ISSN:0975-9646

An Epistemological Text Mining Method by Generating D-Matrix for Unstructured Text

Radhika Y. Deore
Student, Computer Engg. Department, MCOERC Nashik
Pune University, India

***Abstract*— Fault dependency (D)-matrix is a systematic diagnostic model to capture the hierarchical system-level fault diagnostic information.it consisting of dependencies between observable symptoms and failure modes associated with a system. whenever user type any query for searching any file or data, most probable all the files or data trying to match its search query with title of available data and Constructing a D-matrix from first principles and updating it using the domain knowledge 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. Proposed system describes an ontology based text mining method for automatically constructing and updating a D-matrix by mining hundreds of thousands of repair verbatim (typically written in unstructured text) collected during the diagnosis episodes. In proposed approach, firstly construct the fault diagnosis ontology consisting of concepts and relationships commonly observed in the fault diagnosis domain. Next, employ the text mining algorithms that 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.**

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

I. INTRODUCTION

With the rapid growth of the World Wide Web and electronic information services, digital information is increasing at an incredible rate, causing the unprecedented problem of information overload. No one has time to read everything, yet we often have to make critical decisions based on what we are able to assimilate. Thus effective management of electronic documents, especially management of complexity and specialization of knowledge expressed in those text documents, is essential to enterprise knowledge management. A complex system interacts with its surrounding to execute a set of tasks by maintaining its performance within an acceptable range of tolerances. One challenge that managers face is how to construct deep knowledge from a collection of documents to support problem solving. How can we use information technology to gain insights or to extract useful knowledge about this phenomenon from those documents so that we can handle it better in the future or prevent it. The fault detection and diagnosis (FDD) is performed to detect the faults and diagnose the root-causes to minimize the downtime of a system the process of FDD becomes a challenging activity in the event of component or system malfunction. Not surprisingly, after every diagnosis episode the lessons learn are maintained in several databases to detect and diagnose the faults. Big amount of information is available in textual form in databases and online sources. In this context, manual analysis and effective extraction of useful information are not possible it is relevant to provide automatic tools for analyzing large textual collections Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. The task of data mining is to automatically classify documents into predefined classes based on their content. Hundreds of thousands of such repair verbatim are collected and we argue that there is an urgent need to mine this data to improve fault diagnosis (FD). However, the overwhelming size of the repair verbatim data restricts an ability of its effective utilization in the process of FD. Automatically discover the knowledge assets buried in unstructured text. A text mining method to map the diagnostic information extracted from the unstructured repair verbatim in a D-matrix.

II. LITERATURE SURVEY

A. Visualization tool for knowledge discovery in maintenance event sequence:

In this approach to Faults are remove and provide ontology-guided data mining and data transformation But Discovery is loss because result is not in form of matrix [3].

B. frequent Item sets Mining Algorithm with Tags:

It is transaction reduction for finding for finding item sets based on tags and shows result in matrix [1] it does not give accurate result. Its search is only based on tags.

C. Ontology extraction for knowledge reuse:

It provide an easy to use interface that generates relevant sequences of data in meaningful context and retrieve and display similar information Only shows similar information not accurate result in this form like D-MATRIX.

D. Intelligent data warehouse mining:

It builds useful data mining models and it present prototype multidimensional mining system mining hundreds of thousands of repair verbatim (typically written in unstructured text).

III. EXISTING SYSTEM

In Existing work it described various ways to construct D-matrices by using the data sources, such as service

procedures and engineering design The associations between the failure modes and symptoms are mapped in a D-matrix by using the signal flow diagrams and engineering knowledge of a system, such as a failure mode, effects, criticality analysis data, signal data, and engine control unit data and viewing the overall data that is saving all database and frstly parse that data and after that scan overall data so its takes more data base memory and its very much time consuming for parsing and scanning that overall databases. However, the development of D-matrix from scratch takes significant engineering effort and time, while the data mining techniques have shown to save the construction time of D-matrix from the field fail-lure data. The fidelity of the data-driven D-matrix is lower due in part to the noise in the field failure data, whereas the service procedures- based D-matrix is of higher fidelity, but of lower fidelity when compared to the engineering design-based D-matrix. The data driven framework detects anomalies by using the system level fault model and diagnostic reasoned built by mining the operating sensory parameter identifiers data.

