A Survey on Mining of Unstructured Text for Development of D-Matrix

Swati S Hinge¹, B. R. Nandwalkar²

¹ M.E Student, Kalyani Charitable Trust’s Late G.N. Sapkal College of Engineering, Nashik
² Assistant Professor, Kalyani Charitable Trust’s Late G.N. Sapkal College of Engineering, Nashik

swatih110@gmail.com, nandwalkar.bhushan@gmail.com

Savitribai Phule Pune University

Abstract

This survey paper aims to provide the multimedia researcher with the different fusion approaches. The different fusion approaches are used to have multiple modalities are used to accomplish various multimedia related analysis tasks. Fusion method are describe by using different perspective. The problem of metadata based image retrieval system and a very rapid growth in the quantity and availability of digital image involves the researches into the multimedia information system. Finally in last section we are discussing on the multimedia information retrieval based on late semantic fusion approaches.

Index Terms — Late Fusion, Multimedia Information Retrieval, Content based retrieval, Text based retrieval.

I. Introduction

To minimize the downtime of a system the fault detection and diagnosis (FDD) is performed to detect the faults and diagnose the root-causes.¹ Recent overwhelming abundance of data can generate modern diagnostic systems. This data is distributed across multiple heterogeneous systems and cannot be or easily collected and aggregated for use. Even with easy access to data, much of it is uninteresting to the user’s specific inquiry. This puts a large burden on the user to coalesce the data and mine the interesting bits relevant to their current needs.² This complexity may result in failures of multiple components. Hence, there is a need to develop smart on-board diagnostic algorithms that can determine the most likely set of failure causes in a system, given observed test outcomes over time.³ In this system, we need to propose a text mining method to map the diagnostic information extracted from the unstructured repair verbatim in a D-matrix. The D-matrix is one of the standard diagnostic models specified in IEEE Standard 1232. 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, incomplete text entries, and term disambiguation. Typically the process of can FD starts by extracting the error codes from a target system (Fig. 1) and based on the observed error codes the technicians follow specific diagnosis procedure can diagnose the faults along with their practical knowledge. Various data types are collected during fault diagnosis, such as error codes, scanned values of operating parameters associated with faulty component/system, repair verbatim, and many more. The accumulated data can then transferred to the particular database and particularly the repair verbatim data collected over a period of time can be mined to develop the D-matrix diagnostic models.

In particular, our work can fall into the quantitative and data-driven fault diagnosis classifies, where a text-driven D-matrix development methodology on online database can propose where initially the fault diagnosis ontology is constructed by mining the unstructured repair unstructured data.

issn:2229-6093

IJCTA | Nov-Dec 2014
Available online@www.ijcta.com

1968
After that, to discover the dependencies between the symptoms and the failure modes by the developed text mining algorithms can help, which uses this ontology. The qualified relations can used to develop the D-matrix diagnostic model. A text-driven D-matrix as a diagnostic model is the output of our process. In section II, we discuss related work, Proposed System in section III, in section IV, present the conclusion.

II. Related Work

In 1999, for previous system the first problem it was unable to help users in finding out useful information on the online web pages and in locating knowledge about a particular domain that was represented by a bunch of Web-documents. The second problem, it wasn’t to optimize the system or not finding information about the clients using the system, it was not analysing the operations run in a Web-based system. The Marti A. Hearst in [5] proposed a Text Data Mining (TDM) method data mining was to discover new information from data, finding patterns across datasets, and distinguishing signal from noise. The information retrieval system could return a document that was consist of the information a user requested implies that no new discovery was being made. In LINDI project the main tools for locating new information were of two types: maintain for issuing series of queries and related operations across text accumulations, and tightly paired statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents. Make use of attributes associated specifically with text collections both sets of tools and their metadata.

In 2003, [6] the development of successful application there was various latent model element for discrete data. In this method they report a new model for gathered of discrete data that provided full generative probabilistic semantics for documents. Documents are modelled via a hidden Dirichlet random variable that specifies a probability distribution on a latent low measurement topic space. The distribution over words of an unseen document was a continuous mixture over document area and a discrete mixture over the document area and a discrete mixture over all possible subjects. The generative nature of LDA makes it was easy to utilize as a module in more complex architectures and expand it in different directions. If the categorization variable of LDA is treated as a latent variable they obtain a mixture of LDA models, a useful model for situation in the document bunch not only according to their topic overlap, but along the proportion as well.

