VOL. 10, NO. 10, JUNE 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

DISCOVERY OF RELATED E-BOOKS USING DEPENDENCY STRUCTURE MATRIX (DSM) AND LIBRARY ONTOLOGY

K. Nanthini and R. Raja Ramya
Department of Information Technology, Saveetha Engineering College, Chennai, India
E-Mail: nanthinikalanchiam@gmail.com

ABSTRACT

Ontology is an apparent prerequisite for conceptualization. It can be organized in hierarchical manner that provides whole description about domain and their class interrelationship. Now-a-days colleges maintain lot of e-books for each department because students, research scholars and staffs are preferred to use e-books. This paper mainly focuses on finding associated e-books for their source e-book. The discovery of related e-books is recognized by using ontology based text mining. First, specified pages of e-books are extracted by using java library. An extracted page contents are saved in text file. After that, content pages (tables of contents) are mined using file concept. Next employ filtering to that mined content, because that may holds special characters, chapter number and page number. Using this content page wording, identify the class of e-books ontology. Ontology is used to provide inferences about them. Furthermore perform analysis of e-book's content pages with one another and update the DSM. DSM shows whether e-books are processed or not. This DSM is converted into graph structure along with their dependency level. Dependency level specifies the map of each e-book. DSM is automatically updated when new books came to library. Graph structure also updated with their dependency levels.

Keywords: e-books, ontology, text mining, dependency structure matrix (DSM).

1. INTRODUCTION

An effective tool for complex and large systems analysis is Dependency Structure Matrix (DSM) which allows user to analyze the dependencies of system elements/entities and visualize them. It can provide proposal for system development/enrichment. DSM shows system elements in matrix format. Therefore, complex and large systems can be analyzed straightforwardly. DSM consists of rows and columns as system elements. Here, the DSM is used to analyze the books in libraries. Decompose the system into finer subsystems or modules or system elements to clearly examine the system. As a tool for system analysis, DSM captures interactions/interdependencies/interfaces between system elements [9]. Decomposition of system and dependence relationships correctness is the victory of DSM. Ontology clearly specifies the relationships between each subsystem. Therefore, DSM is constructed by using ontology for any domain in accurate way.

Text mining seeks to mine required data from unformatted textual records through the recognition of interesting patterns and an examination of interesting patterns. In this paper, table of contents of library e-books can be extracted by using text mining.

Ontology is a clear requirement of a conceptualization and it is creative process. It provides understandable description regarding complex system. Ontology is represented in hierarchical manner by decomposing the system which has relationship among them. Extensive domain knowledge is required for manual acquisition of developing ontologies. In nearly all cases, the result of ontology development could be incomplete or inaccurate [10]. Semi-automatic or automatic methods for building the ontology is used instead of manual building of ontology to overcame this disadvantage. Development of ontology need complete domain expert. No particular ontology can be developed in accurate way for any domain. Two ontologies developed by different community would not be similar one. The word ontology is applied in both a theoretical and non- theoretical context [8].

2. LITERATURE REVIEW

Satnam Singh *et al.* [13] provided a survey on how to organize D-matrices for various systems. They focus on industrial view of D-matrices that portray the advantages and disadvantages of different D-matrices types, since every D-matrix has its own syntax and set of discrete element that have high level of details to illustrate the symptoms and failure modes of system [13]. They classified the D-matrices by considering data source and symptoms imperfectness:

- Based On Sources:

1. Engineering D-matrix →EDx
2. Historical data D-matrix →HDx
3. Documents D-matrix →DDx

- Based On Symptoms Imperfectness:

1. Hard D-matrix
2. Soft D-matrix

Dnyanesh G. Rajpathak *et al.* [5] proposed an ontology based text mining method for automatic

construction of D-matrix and updating D-matrix by mining source reports. The primary purpose of D-matrix is to capture the fundamental associations between system failure modes and symptoms [13]. D-matrices are developed by analyzing historical data of failures, documents gathered during services, etc. [13]. In [5], initially they build ontology for fault analysis that consists of concepts and relations which usually occur in fault analysis domain. Subsequently identify failure modes, symptoms and their dependencies by employing the text mining algorithms that make use of fault diagnosis domain [5].