The System performed fault detection and diagnosis (FDD) to detect the faults and diagnose the root-causes to minimize the down-time of a system. It downloaded Whole html-pages so it requires un-wanted databases. Also html-like tags and non-textual information like images, commercials, etc are cleaned from the downloaded text so its time consuming task. Also separate processing of Phrase merging.

The Existing Text Mining Process:

- Whole html-pages are downloaded from a given forum site.
- Html-like tags and non-textual information like images, commercials, etc are cleaned from the downloaded text.
- The textual parts are divided into informative units like threads, messages, and sentences.
- Products and product attributes are extracted from the messages.
- Generate D-matrix only for one result.

A. Manual System

Comparisons are made either by using co-occurrence analysis or by utilizing learned comparison patterns The selection of research proposals in existing is done manually means the proposals are submitted to funding urgency and according to the name of research proposals or paper and the keywords the research proposals are classified into groups or domain this done manually means by the human. Following block diagram shows the process of manual clustering of research proposals.

But this is not suitable for large data. It makes misplacement of research proposals due to manual process and classification according only the name of research proposals. So this misplacement makes the reviewers or experts more confuse of the research proposals which are not from their area of research. There exists the software which also cannot handle the large data and misplacement of research proposals.

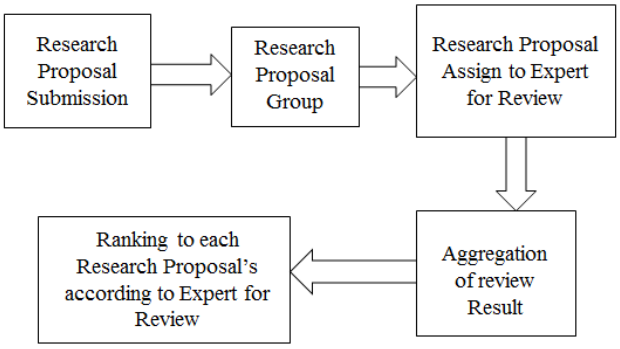

Fig No 1: Existing Manual System

B. Automatic Existing System

The Existing system performed fault detection and diagnosis (FDD) to detect the faults and diagnose the root-causes to minimize the down-time of a system. It downloaded Whole html-pages so it requires unwanted databases. Also html-like tags and non-textual information like images, commercials, etc are cleaned from the downloaded text so its time consuming task. Also separate processing of Phrase merging.

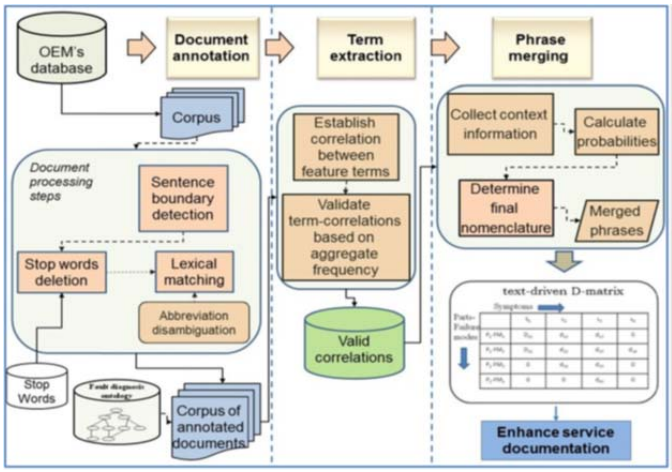

Fig No 2: Text-driven D-matrix development methodology from unstructured text data.

IV. Proposed System

The proposed system is based on the Epistemology that is form of the Ontology in text mining. It forms three phases to process. That consist Document Annotation and combined The Term Extraction and Phrase Merging for Optimization of time. We propose a text mining method to map the diagnostic information extracted from the unstructured repair verbatim in a D-matrix. However, the construction of a D-matrix by using text mining is a challenging task partly due to the noises observed in the repair verbatim text data abbreviated text entries the abbreviation are used to record the terms and it is crucial to disambiguate their meaning, incomplete text entries the incomplete repair information makes it difficult to derive the precise knowledge from the data; term disambiguation the same term is written by using inconsistent vocabulary. Typically 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 are collected, such as error codes, scanned values of operating parameters associated with faulty component/system, repair verbatim, and so on. 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. Such models can be used to perform accurate FDD. The D-matrix captures component and system level dependencies between a single or multiple failure modes1 (or root-cause of failures) with a single and multiple symptoms (a set of fault codes, observed symptoms, etc.) 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. Also, 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 of a specific failure mode using a specific symptom. The values between 0 and 1 indicate the level of strength of detecting a failure mode by using a symptom. In particular, our work falls into the quantitative and data-driven fault diagnosis categories, whereby a text driven D-matrix development methodology is proposed where initially the fault diagnosis ontology is constructed by mining the unstructured repair verbatim data. Subsequently, the text mining algorithms are developed, which uses this ontology to discover the dependencies between the symptoms and the failure modes. The qualified associations are used to construct the D-matrix diagnostic model.

