Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.156 – 159

REVIEW ARTICLE

A Review on Improved Text Mining Approach for Conversion of Unstructured to Structured Text

Pranita Baitule[1], Prof. Vikrant Chole[2]

[1]Department of Computer Science and Engineering, GHRAET, Nagpur, India

[2]Department of Computer Science and Engineering, GHRAET, Nagpur, India

[1] pranitabaitule@gmail.com; [2] Vikrantchole@gmail.com

Abstract- In text mining, the purpose is to process unstructured (textual) information, extract meaningful numeric indices from the text, and thus make the information contained in the text accessible to the various data mining. A new approach to extracting information from unstructured documents based on an application ontology that describes a domain of interest. More knowledge and information eventually available into computers, critical capability of systems supporting learning management is classification of documents into ordering that are meaningful to the user. Ontology based text mining method used for accordingly constructing and renovating D-matrix by mining hundreds of thousands of repair verbatim collected during diagnosis period. In this paper we will proposed text mining algorithms to process data on the unstructured text to use of this ontology for identify the necessary artifacts, like parts, symptoms, failure modes and dependencies from the unstructured repair verbatim text.

Keywords- Text mining, ontology, data mining, information retrieval

I. Introduction

To maintain the performance of an acceptable range of tolerances the system must interacts with surrounding for executing some set of task. A fault can be defined as detecting abnormal process behavior. FDD is fault detection and diagnosis performance for detecting the fault and diagnose root causes of the system. Commonly book kept diagnosis data comes in the form of unstructured repair verbatim that provides rich source of diagnostic information. Hundreds of thousands of repair verbatim are collected and argue that there is an urgent need to mine this data to improve fault diagnosis (FD). The size of the repair verbatim data restricts an ability of its effective utilization in process of FD. Generally the process of fault diagnosis starts by extracting error codes from a target system and based on the recognized error codes technicians follow specific diagnosis procedure along with their experience to diagnose the faults. During fault diagnosis, many data types are collected, such as error codes, scanned values of operating parameters associated with faulty component system, repair verbatim. The collected data transferred to the database and particularly a repair verbatim data collected over a period of time can be mined to develop the D-matrix diagnostic models. These models can be used by field technicians and other stakeholders for performing accurate FDD. The D-matrix captures component and system level dependencies between a single or multiple failure modes with a single and multiple symptoms in a structured way.

Text mining is achieving an attention due to its ability to automatically discover the knowledge assets buried in unstructured text. Text mining is an important step of knowledge discovery process. In this paper, propose a text mining method to map 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 challenging task partly due to noises recognized in the repair verbatim text data – *abbreviated text entries:* the abbreviation are used to record the terms and it is insistent to determine their meaning, for example, loose *fr*door *chked*repair performed; *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 unpredictable vocabulary, e.g., FTPS_Inop and FTPS_Internal Short, and inconsistency must be extracted to maintain uniformity. The D-matrix obtains 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 way. These dependencies with 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 enclosing 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 this 0 or 1, where 0 indicates no detection and 1 indicates complete detection of a particular failure mode using a specific symptom. The values between 0 and 1 indicate the level of power of detecting a failure mode by using a symptom. In particular 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. The output of our process is a text-driven D-matrix as a diagnostic model.

A principle approach is proposed to develop a D-matrix diagnostic model by detecting an unstructured repair verbatim data associated with the multiple systems in parallel through the development of ontology-based text mining algorithms. It overcomes the limitation faced in the real life industry of having to construct the D-matrix diagnostic models manually or using first principles. Further, in our approach we are able to capture the cross-system dependencies, which helped to significantly improve the performance of FDD. The relations from the fault diagnosis ontology are used to discover the dependencies between the symptoms and the failure modes corresponding to different systems. It improved the performance of system when compared with the Latent-Dirichlet Allocation (LDA) technique. While constructing the D-matrix models, the abbreviations are disambiguated by merging the inconsistent failure modes into a single, consistent term, which provides homogeneous, consistent fault models.

