International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

An Ontology-Based Text Mining Method to Develop D-Matrix

Poonam Jagdale[1], Devendra P Gadekar[2]

[1]Pune University, Pune, Maharashtra, India

[2]Assistannt Professor, Pune University, Pune, Maharashtra, India

Abstract: ***In this point, we demonstrate an ontology based text mining method for naturally developing and updating a D-matrix by mining a huge number of repair verbatim (written in unstructured text) gathered during the analysis scenes. Fault dependency (D)-matrix is a systematic demonstrative model [8] which is used to catch the progressive system level deficiency symptomatic data comprising of dependencies between observable symptoms and failure modes connected with a framework. D- Matrix is a time consuming process. Developing a D-matrix from first standards and updating it utilizing the domain information is a work concentrated. Further, in-time increase of D-matrix through the disclosure of new symptoms and failure modes watched for the first run is a challenging task. In this methodology, we first develop the fault diagnosis ontology comprising of concepts and relationships regularly saw in the fault diagnosis domain. Next, we utilize the text mining algorithm that make utilization of this ontology to distinguish the fundamental artifacts, for example, parts, symptoms, failures modes, and their conditions from the unstructured repair verbatim text. The proposed technique is implements as a prototype tool and accepted by utilizing real - life information gathered from the automobiles space.***

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

1. Introduction

A complex system collaborates with its surrounding to execute a set of assignments by keeping up its performance inside a acceptable range of tolerances. Any deviation of a framework from its worthy execution is dealt with as a fault [1]. The fault detection and diagnosis (FDD) is performed to distinguish the faults and diagnose the root-causes to minimize the downtime of a system. Because of constantly becoming technological complexity that is inserted in the vehicle system, for case complex programming installed system [2], diagnostic sensors, web, and so on the methodology of FDD gets to be a challenging activity in the occasion of part or system malfunction. As anyone might expect, after every judgment scene the lessons learnt are kept up in a few databases (e.g., the error codes are put away in on-board PCs of aircraft) to detect and diagnose the faults. One generally book kept diagnosis information comes as unstructured repair verbatim (additionally alluded to as patient medicinal records in medical industry), that gives a rich wellspring of diagnostic data. It comprises of indications relating to the faulty parts, the watched disappointment modes, and the repair moves made to fix faults. More such repair verbatim is gathered. However, we contend that there is a pressing need to mine this information to, enhance fault diagnosis (FD). Nonetheless, the staggering size of the repair verbatim information limits a capacity of its powerful use during the time spent FD. Text mining [3] is picking up a genuine consideration because of its capacity to naturally find the knowledge assets covered in unstructured text. In this paper, we propose a text mining strategy to guide the analytic information separated from the unstructured repair verbatim in a D-matrix [4]. The D-matrix is one of the standard analytic models determined in IEEE Standard 1232 [5]. Be that as it may, the development of a D-matrix by utilizing text mining is a challenging task partly because of the noises observed in the repair verbatim text information. The abbreviated text entries: it is used to record the terms and it is essential to disambiguate their meaning, for example, loose fr door chked repair performed. Incomplete text entries: the incomplete repair information makes it hard to determine the exact learning from the data. Term disambiguation: the same term is composed by utilizing inconsistent vocabulary, e.g. FTPS−Inop and FTPS−Internal Short. A principled methodology is proposed to build up a D-matrix analytic model by dissecting the unstructured repair verbatim information connected with the various systems in parallel through the advancement of ontology-based text mining algorithms. It defeats the constraint confronted in the genuine business of needing to build the D-matrix diagnostic models utilizing first principle. Further, in our methodology we have the capacity to catch the cross-system dependencies, which made a difference to essentially improve the execution of FDD. Generally, the D-matrix built by utilizing the history information, engineering information, and sensory information, for sample, [6]. However a practically nothing understanding is given about the disclosure of new symptoms furthermore, failure modes watched first time and their incorporation in the D-matrix models. In our methodology the occasional growth of the deficiency finding cosmology helps the content mining calculation to develop the right D-matrix.

2. Literature Review

In paper [22], this paper near-optimal algorithm for dynamic multiple fault diagnosis (DMFD) issues in the vicinity of in perfect test results. The DMFD issue is to focus the undoubtedly advancement of component states, the particular case that best clarifies the observed test results. Here, we examine four details of the DMFD problem. These incorporate the deterministic circumstance relating to impeccably observed coupled Markov decision processes to a few part of the way watched factorial covered up Markov

models ranging from the situation where the blemished test results are elements of tests just to the situation where the test results are elements of shortcomings and tests, and additionally the case where the false alerts are connected with the ostensible (fault free) case just. All these definitions are unmanageable NP-hard combinatorial optimization problems.

