Indian J.Sci.Res. 6(1) : 47-52, 2015

ISSN : 0976-2876 (Print)
ISSN : 2250-0138 (Online)

AN ONTOLOGY TEXT MINING TO CONVERSION OF UNSTRUCTURED TO STRUCTURE TEXT IN D-MATRIX

RADHIKA Y. DEORE[1]

Department of Computer Engg., Matoshree College of Engineering and Reaserch Centre, Nashik, Pune University, India

ABSTRACT

In the general context of knowledge discovery, that technique called textual content mining technique, are important to extract facts from unstructured textual information. The extracted records can then further used to classify the content of massive textual bases. This paper discusses Fault dependency (D)-matrix is a systematic diagnostic version that desires to capture the hierarchical device-degree fault diagnostic information.It including dependencies between observable symptoms and failure modes related to a system. Every time person type any query for looking any report or information, most possibly all of the files or records looking to search query with title of to be had information and building a D-matrix from first concepts and updating it the usage of the domain know-how is a hard work intensive and time consuming project. Similarly, in-time augmentation of D-matrix through the revelation of new signs and symptoms and failure modes found for the first time is a hard challenge. Proposed machine describes an ontology primarily based text mining approach for automatically constructing and updating a D-matrix with the aid of mining loads of thousands of restore verbatim(generally written in unstructured text). In proposed technique, firstly construct the fault analysis ontology such as ideas and relationships typically observed within the fault prognosis area. Next, hire the textual content mining algorithms that employ ontology concept to become aware of the wanted artifacts, inclusive of additives, signs, failure modes, and their dependencies from the unstructured restore verbatim textual content.

KEYWORDS : Data Mining, Fault Analysis, Fault Diagnosis, Information Retrieval, Text Processing

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. 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.. 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

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.

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

A text mining method to map the diagnostic information extracted from the unstructured repair verbatim in a D-matrix Text Mining (Hearst, 1999) is automatically discovers the knowledge assets buried in unstructured text. Text mining (Michael, 2004) is similar to data mining, difference is that data mining tools (Navathe and Ramez

[1]**Corresponding author**

2000) are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc..

Literature Survey

According to Schuh et al., 2013 in this approach to Faults are removing and provide ontology-guided data mining and data transformation. But Discovery is loss because result is not in form of matrix.

According to Singh and Dhir, 2013 in this paper it is transaction reduction for finding for finding item sets based on tags and shows result in matrix. It does not give accurate result. Its search is only based on tags. No use of ontology In this paper 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.

According to Sobhani and Poshtan, 2013 this paper investigates the detection and isolation of faults using structured residuals. Actuator and sensor faults are considered. Residuals are generated using a bank of unknown input observers (UIO). Three-tank benchmark system was used as a prototype of many process industries. Simulation results show the effectiveness of the studied method.

According to Gaeta et al., 2012 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 to query not accurate result in this form like D-MATRIX.

Ching-AngWu qt al., 2011 in this paper ,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) but its very time consuming

According to Wen Zhang et al., 2010 in this paper text mining such as document clusterization and assign cluster topic. It only cluster the frequent data but not showing result in D-Matrix

According to Marti, 2011 it presented first data mining, information access, and corpus-based computational linguistics, and then discusses the relationship of these to text data mining. The intent behind these contrasts is to draw attention to exciting new kinds of problems for computational linguists.

According to Venkatasubramanian et al., 2012 history based method in this paper One realizes that no single method has all the desirable features one would like a diagnostic system to possess. It is our view that some of these methods can complement one another resulting in better diagnostic systems. Integrating these complementary features is one way to develop hybrid systems that could overcome the limitations of individual solution strategies. The important role of fault diagnosis in the broader context of process operations is also outlined. We also discuss the technical challenges in research and development that need to be addressed for the successful design and implementation of practical intelligent supervisory control systems for the process industries.

Singh et al., proposed Dynamic Multiple Fault Diagnosis: Mathematical Formulations and Solution Techniques. In this paper, they discuss four formulations of the DMFD problem. These include the deterministic situation corresponding to a perfectly observed-coupled Markov decision processes, to several partially-observed factorial hidden Markov models ranging from the case where the imperfect test outcomes are functions of tests only to the case where the test outcomes are functions of faults and tests, as well as the case where the false alarms are associated with the nominal (fault-free) case only. All these formulations are intractable NP-hard combinatorial optimization problems. We solve each of the DMFD problems by decomposing them into separable sub problems, one for each component state sequence.