In 2004, Resulted data retrieval problem was solved but modern diagnostic systems could create a prodigious abundance of data. Often this data was distributed to the other side of multiple heterogeneous systems and could not be easily collected and aggregated for manipulate. Even, there can be an extensive amount of automatically logged information with access to data and interesting content sparsely intermixed. Therefore, a large load was put on the user to combine the data and mine the interesting bits relevant to their current requirements. Depending on the labour at hand, this amount of effort may not be valid or practical, and the potential for knowledge finding was lost. So in [7] Michael Schuh, John Sheppard, Shane Strasser, Rafal Angryk, and Clemente Izurieta proposed OWL ontologies situated on formally defined IEEE standards, and manipulate these ontologies to guide the data mining and data renew processes. Their application detached much of the user’s load for data look-up and greatly grows the potential for knowledge discovery from data (KDD) in this domain. Authors produce an easy-to-use interface that causes relevant sequences of data in meaningful circumstances in a fraction of the time it was take domain experts for retrieve and exhibition of similar information. Transforming data records into ontology-based event graphs, and providing a filterable visualization of event sequences through time were improved knowledge discovery within maintenance data.

In 2009, Unstructured data was collected efficiently but the problem was in safety critical systems, such as automobiles, aircraft, space vehicles and nuclear power plants are becoming unusually more interconnected and complex. The recent move
forwarded in wireless technology, remote communication, computational capabilities, standardized hardware/software interfaces and sensor technology had increased the complexity problem of these systems. This complexity may result in failures of multiple modules. Hence, there was a need to construct smart on-board diagnostic algorithms that can decide the most likely set of failure causes in a system, given observed test result over time. In [8], the Satnam Singh, Kihoon Choi, Anuradha Kodali and Krishna Pattipati proposed a solution scheme that could be viewed as a two-level correlated solution framework for the DMFD problem. At the top (correlation) level, they update the Lagrange multipliers (correlation variables, dual variables) by using the subgradient method. The top level facilitates correlation between each of the subproblems, and could thus reside in a vehicle-level diagnostic control unit. At the bottom level, they used a dynamic programming approach to solve each of the subproblems. The key merit of their approach was that it provided an estimated duality gap, which was a determine of suboptimality of the DMFD solution. Interestingly, the perfectly-observed DMFD problem guided to a dynamic set covering problem, which could be approximately resolved via Lagrangian relaxation and Viterbi decoding. All the DMFD formulations are NP-hard. The easiest one, i.e. the deterministic formulation (Problem 4) is also NP-hard.

In [1] Dnyanesh G. Rajpathak, Satnam Singh, 2014 they first developed the fault diagnosis ontology comprising of concepts and relationships commonly observed in the fault diagnosis domain. Next, employ the text mining algorithms that construct utilize of this ontology to recognize the essential artifacts, such as parts, symptoms, failure modes, and their relations from the unstructured repair verbatim data. The method was implemented as a original tool and validated by using real-life data gathered from the automobile domain. They validate how the proposed method has been used to develop the D-matrices using real-life data communicated to the automobile evaporative emission (EVAP) control system. The method described in [1] was applied to develop a text-driven D-matrix and it was observed that our method successfully removed the failure modes and symptoms along with their relationship from all the related systems to construct the D-matrix. The LDA relations revealed that all the failure modes did not have a like probability and a high probability was given to all such symptoms illustrate failure modes corresponding to the similar system, whereas a little probability was allocated to the symptoms coming from the contrast systems. In fact, it was essential to trace the symptoms and failure modes coming from all the associated systems for accurate FD. They feel that there was incorporating the domain specific relations while constructing the matrices to further enhance the performance with a need to provide an addition to the LDA model by.

In the existing methods of fault modelling [3], [6], [9], [10] the limited work done we observed that by analysing unstructured repair verbatim data to develop a D-matrix. Only recently [2] the tool is proposed that locates the knowledge by collecting relevant sequences from the on-board diagnosis and conservation data by using the ontology-based data mining on the online database. Our system may improve the performance of our system when differentiate with the Latent-Dirichlet Allocation (LDA) technique. Traditionally, the D-matrices were developed by using the sensory data, engineering data, and history data, for example,[3], [6], [7], [11], [12] but a very small insight is supplied about the discovery of recently 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. Analysis of Our Work

The proposed structure mainly focuses on following areas:
Module 1:- Document Annotation.
Module 2:- Term Extraction.
Module 3:- Text Mining.

Module 1. Document Annotation

In module 1, first during field FD the repair verbatim data points are gathered by retrieving them from the OEM’s database are recorded. In the initial step, the terms, are annotated like part, failure mode and symptom relevant for the D-matrix are from each repair verbatim by constructing the document annotation algorithm. A repair verbatim consists of several parts, symptoms, failure modes and actions and the correct relations must be organized between the relevant terms based on their existence with each other. Here, using the sentence boundary detection rules a repair verbatim is first split in different 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.

A typical repair verbatim composed of multiple parts, symptoms, failure modes, and actions and it is necessary to identify the accurate relations between them such that only the accurate causal relations are examined for developing a D-matrix.