In paper [23], principal component analysis (PCA) has discovered wide application in process monitoring, slow and normal process frequently happen in real methods. In this paper, we propose two recursive PCA algorithms for adaptive process monitoring. The paper begins with an efficient way to deal with overhauling the relationship matrix recursively. The calculations, utilizing rank-one modification and Lanczos tridiagonalization, are then proposed. The quantities of principal components and as far as possible for process monitoring are likewise decided recursively. A complete adaptive monitoring algorithm that addresses the issues of missing values and outlines is displayed. At last, the proposed calculations are connected to a rapid thermal annealing process in semiconductor processing for adaptive monitoring.

In paper [24], in this paper, we lead a study on content clustering utilizing regular item sets. The principle commitment of this paper is three manifolds. First, we exhibit an review on existing methods for record clustering utilizing frequent pattern. Second, another strategy called maximum capturing is proposed for document clustering. Third, tests are completed to assess the proposed technique in comparison with CFWS, CMS, FTC and FIHC methods. Moreover, topics produced by Maximum Capturing distinguished clusters from each other and can be used as labels of document clusters.

In paper [25], we depict latent Dirichlet allocation (LDA), a generative probabilistic model for accumulations of discrete data, for example, text corpora. LDA is a three-level hierarchical Bayesian model, in which each set of an accumulation is demonstrated as an infinite mixture over a hidden arrangement of subjects. Every subject is, in turn, demonstrated as an unending mixture over a hidden arrangement of point probabilities. In the connection of content demonstrating, the subject probabilities give an unequivocal representation of an archive. We exhibit effective rough inference techniques in view of variational methods and an inference technique for exact Bayes paper estimation. We report brings about text classification, and collaborative filtering, also, community oriented sifting, contrasting with a mixture of unigrams model and the probabilistic LSI model.

In paper [26], here we display a trainable model for distinguishing sentence limits in raw text. Given a corpus clarified with sentence boundaries, our model figures out how to group every occurrence of.,?, and / as either a valid or invalid sentence limit. The preparation system requires no hand-crafted rules, lexica, part-of-speech tags, or area particular information. The model can thusly be trained effortlessly on any sort of English, and should to be trainable on whatever other Roman alphabet language. Performance is practically identical to or better than the execution of similar systems, however we stress the effortlessness of retraining for new domain.

In this article [27], we show a language-independent, unsupervised way to deal with sentence limit detection. It is in light of the assumption that an extensive number of ambiguities in the determination of sentence boundary can be disposed of once abbreviations have been identifying. Rather than depending on orthographic pieces of information, the proposed framework has the capacity identify accuracy with high exactness utilizing three criteria that just oblige data about the candidate sort itself and are independent of context: Abbreviations can be characterized as a tight collocation comprising of a truncated word and a last period, abbreviations forms are normally short, and condensing sometimes contain internal periods. We likewise demonstrate the capability of collocation proof for two other vital sub tasks of sentence boundary disambiguation, such that detection of initials and ordinal numbers. The proposed framework has been tried widely on eleven different languages and on diverse text classifications. It accomplishes great results with no further alterations on the other language-specific resources. We find its execution against three separate baselines and contrast it with different frameworks for sentence boundary detection proposed in the literature.

In this paper [28], with the developing utilization of Natural Language Processing (NLP) techniques for data extraction and idea indexing in the biomedical area, a strategy that quickly and efficiently relegates the right feeling of an ambiguous biomedical term in a given setting is required simultaneously. The current status of word sense disambiguation (WSD) in the biomedical domain is that handcrafted rules are utilized in view of relevant material. The disservices of this methodology are (i) generating WSD rules manually is a time-consuming and tedious task, (ii)maintenance of rule sets becomes increasingly difficult over time, and (iii) handcrafted rules are often incomplete and perform poorly in new domains comprised of specialized vocabularies and different genres of text.. This paper displays a two-stage unsupervised strategy to manufacture a WSD classifier for an equivocal biomedical term W. The principal stage naturally makes a sense-labeled corpus for W, and the second phase infers a classifier for W utilizing the determined sense-labeled corpus as a preparation set. A developmental examination was performed, which devil started that classifiers prepared on the determined sense-labeled corpora attained to a general exactness of around 97%, with more prominent than 90% precision for every individual uncertain term.