Literature Survey

In present work it described diverse methods to assemble D-matrices by way of using the information resources, together with provider tactics and engineering design The associations among the failure modes and symptoms are mapped in a D-matrix by way of using the signal go with the flow diagrams and engineering expertise of a system, consisting of a failure mode, consequences,

criticality analysis information, signal information, and engine control unit facts and viewing the overall statistics this is saving all database and first of all parse that facts and after that scan ordinary records so its takes more information base reminiscence and its very plenty time consuming for parsing and scanning that standard databases. But, the improvement of D-matrix from scratch takes massive engineering effort and time, at the same time as the facts mining strategies have proven to shop the construction time of D-matrix from the sphere fail-trap statistics. The fidelity of the records-driven D-matrix is decrease due in component to the noise inside the discipline failure information, while the service procedures- primarily based D-matrix is of higher constancy, however of lower fidelity when in comparison to the engineering layout-primarily based D-matrix. The information pushed framework detects anomalies by way of the use of the system level fault version and diagnostic reasoned built via mining the running sensory parameter identifiers statistics.

A. Manual System

In guide system [Pankaj Chandre, Bharat Vishe, Hemant Vishe, Pralhad Lengule and Ankush Shah 2014], whilst getting to know any content on in step with the call of that unique studies content material. The contrast of studies proposals in present is achieved manually. This is the proposals are submitted to finding urgency and in keeping with the name of studies proposals or paper and the keywords the research proposals. In a while it classifies into the businesses or under particular domain this all process performed manually means by means of the human. Following diagram indicates the manner of guide clustering of research proposals.

However that is not possible for huge information. It makes misplacement of research proposals due to guide process and type according only the call of research proposals. So this misplacement makes the specialists extra confuse of the research proposals which are not from their region of studies. There exists the software which also can't cope with the big records and misplacement in studies proposals.

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. In existing work natural language processing has produces different no of technologies that teach to computers natural languages that they can understand the natural language so they can generate text. Some of technologies [Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju] [Vishal Gupta, Gurpreet S. Lehal, 2009] can be used in the text mining process are information extraction, summarization, categorization, Clustering.

The Existing Text Mining Process

Entire html-pages are downloaded from a given discussion board web site.

Html-like tags and non-textual records like photos,

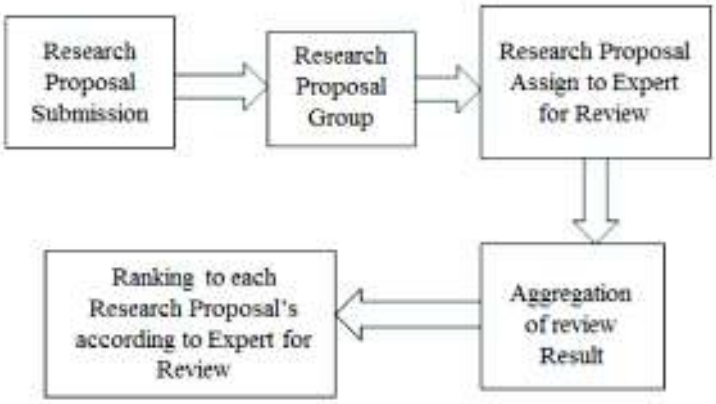

Figure 1: Existing Manual System

advertisements, and so forth. Are cleaned from the downloaded textual content.

The textual parts are divided into informative gadgets like threads, messages, and sentences.

it's miles sequential manner so its takes lot of time for processing.

Its needs lot of reminiscence.

Little need of Re-rating so it result are not accurate.