Text mining involves the application of techniques from areas like information retrieval, natural language processing, information extraction and data mining. *Information Retrieval* (IR) systems identify the documents in a collection which match a user's query. The most well-known IR systems are search engines such as Google, which allows identification of a set of documents that relate to a set of key words. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the discovery cycle considerably by reducing the number of documents found for analysis. consider, if a researcher is interested in mining information only about protein interactions, he/she might restrict their analysis to documents that contain the name of a protein, or some form of the verb 'to interact', or one of its synonyms. Already, through application of IR, the vast accumulation of scientific research information can be reduced to a smaller subset of relevant items. *Natural Language Processing* (NLP) is the analysis of human language so that computers can understand research terms in the same way as humans do. *Information Extraction* (IE) is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that the researcher is interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generated by NLP systems. *Data Mining* (DM) (often known as knowledge discovery) is the process of identifying patterns in large sets of data. When used in text mining, DM is applied to the facts generated by the information extraction phase.

Continuing with the protein interaction example, the researcher may have extracted a large number of protein interactions from a document collection and stored these interactions as facts in a separate database. By applying DM to this separate database, the researcher may be able to identify patterns in the facts. This may lead to new discoveries about the types of interactions that can or cannot occur, or relationship between types of interactions and particular diseases, and so on. Ontology can be defined as particularization of a representational vocabulary for shared domain of conversation which may include definitions of classes, relations, functions and objects. An ontology-based computer system does not interact directly with real world but rather with internal models of the relationships between concepts and objects in the real world. Such models represent problem domains, and development of these models in computers is referred as ontology building. Most of the information is available in the form of unstructured natural language documents due to the growth of the web, digital libraries, technical documentation, etc. It is need of time to discover non-trivial, previously unknown, and potentially useful knowledge

from such unstructured natural language documents. A Large number of ontology is needed for describing the world wide knowledge in different domains and inferring new knowledge from them.

II. EXISTING TECHNIQUES

In this paper different methods are discussed. Those methods have some advantages and disadvantages.

Dnyanesh G. Rajpathaket. al [1] have proposed to construct the D-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis. In real-life, manual construction of D-matrix diagnostic model corresponding to the complex systems is not practical as it would involve significant effort to integrate the knowledge from SMEs and represent it in a D-matrix. This approach overcame these limitations where natural language processing algorithms were proposed to automatically develop the D-matrices from the unstructured repair verbatim. They compared the testability and diagnosability metrics of the historical data-driven D-matrix and the text-driven D-matrix. They have also proposed naïve Bayes probability model for developing abbreviated terms by considering context. Their methodology for D-matrix construction consists of three building blocks document annotation, term extraction, and phrase merging.

JantimaPolpinij [2] has introducing Knowledge discovery database KDD process for finding knowledge from unstructured textual data is major problem in area of knowledge discovery in database. The problem becomes large due to ambiguity and lexical variations in natural language. To solve this problem they have present a unified methodology is called ontology based knowledge discovery in unstructured text (ON-KDT) methodology, to discover knowledge from unstructured text. This approach leverages semantic information encoded with ontology to improve the effectiveness of knowledge extraction process. A good knowledge discovery plan has to be established to achieve the project goals. The plan should be as detailed as possible that have step-by-step to perform during project including initial selection of relevant techniques and tools. ON-KDT process model deals with two main manual tasks. The first one is to determine the application domain. It includes defining of the problem and the goals of end-user. And second is to collect an initial data and proceed with activities in order to make more understanding target data and identify data quality problems.

Vishwadeepak Singh et.al [3] has described that the text mining technique for automatically extracting association rule from collection of textual documents. This technique is called as Extracting Association Rules from Text (EART).it depends on keywords feature for discover association rules. The intension is to design and devise approaches for mining potential information and association with large collection of text documents. The important approach for text mining involves use of Natural Language Processing (NLP) for information extraction. They devise approaches either by utilizing some of the effectual Information Extraction and KDD Techniques by developing innovative techniques for mining information and association amongst keywords items in text documents collection.