3. Methodology

In this Methodology D-Matrix development comprises of the accompanying building block of document annotation, term extraction, and phrase merging. At first, the repair verbatim data focuses are gathered by recovering them from the OEM's database, which are recorded during field FD. In the first step, the terms, such as, part, symptom, and failure mode, relevant, applicable for the D-matrix are explained from every repair verbatim by building up the document annotation algorithm. A repair verbatim comprises of a

several sections, symptoms, failure modes and actions and the right affiliations must be built between the important terms in view of their proximity with each other. Here, a repair verbatim is first part in different sentences by utilizing the sentence boundary detection rules and the terms showing up in the same sentence are co-related with one another.

(i). Fault Diagnosis Ontology: The field of ontology building has collected a genuine attention of the artificial intelligence community in 90s. An ontology can be seen as a information model that

Expressly depicts different concepts that exist in a domain of discourse, along with their properties. The fault diagnosis ontology specifically is a lightweight ontology which is formalized by utilizing the ontology development methodology. It catches the terms and the relations saw in the domain of vehicle fault diagnosis. The cases are formalized utilizing the asset depiction schema which facilitate exchange, storage and machine readability of ontology.

(ii). **Ontology-Based Text Mining:** In this model we describe some few steps viz. term extractor, document annotation, and phrase merging involved in the ontology based text mining construction of a D-matrix

(a). **Document Annotation:** Due to the several types of we saw in our information (Section I) the undertaking of distinguishing the principle building blocks of D-matrix, for example. parts, symptoms, and failure modes into a non-trifling activity. The documents annotation serves to channel out the data that is irrelevant for our investigation and it gives a particular connection for the consistent and shared interpretation information. Initially, the following preprocessing steps—the sentence boundary detection (SBD), are utilized to part a repair verbatim into partitioned sentences, the stop words are erased to the non-expressive terms, and the lexical matching identifies the right significance of abbreviations. Therefore the terms from the prepared verbatim are coordinated utilizing the examples in the fault diagnosis ontology.

(b). **Term Extractor:** In this method we expounded the terms, the critical terms required for the development of a D-matrix i.e. symptoms and failure modes are extricated by utilizing the term extractor algorithms. Initially, the causal connection between the relevant symptom-failure mode sets is distinguished to verify that just the right matches are extricated. The existing methodologies for frequent item sets mining disregards the request in which the term expressions are recorded in document, anyway we must keep up such requesting to see how the fault diagnosis is performed.

(c). **Phrase Merging:** In this Point, Because of the human intervention while catching the repair verbatim data an irregularity has been seen in the information regarding the term used to record the failures modes, e.g., Tank Sensor_Inop or Tank Pressure Sensor Internal Short. So, we check whether two distinctive disappointment modes are the varieties of basically the same failures mode such that they can be converged before populating a D-Matrix. The phrase merging algorithms takes as enter an arrangement of failures modes extricated by the term extractor algorithm. A priori our system does not have the information about which two failure modes can be combined; thus the closeness of every failures mode is merged with each other for example, Evap leak at fuel tank and Tank draining at neck and we have to further process these phrases. A few procedures are proposed to study the vocabulary mismatch problem, stemming, translation models, query expansion, LSI, among others. The phrase merging algorithm is comparative in soul with the query expansion procedures, where every failure mode is dealt with as a potential query and amid the development stage extra data is gathered as the attributes.

4. Conclusion

With the help of ontology–based text technology we developed D-Matrices by using automatically mining the different structure repair verbatim information is gathered during the time of fault diagnosis. The manual developing of D-matrix diagnostic model it would be participates to integrate the information from SME's. In most of the cases SME may not able to realize all dependencies between fault diagnosis and failure nodes. The main approach of these system is to overcome the limitations where natural language algorithms represent the automatically develop the D-matrices from the unstructured repair verbatim. We also comparison to diagnosability metrics of the historical data-driven D-matrix and testability. In Section four we have discussed the text-driven D-matrix approach achieved higher fault isolation, and lower ambiguity group size due to the DTC Symptoms. And finally we discussed the efficiency of text-driven D-matrix which is compared with LDA. The main used of LDA is improving fault isolation rate and fault detection. In a future point of view the main goal of each D-matrix as a graph and develop the graph comparisons algorithms.