Proposed Model

The proposed device is primarily based on the Epistemology this is shape of the Ontology in textual content mining. It paperwork 3 phases to manner. That consist file Annotation and mixed The term Extraction and phrase Merging for Optimization of time. We advocate a textual content mining method to map the diagnostic data extracted from the unstructured repair verbatim in a D-matrix. But, the development of a D-matrix via using text mining is a hard project partially because of the noises determined inside the restore verbatim text information abbreviated text entries the abbreviation are used to document the phrases and it's far vital to disambiguate their that means, incomplete textual content entries the unfinished repair data makes it hard to derive the right knowledge from the information; term disambiguation the same time period is written with the aid of the usage of inconsistent vocabulary. Typically the procedure of FD starts via extracting the mistake codes from a goal device and based on the found blunders codes the technicians observe particular diagnosis procedure together with their experience to diagnose the faults. At some point of fault analysis, several information sorts are gathered, which include blunders codes, scanned values of operating parameters associated with faulty component/machine, restore verbatim, and so forth. The accrued statistics is then transferred to the OEM database and in particular the restore verbatim records accumulated over a period of time may be mined to expand the D-matrix diagnostic fashions. Such fashions can be used to perform accurate FDD. The D-matrix captures element and gadget degree dependencies between a single or more than one failure modes1 (or root-cause of failures) with a single and more than one symptoms (a set of fault codes, determined signs, and so forth.) in a established style. These dependencies among failure modes (f1, f2, etc.) in elements (p1, p2, etc.) and symptoms (s1, s2, and etc.) permit us to country a hard and fast of failure modes causing signs and symptoms. Also, the causal weights (d11, d12, etc.) are contained on the intersection of a row and a column suggests a probability of detection. Inside the binary D-matrix, all the probabilities have a price of both 0 or 1, wherein 0 indicates no detection and 1 indicates complete detection of a specific failure mode using a particular symptom. The values between 0 and 1 indicate the level of power of detecting a failure mode through the usage of a symptom. Especially, our paintings falls into the quantitative and statistics-driven fault diagnosis classes, whereby a textual content driven D-matrix development technique is proposed where first of all the fault prognosis ontology is constructed by means of mining the unstructured restore verbatim records. Sooner or later, the textual content mining algorithms are advanced, which uses this ontology to find out the dependencies among the signs and symptoms and the failure modes. The certified institutions are used to assemble the D-matrix diagnostic model.

The proposed device will use the equal Mathematical programming model but with optimized by blended the 2 phases term Extraction and word merging so optimize the greater time for processing. Proven correlated result instantly offers for calculation possibilities not anything however paintings like multithreaded.

Growing a vocabulary for ontology is to extract critical terms from textual content documents associated with a particular domain.

The corpus is then parsed into tokens or terms.

Unstructured textual content inside the corpus turns into a established records object via the advent of a term-through-record frequency matrix.

Frequency weights of those concepts can be adjusted to account for the distribution of phrases throughout documents.

Natural language processing (NLP) and text mining strategies are powerful for statistics extraction from text documents.

Essentially, final bring about D-matrix which makes use of for contrast between or extra outcomes which isn't always in existing machine.

Document Annotation Algorithm

Inside the first step, the terms, together with part, symptom, and failure mode, relevant for the D-matrix are annotated from every repair verbatim by way of growing the report annotation algorithm. The document annotation allows to clear out the facts this is beside the point for our evaluation and it gives a specific context for the regular and shared interpretation of the records. To start with, the subsequent preprocessing stepsthe sentence boundary detection (SBD), are used to cut up a repair verbatim into separate sentences, the prevent words are deleted to put off the non-descriptive terms, and the lexical matching identifies the precise that means of abbreviations. Ultimately the terms from the processed verbatim are matched using the times inside the fault prognosis ontology.

Term Extraction Algorithm

Time period Extraction extracts pieces of information which might be silent to the consumer want. It extracts the precise information which consumer wants. Having annotated the terms, the crucial terms wanted for the construction of a D-matrix, i.E. Via the usage of the time period extractor set of rules signs and symptoms and failure modes are extracted.

CONCLUSION AND FUTURE WORK

This paper has offered an Ontology text mining manner. It constructs the D-matrices through routinely mining the unstructured restore verbatim data accrued in the course of fault diagnosis. Text facts Mining or information-Discovery in textual content keywords in unique place which might be refers back to the technique of extracting interesting and non-trivial statistics and understanding from unstructured textual content. In actual-lifestyles, the guide production of a D-matrix is complicated structures and time eating the present strategies require more statistics for training as well as the computational time of those strategies and there no rating manner.