Shaboo Wang et.al [5] has addressed automatic hierarchical domain ontology generation from semi structured data, specifically, from HTML and XML documents. The main process of their work includes domain terms extraction, pruning, union and hierarchical structure representation they illustrate study based on Artificial Intelligence related conference data represented in HTML and XML documents. To make the implicit ontology on the Web explicit and integrate them together, in this paper, they provide an approach of building domain ontology hierarchy from semi-structured data such as XML and HTML files. They divide the whole process into two steps, domain Concept Extraction and ontology pruning and union. Implicit ontologies in the conference and workshop information are selected for investigation. They have select XML version of the DBLP data set and several workshop Web pages as data sources. They have introduced concept duplication and relevant pruning methods.

David W. Embley et.al [6] they have proposed a framework for an ontology-based system that extracts and structure information found in data-rich unstructured documents. Except for ontology creation, the processes in their framework are renovate and do not require human intervention. They also built a prototype system based on this framework. For applications that are data rich and narrow in ontology breadth, the approach presented here shows great promise. They noticed that most of the errors in recall and precision were due to incomplete lexicons and incomplete ontology without changing the framework, better lexicons and rich ontology will overcome both of these shortcomings.

A.M.Abirami [7] et.al they have proposed model for extracting resume information from different websites and make job of job recruiter easier by catching suitable resume to fit their needs. Ontology is created with suitable entities and their relationships for this domain. Each resume is split into four different sections – personal, education, skills and work experience. Attribute values extracted from resume documents. These values are updated in four different Resource Description Framework (RDF) files for each resume.

III. CONCLUSION

There are so many researches initiated in these topics to devise innovative technique for text mining different method that is used for the ontology and extracting information from the unstructured text. The main purpose of this paper is to discuss the challenges of previous methods. With the help of these methods, we can compare different method for text mining.

REFERENCES

[1] Dnyanesh G. Rajpathak, et.al, "*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] JantimaPolpinij , "*Ontology-based Knowledge Discovery from Unstructured Text*" International Journal of Information Processing and Management(IJIPM) Volume4, Number4, June 2013

[3] Vishwadeepak Singh et. al, "*Text Mining Approaches To Extracts Interesting Association Rules From Text documents*" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012

[4] SandeepSirsatet. al "*Mining Knowledge From Text Repositories using Information Extraction*" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 53–62. _c Indian Academy of Sciences.

[5] Shaobo Wang1, et. Al "*Ontology Extraction and Integration from Semi-structured Data*" International WIC Institute, Beijing University of Technology, Maebashi Institute of Technology, Japan.

[6] David W. Embleyet. al "*Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents*", Brigham Young University, Provo, Utah 84602, U.

[7] A. M. Abiramiet. al "*Ontology Based Ranking of Documents using Graph Databases: A Big Data Approach*", Dept. of Information TechnologThiagarajar College of Engineering Madurai, Tamilnadu India.

[8] DimitriosSkoutaset. al, "*Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data*", International Journal on Semantic Web & Information Systems, Volume 3, Issue 4,2007.

[9] V. VenkataSubramanianet .al, "*A review of process fault detection and diagnosis Part I: Quantitative modelbased methods,*" Comput. Chem. Eng., vol. 27, no. 3, pp. 293–311, 2003.

[10] V. Venkatasubramanian, R. Rengaswamyet. al, "*A review of process fault detection and diagnosis PartIII: Process history based methods,*" Comput. Chem. Eng., vol. 27, no. 3, pp. 327–346, 2003.

[11] Martin Meliket. al, "ontology Based Fault Diagnosis for Industrial Control Application" Automation and Control InstituteVienna University of Technology Gushausstr. 27-29/E376A-1040 Vienna, Austria.

[12] HmwayHmway Tar et. al, "*Ontology Based Concepts Weighting for text Documents*" World Academy of Science, Engineering and Technology Vol:5 2011-09-21.

[13] DonghuiFenget. al, " *Extracting data Records from Unstructured Biomedical full Text*" Information Sciences Institute University of Southern California Marina del Rey, CA, 90292.

[14] TodSedbrooket. al, "*DEAR- A New Technique for Information extraction and Context –Dependent Text Mining*" University of Northern Colorado, Monfort College of Business Communications of the IIMA © 2010 Volume 10 Issue.

[15] EhsanAsgarianet. al, "*Designing an integrated Semantic framework for Structured opinion Summarization*" Ferdowsi University of Mashhad, Mashhad, Iran.