Module 2. Term Extraction

Module 2 takes tuples are developed from take an input annotated repair verbatim form module 1 by using the term extraction algorithm to inhabit a D-matrix. Having annotated the terms, the critical terms needed for the development of a D-matrix, i.e., symptoms and failure modes are extracted by using the term extractor algorithm. Initially, the causal association between the relevant symptom-failure mode pairs is discovered to make sure that only the accurate pairs are extracted. The existing methods for frequent item sets mining push aside the order in which the term phrases are recorded in documents, but we must preserve such ordering to understand how the fault diagnosis is executed. The annotated terms are extracted tuples are developed first by identifying the all the failure modes, and symptom. Next, the tuples are combined using parts as the common tuple member and the tuples are developed. Typically, several tuples are developed from the corpus, but only the relevant tuples must be maintained while developing a D-matrix correlating to a specific system. At the end of this step, assorted tuples are developed but all of them are not censorious to diagnose the faults observed in relations with a specific system. The clustered normalized frequency of the tuples is calculated and the tuples with their frequency above a particular threshold are kept as the reasonable tuples.

Module 3. Text Mining

Next, the text mining is used to avoid ambivalent references of the failure mode phrases. 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 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 are merged. A priori our system does not have the knowledge about which two failure modes can be merged; hence the similarity of each failure mode is calculated with every other failure mode to derive the similarity score. Finally, the newly constructed D-matrix is audited by subject matter experts (SMEs) to identify the discovery of new symptoms and failure modes for in-time FDD.

Model when solving problems we have to decide the difficulty level of our problem. There are three types of classes providing for that as follows:

- **P Class.**
- **NP-hard Class.**
- **NP-Complete Class.**

**P:** Informally the class P is the class of decision problems solvable by some algorithm within a number of steps bounded by some fixed polynomial in the length of input. Turing was not the concerned about the efficiency of the machines, but rather his concern is whether they can simulate arbitrary algorithms given appropriate time. Whenever it turns out Turing machines can generally simulate more efficient computer models (for example machines equipped with many tapes or an unbounded random access memory) by at most squaring or cubing the calculation time. Thus P is a robust class and has identical definitions over a large class of computer models. Here I follow standard exercise and define the class P in terms of Turing machines.

**NP Hard:** A problem is NP-hard if solving it in polynomial time would make it possible to solve all...
problems in class NP in polynomial time. Some NP-hard problems are also in NP (these are called "NP-complete"), while some problems are not. If you could shrink an NP problem to an NP-hard problem and then solve it in polynomial time, you would be solving all NP hard problems. And there are also some decision problems in NP-hard but these are not NP-complete, likewise the infamous halting problem. What does NP-hard mean? A lot of times you can solve a problem by reducing it to a distinct problem. I can bring down Problem B to Problem A if, specified a result to Problem A, I can easily develop a solution to Problem B. (In this case, "easily" means "in polynomial time.")

Such That $S = \{I, F, O\}$ Where,

**I represent the set of inputs**

In the input set I system takes an input D is a number of documents and words for offline as well as online working.

$I = \{D, W\}$

$D =$ Set of documents.

$W =$ Set of words.

**F is the set of functions**

In this set initially, token can be generate i.e. $T$, then calculate the frequency of the parts, symptoms and failure modes i.e. $F$ and after that it generates a D-matrix i.e. $M$.

$F = \{T, F, M\}$

$T =$ generate token.

$F =$ calculate frequency.

$M =$ generate D-Matrix.

**O is the set of outputs**

Finally after processing output will be in the form of cluster $C$ which in a structured manner.

$O = \{C\}$

$C =$ form clusters

The previous technique was applied on offline database only but this method can be efficient for not only offline database but also online database too. It also compares both dataset

**IV. Conclusion**

From a study of existing system we can conclude, in this system, ontology-based text mining technique has been proposed to develop the D-matrices by collecting data during fault diagnosis automatically mining the unstructured repair verbatim data. In previous approach, D-matrix diagnostic model for the manual construction corresponding to the complex systems is impractical as it would involve effort to integrate the knowledge and represent it in a D-matrix. In many systems, to execute fault diagnosis not even be able to grasp all the relations between failure modes.
and symptoms resulting into restricted support. This disadvantage will overcome these barriers where natural language processing algorithms were proposed to automatically develop the D-matrices from the unstructured repair verbatim. In this system, we compared the data-driven D-matrix and the text-driven D-matrix with its analysis and identifiable metrics on online database as well as on offline database, where the text-driven D-matrix method we will get more fault detection, more fault isolation, and less ambivalence group size due to textual symptoms and the equivalent failure modes included in the text-driven D-matrix. Finally, the computation performance of the text-driven D-matrix when compared with LDA substantiation better fault detection and fault isolation rate while presenting lower error rate on the comparison of online database and offline database.

References