The proposed method overcame these barriers wherein natural language processing algorithms were proposed to routinely increase the D-matrices from the unstructured repair verbatim. In comparison to the existing algorithms, the proposed hybrid algorithm calls for much less schooling facts and much less computational time. The proposed work encourages the efficiency within the inspiration clustering procedure and the resulted clusters are

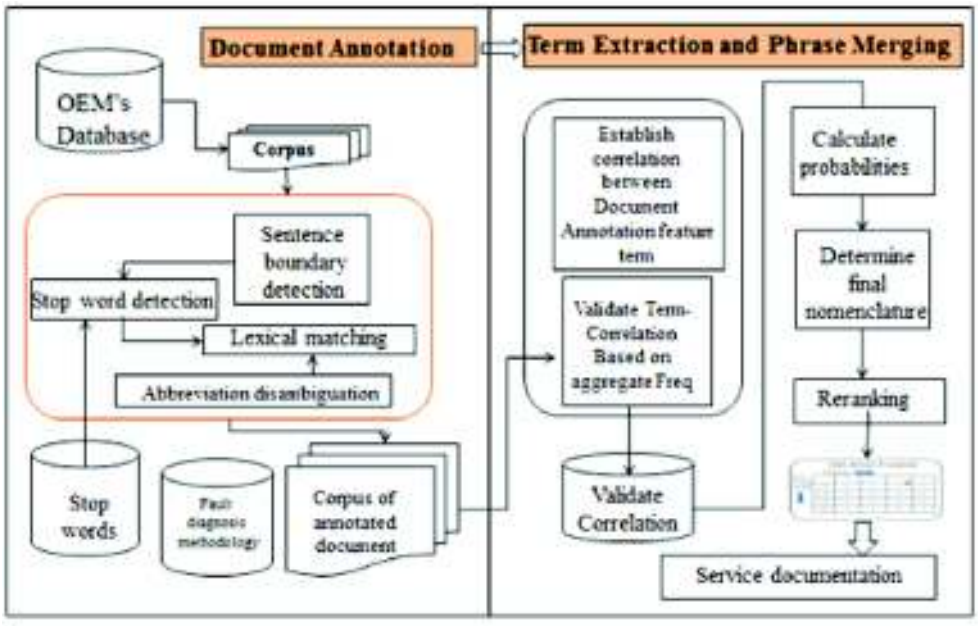

Figure 2: Proposed Text-driven D-matrix Development Methodology

in D-matrix shape that is simple for evaluation cause. In future we can combine unique clustering and re-ranking algorithms and evaluate the consequences.

REFERENCES

Berry Michael W., 2004 .Automatic Discovery of Similar Words, in Survey of Text Mining: Clustering, Classification and Retrieval. Springer Verlag, New York, LLC: 24-43.

Chandre P., Vishe B., Vishe H., Lengule P.and Shah A.,2014. Ontology in Text Mining To Cluster Research Project Proposals, **4**(4).

Ching-AngWu,Wen-Yang Lin,Chang-long Jiang, 2011.Towards intelligent data warehouse mining: An ontology integrated approach for multidimensional association. In September 2011.

Gupta V. and Gurpreet S. L.,2009. A Survey of Text Mining Techniques and Applications.Journal of emerging technologies in webintelligence. August 2009.

Mohammad H.S. and Poshtan J., 2012. Fault Detection and Isolation Using Unknown Input Observers With Structured Residual Generation. In April, **3**(3).

Hearst T., 1999. Untangling text data mining. In Proc. 37th Annu. Meeting Assoc. Comput. Linguist.: 310.

Gaeta M. , Orciuoli F. , Paolozzi S. and Salerno proposed S. , 2012.Ontology extraction for knowledge reuse: The e-learning perspective. In July, 2012.

Schuh M., Sheppard J. W., Strasser S., Angryk R., and Izurieta C., 2013. A Visualization tool for knowledge discovery in maintenance event sequences.In July 2013.

Marti A. Hearst ,2011.Untangling Text Data Mining .In July,2011.

Navathe, Shamkant B., and Elmasri Ramez, 2000. "Data Warehousing And Data Mining", in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, **2**(2):841-872.

Singh H. and Dhir R., 2013.A New efficient Matrix based frequent Item sets Mining Algorithm with Tags. In August 2013.

Strasser S., Sheppard J., Schuh M., Angryk R., and Izurieta C. 2011.Graphbased ontology-guided data mining for d-matrix model maturation. in Proc. IEEE Aerosp, **21**: 1-12

Venkat V. , Raghunathan R., Surya N. K. and Y. Kewen, 2002. A review of process fault detection and diagnosis Part III: Process history based methods. In July, 2002.

Venkatasubramanian V., Rengaswamy R., Surya N., Kavuri, Kewen Y., 2002. A review of process fault detection and diagnosis Part III: Process history based methods **4** (4), In July, 2002.

Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju, Zhang, 2005. "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, **4** (4).

Wen Zhang, Taketoshi, XijinTang, Qing Wang, 2010.Text clustering using frequent item sets. In July, 2010.