Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Text Mining on Unstructured Data using D-matrix

**Poonam B. Kucheria, Prof. Kiran. P. Gaikwad**
Department of CSE, S.N.D.C.O.E.R.C, Yeola, Dist-Nasik,
Savitribai Phule University of Pune, Pune, Maharashtra, India

***Abstract— Fault dependency (D)-matrix is a systematic diagnostic model which is used to capture the fault system information at the hierarchical system-level. It consist dependencies between observable symptoms and failure modes associated with a system. Whenever user type any query for searching any file or data, most probably all the files or data trying to match its search query with title of available data and constructs a D-matrix. 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. The proposed method will be implemented as a prototype tool and validated by using real-life data collected from the automobile domain.***

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

## I. INTRODUCTION

A complex system interacts with its surrounding to execute a set of tasks within an accepted range of tolerances by maintaining its performance. If anything goes wrong then it is considered as a fault. The fault detection and diagnosis (FDD) is done to detect the faults and diagnose the root-causes to minimize the downtime of a system. However, the overwhelming size of the repair verbatim data restricts an ability of its effective utilization in the process of FDD. Further, Text mining is gaining a serious attention due to its ability to automatically discover the knowledge assets buried in unstructured text.

Ontology learning systems for concept extraction were 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.

Automatically constructing and updating results by mining of thousands of repair verbatim (typically written in unstructured text) collected during the diagnosis episodes. And it will also construct result for unstructured PDF and document files in the form of D-Matrix fastly and accurately. Ontology is the philosophical study of the nature of being, becoming, reality and existence, and also of the basic categories of being and their relations.

## II. LITERATURE SURVEY

M. Schuh, J. W. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, personalized search has been proposed for many years and many personalization strategies have been investigated, to remove Faults and provide ontology-guided data mining and data transformation but Discovery is loss because result is not in form of matrix.

Harpreet singh and renu Dhir also did study on transaction reduction for finding item sets based on tags and shows result in matrix but it does not give accurate result. Its search is only based on tags. There was no use of ontology.

M. Gaeta, F. Orciuoli, S. Paolozzi, and S. Salerno, provide an easy to use interface that generates relevant sequences of data in meaningful context and retrieve and display similar information but it only shows similar information not accurate result in this form like D-MATRIX.

Ching-Ang Wu,Wen-Yang Lin,Chang-long Jiang, has proposed which builds useful data mining models and it present prototype multidimensional mining system,but mining hundreds of thousands of repair verbatim (typically written in unstructured text).

Wen Zhang,Taketoshi,Xijin Tang,Qing Wang, proposed on text mining such as document clusterization and assign cluster topic but it only cluster the frequent data but not showing result in D-Matrix.

Till now, there is no any accurate service available for the data retrieval system using text information. Existing systems are depends on the title which is given to Files/ Data. Title of each File is used as a main parameter for sorting the number of Data against the search query.

 

## III. EXISTING SYSTEM

We develop our windows based application on a computer system. There are 2 parts online and offline parts.

- At online stage, we are passing web URL in text box as an input parameter. After getting URL we are validating URL with null URL or Flipkart URL. If the URL is valid then html contents are fetched of that URL. Further html contents are parsed into HTMLAgibility object. As per the requirement we are converting unstructured data into structured data. Then structured data is stored in the database for further operations (i.e. sorting, searching etc.)
- In offline stage, first we are passing pdf file path as an input parameter. After getting the path then it is validated. If the path is valid then data is fetched from pdf into text format, then the text data is transferred into XML. As per the requirement we are converting unstructured data into structured data. Then structured data is stored in the database for further operations (i.e. sorting, searching etc.)

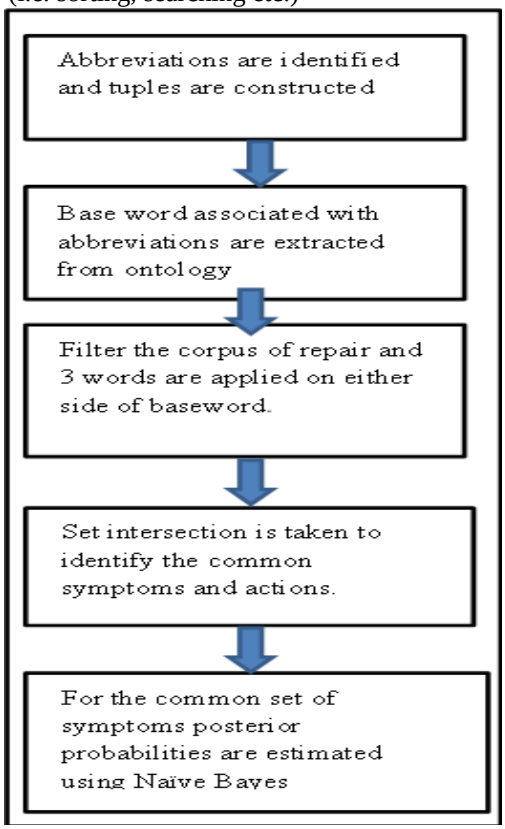


Figure1: Process of filtering

## IV. PROPOSED SYSTEM

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.

Automatically constructing and updating results by mining of thousands of repair verbatim (typically written in unstructured text) collected during the diagnosis episodes. And it will also construct result for unstructured PDF and document files in the form of D-Matrix fastly and accurately. To implement a model which captures the Title and Description all the captured data are then classified according to the duplication property. It is used for further process of data retrieval system.

## V. SYSTEM DESIGN

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.


