "Ontology-based Knowledge Discovery from Unstructured Text" International Journal of Information Processing and Management(IJIPM) Volume4, Number4, June 2013.
[3] Vishwadeepak Singh et. al, "Text Mining Approaches To Extracts Interesting Association Rules From Text documents" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012.
[4] SandeepSirsatet. al "Mining Knowledge From Text Repositories using Information Extraction" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 53–62. _c Indian Academy of Sciences.
[5] 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 ., 2011, pp. 1–12..
[6] S. Singh, S. W. Holland, and P. Bandyopadhyay, "Trends in the development of system-level fault dependency matrices," in Proc. IEEE Aerosp. Conf ., 2010, pp. 1–9.
[7] 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..
[8] O. Benedittini, T. S. Baines, H. W. Lightfoot, and R. M. Greenough,"State-of-the-art in integrated vehicle health management," J. Aer. Eng.,vol. 223, no. 2, pp. 157–170, 2009..

BIOGRAPHY

Padmini Magdum Student of ME Computer Engineering second year from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune, India.

Prof. V. S. Nandedkar is a faculty in the Computer Engineering from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune, India.