References:

[1] Dnyanesh G. Rajpathak, "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] O. Benedittini, T. S. Baines, H. W. Lightfoot, and R. M. Greenbush, "State-of-the-art in integrated vehicle health management," *J. Aer. Eng.*, vol. 223, no. 2, pp. 157–170, 2009.
[3] T. Hearst, "Untangling text data mining," in *Proc. 37th Annu. Meeting Assoc. Comput. Linguist*, 1999, pp. 3–10.
[4] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. Kavuri, "A review of process fault detection and diagnosis Part I: Quantitative model based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 293–311, 2003.
[5] *IEEE standard for artificial intelligence exchange and service tie to all test environments (AI-ESTATE)*, IEEE Std. 1232–2002, 2002.
[6] 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.*, 2008, pp. 1–6.

[7] T. Felke, "Application of model-based diagnostic technology on the Boeing 777 airplane," in *Proc. 13th AIAA/IEEE DASC*, 1994, pp. 1–5.
[8] G. Ramohalli, "The Honeywell on-board diagnostic and maintenance system for the Boeing 777," in *Proc. IEEE/AIAA DASC*, 1992, pp. 485–490.
[9] P. M. Frank and J. W¨unnenberg, "Robust fault diagnosis using unknown input observer schemes," in *Proc. Fault Diagnosis Dynamical Syst.: Theory Appl.*, 1989, pp. 47–98.
[10] N. Viswanadham and R. Srichander, "Fault detection using unknown Input observers," *Control-Theory Ad Tech.*, vol. 3, pp. 91–101, 1987.
[11] 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.
[12] Y. Dingli, J. B. Gomm, D. N. Shields, D. Williams, and K. Disdell, "Fault diagnosis for a gas-fired furnace using bilinear observer method," in *Proc. Amer. Control Conf.*, 1995, pp. 1127–1131.
[13] H. Yang and M. Saif, "Nonlinear adaptive observer design for fault detection," in *Proc. Amer. Control Conf.*, 1995, pp. 1136–1139.
[14] 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.
[15] C. Charniak and D. McDermott, *Introduction to Artificial Intelligence*. Reading, MA, USA: Addison Wesley, 1985.
[16] 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.
[17] 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.
[18] M. A. Kramer and B. L. Palowitch, "A rule based approach to fault diagnosis using the signed directed graph," *AIChE J.*, vol. 33, no. 7, pp. 1067–1078, 1987.[26] R. F. Li and X. Z. Wang, "Qualitative/quantitative simulation of process temporal behavior using clustered fuzzy digraphs," *AIChE J.*, vol. 47, no. 4, pp. 906–919, 2001.
[19] J. F. McGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Eng. Practice*, vol. 3, no. 3, pp. 403–414, 1995.
[20] W. Li, H. Yue, S. Valle-Cervantes, and S. J. Qin, "Recursive PCA for adaptive process monitoring," *J. Process Control.*, vol. 10, no. 5, pp. 471–486, 2000.
[21] S. Singh, H. S. Subramania, and C. Pinion, "Data-driven framework for detecting anomalies in field failure Data," in *Proc. IEEE Aerosp. Conf.*, 2011, pp. 1–14.
[22] Satnam Singh, *Member, IEEE*, Anuradha Kodaly, Kihoon Choi, Krishna R. Pattipati," Dynamic Multiple Fault Diagnosis: Mathematical Formulations and Solution Techniques"
[23] Weihua Li, H. Henry Yue, Sergio Valle-Cervantes, S. Joe Qin," Recursive PCA for adaptive process monitoring", Department of Chemical Engineering, The University of Texas at Austin, Austin, TX 78712, USA .
[24] Wen Zhang, Taketoshi Yoshida , Xijin Tang , Qing Wang," Text clustering using frequent itemsets", Internet Software Technologies, Institute of Software, Chinese Academy of Sciences, Beijing 100190.
[25] David M. Blei, Andrew Y. Ng, Michael I. Jordan," Latent Dirichlet Allocation", Computer Science Division and Department of Statistics University of California Berkeley, CA 94720, USA .
[26] Jeffrey C. Reynar and Adwait Ratnaparkhi, "A Maximum Entropy Approach to Identifying Sentence Boundaries", Department of Computer and Information Science University of Pennsylvania,USA
[27] Tibor Kiss, Jan Strunk ," Unsupervised Multilingual Sentence Boundary Detection"
[28] Hongfang Liu,*,1 Yves A. Lussier,†,2 and Carol Friedman," Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method "