V. System Module

The proposed system will use the same Mathematical programming model but with optimized by combined the two phases Term Extraction and Phrase merging so optimize the more time for processing. Validated correlated result instantly gives for calculation probabilities nothing but work like multithreaded.

- Creating a vocabulary for ontology is to extract important terms from text documents related to a particular domain.
- The corpus is then parsed into tokens or terms.
- Unstructured text in the corpus becomes a structured data object via the creation of a term-by-document frequency matrix.
- Frequency weights of those concepts can be adjusted to account for the distribution of terms across documents.
- Natural language processing (NLP) and text mining techniques are effective for information extraction from text documents.
- Basically, Final Result in D-matrix which uses for Comparison between two or more results which is not in existing system.

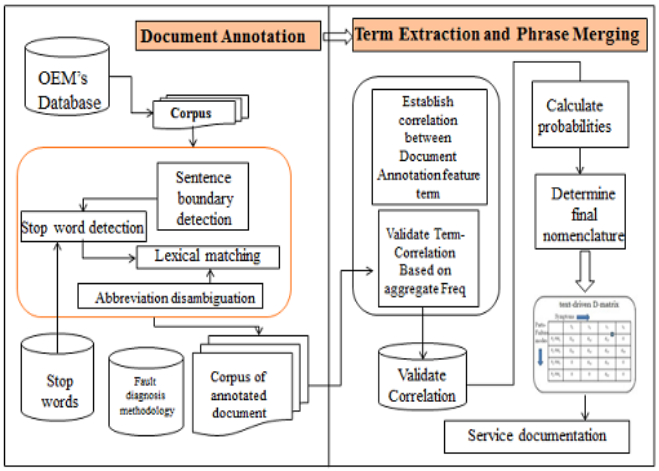

Figure No 3: Proposed Text-driven D-matrix development methodology

VI. Conclusions

It is concluded to construct the D-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis.

Text Data Mining or Knowledge-Discovery in Text refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text

The existing techniques require more data for training as well as the computational time of these techniques is also large. In contrast to the existing algorithms, the proposed hybrid algorithm requires less training data and less computational time.

References

[1] Harpreetsingh and renuDhir"A New efficient Matrix based frequent Item sets Mining Algorithm with Tags" August,2013

[2] S. Strasser, J. Sheppard, M. Schuh, R. Angryk, and C. Izurieta (2011). Graphbased ontology-guided data mining for d-matrix model
maturation. in Proc. IEEE Aerosp, 21, (pp. 1f12).

[3] M. Schuh, J. W. Sheppard, S. Strasser, R. Angryk, and C. Izuri-eta (2013).A Visualization tool for knowledge discovery in maintenance event sequences, July.

[4] Dnyanesh G. Rajpathak, Member, IEEE and Satnam Singh An Ontology-Based Text Mining Method Develop D-Matrix from Unstructured TextIEEE Trans. Syst., Man Cybern. A, Syst. Humans, VOL.44, NO. 7, 2014 ,pp. 966-977.

[5] Santosh Kumar Paul, Madhup Agrawal, Shyam Rajput, 3Sanjeev Kumar (2014). An Information Retrieval(IR) Techniques for text Mining on web for Unstructured data(pp. 68f77).

[6] M. Gaeta, F. Orciuoli, S. Paolozzi, and S. Salerno (2012). Ontology extraction for knowledge reuse: The e-learning perspective(pp.111116).

[7] S. A. Chatzichristo_s, K. Zagoris, Y. S. Boutalis, and N. Papa-markos (2010). Accurate image retrieval based on compact composite descriptors and relevance feedback information, Int. J. Pattern Recog.Artif.Intell.(pp. 207244).