Maryam Hazman *et al*. [10] presented a survey on various methods in ontology learning from semi-formatted and unformatted data. Ontology learning refers to mining ontological elements from source and develops ontology from them. Ontology learning aim to build ontologies in semi-automatic or automatic manner from text that was mined from sources [10]. Natalya F. Noy *et al*. [12] described a methodology to develop ontology for systems and they provided the detailed procedures to develop ontology. It addresses the complicated issues in defining hierarchies of class and class properties and instances. However, No particular ontology can be developed in accurate way for any domain [12]. Ontology learning is granular of techniques and methods to build ontology from beginning or ontology enhancement or adapting a previously available ontology in semiautomatic manner [10]. The steps to develop ontology specified in [12]:

1. Find out the domain and ontology scope
2. Existing ontologies reuse consideration
3. List the terms of ontology
4. Defining classes and its hierarchy
5. Define class properties
6. Facets definition
7. Create instances

Juan C. Rendón-Miranda *et al*. [7] proposed a paper on automatic classification of scientific papers in PDF and it describes the work related to identify different sections of the paper, automatically classify and instantiate them in an ontology in order to perform inferences about them. There are several tools to extract text from PDF files. Ontology population is made to enrich the information storage and it can be used to perform inferences with the store knowledge [7]. Here scientific papers are classified based on paper elements such as: title, author, abstract, keywords.

3. SYSTEM FLOW

The following diagram describes how the related books are identified in library:

1. Extract content from library e-books
2. Using this content, represent ontology for each e-book
3. Subsequently construct DSM. The elements of DSM are library e-books. Primarily it is used to verify and represent whether the books are related or not. Represent diagonal matrix as 0, for the reason, they are same one. If books are related one represent as 1, otherwise it is represented as -
4. After that convert DSM into graph structure and specify percentage level of each book which is assigned to it
5. Finally related e-books are listed to user for their targeted book.

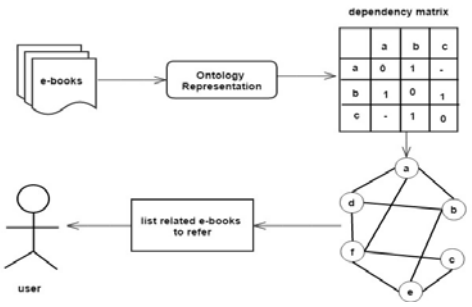

Figure-1. System architecture diagram.

4. METHODOLOGY

Proposed system consists of following methodologies to find the related e-books: content extraction, ontology representation, dependency matrix construction, graph structure representation.

4.1 Content extraction

Content extraction is searching and retrieving a subset of documents. The documents are often unstructured in nature and contain vast amounts of textual data. To mine content pages from e-books, PDFBox is used. The PDFBox library [3] is a Java library for operating with PDF files which is open source. This library performs new PDF file creation, handling of existing documents and extract data from files [3]. Extract particular pages from e-books by using this library. After extraction of pages from e-books using PDFBox library, desired content pages of each e-book extracted using file concept. Furthermore, apply text filtering to filter irrelevant terms. Because content pages of e-books can hold page number, chapter number and special character. Text filtering is primarily used to set the boundaries.

- The sub task involved in this method:

1) Extract pages from PDF
2) Mine Content Page of each Book
3) Filter Content
4) Save Content
5) Save PDF

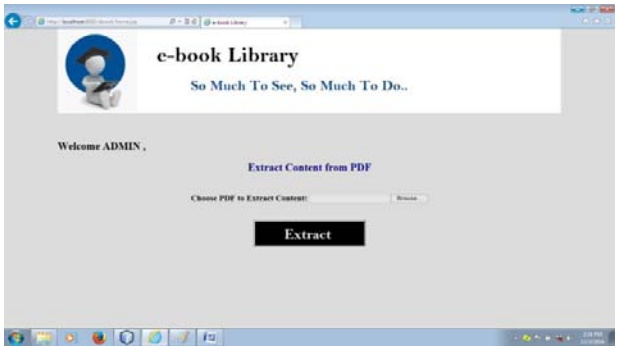

Figure-2. Home page.

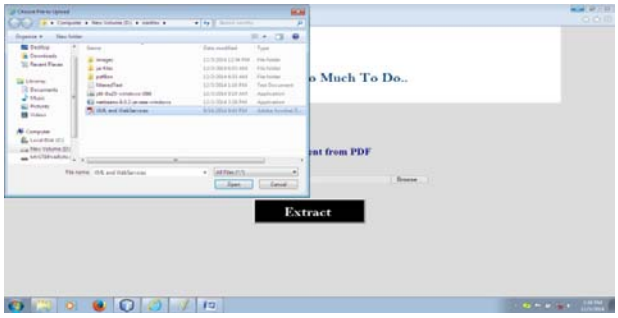

Figure-3. Content extraction page- image I.

Figure-4. Content extraction page- image II.