Figure2: Data Flow Diagram

 

As shown in Figure3, there are three processes in our structuring framework: an ontology parser, a constant/keyword recognizer, and a structured-text generator. The input is application ontology and a set of unstructured documents and the output is a populated relational database. A main program invokes the parser, recognizer, and generator in the proper sequence. The ontology parser is invoked only once at the beginning of execution, while the recognizer and generator are repeatedly invoked in sequence for each unstructured document to be processed.

The input to our system is application ontology and a set of unstructured documents.

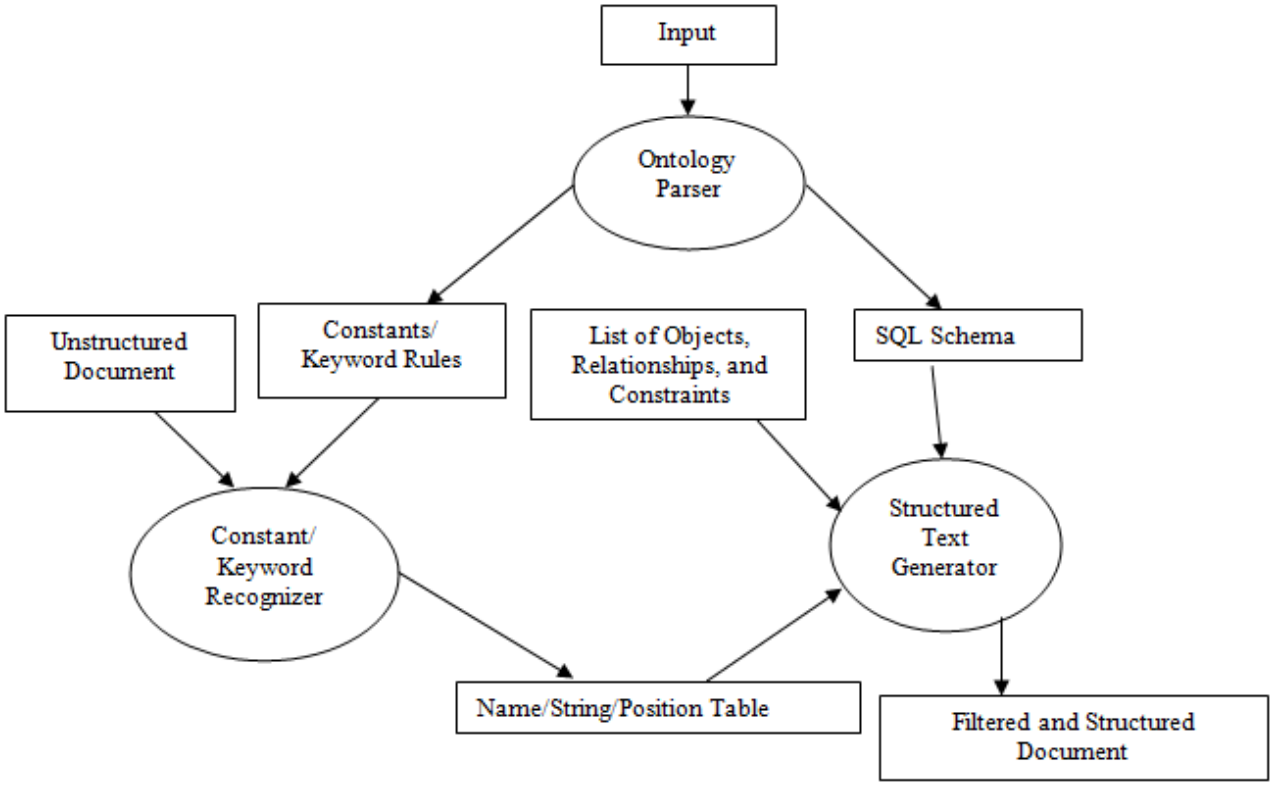


Figure3: Architecture Diagram

## VI. CONCLUSION

An ontology-based text mining methodology has been proposed to construct the D-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis. We have provided a framework for converting data-rich unstructured documents into structured documents. In addition, we have implemented the procedures in our framework, and we have demonstrated that our framework and implemented procedures achieve good results.

## ACKNOWLEDGMENT

Authors are cordially giving thanks to the researchers of different model for Web-Based Structured data and to get D-matrices from Unstructured Data. All other who have tried hard to make their work easy to accomplish.

## REFERENCES

[1] M. Schuh, J. Sheppard, S. Strasser, R. Angryk, and C. Izurieta,"Ontology-guided knowledge discovery of event sequences in maintenance data," in *Proc. IEEE AUTOTESTCON Conf* ., 2011, pp. 279–285.

[2] M. Schuh, J. W. Sheppard, S. Strasser, R. Angryk, and C. Izurieta, "A Visualization tool for knowledge discovery in maintenance event sequences," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 28, no. 7, pp. 30–39, Jul. 2013.

[3] 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.

[4] M. Gaeta, F. Orciuoli, S. Paolozzi, and S. Salerno, "Ontology extraction for knowledge reuse: The e-learning perspective," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 4, pp. 798–809, Jul. 2011.

[5] J. Sheppard, M. Kaufman, and T. Wilmering, "Model based standards for diagnostic and maintenance information integration," in *Proc. IEEE AUTOTESTCON Conf.*, 2012, pp. 304–310.

[6] W. Zhang, T. Yoshida, X. Tang, and Q. Wang, "Text clustering using frequent itemsets," *Knowl.-Based Syst*., vol. 23, no. 5, pp. 379–388, 2010.

[7] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, "Bug localization using latent dirichlet allocation," *Inf. Softw. Technol*., vol. 52, no. 9, pp. 972–990, 2010.

 