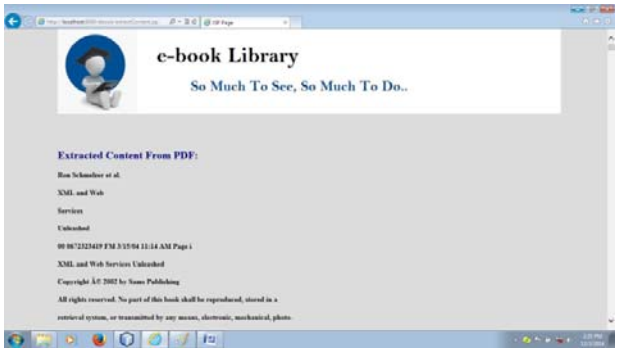

Figure-5. Content extraction page- image III.

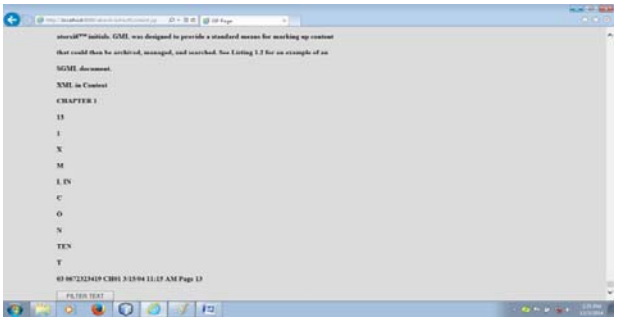

Figure-6. Content extraction page- image IV.

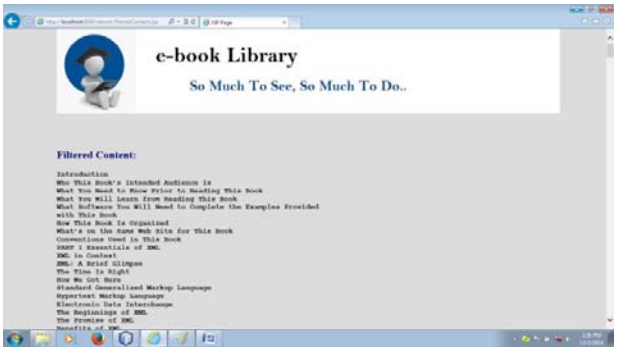

Figure-7. Content filter page.

4.2 Ontology representation

Ontology can be seen as an information model that explicitly describes various concepts that exist in a domain. Library ontology is represented by using classes that are sets, collections, concepts. Library ontology classes are departments, department related topics and department books. In addition classes can contain subclasses. Subclasses can be departments, department related topics. Classes and subclasses have their own attributes and relationships. Attributes are properties or parameters of books and relationships are way in which classes are related to one other. Classes are used to group e-books that are linked and organize these books into non-overlapping sets. Classes are identified using a set of descriptive terms. The number of classes that are identified can be controlled.

- The sub task involved in this method:

1) **Hierarchical representation:** Identify the classes of e-book ontology and specify these classes in hierarchical manner
2) **Identify each book class:** Find each book class by using descriptive terms and save in particular class

4.3 Dependency matrix construction

Dependency Structure Matrix is used to relate entities/elements of one kind to each other. Therefore it is able to analyze dependencies of elements. In library ontology, books are the elements of this matrix. The key

purpose of DSM is to specify process status of e-books that indicate whether the e-book is processed for ontology representation. Arrival of a book to the library, DSM is updated.

- The sub task involved in this method:

1) Compare page content: Compare each book content page with one another. Consequently check whether books are related or not and specify its dependency in matrix
2) Specify dependency in matrix: Diagonal matrix is represented as 0, because they are same one. Books are related means represented as 1, otherwise it is represented as –
3) Update DSM: If fresh book arriving to library, perform the above steps from content extraction to specify dependency in matrix

4.4 Graph structure representation

Graph is used for system representation. The graph consists of nodes and edges. Nodes are used to show a system element and edges are used to show the relationships between system elements by mapping the nodes. Edges are provided with dependency level between two books that is represented in percentage format.

Percentage = (number of lines in related book content page that are same as targeted book content page / number of lines in targeted book content)*100

5. CONCLUSIONS

Proposed paper is effectual in finding related/interrelated/correlated e-books in libraries by using Dependency Structure Matrix (DSM) that is constructed using ontology based text mining. Now-a-days, seek of related e-books are done manually, by comparing the content pages of desired books. This paper overcame this constraint by displaying related e-books for a targeted e-book automatically. The related e-books are listed by using DSM along with graph formation with their dependency level. Dependency level is specified in percentage format. It will be helpful for students and research scholars to find their desirable one. The continuous training of the system will provide effective results.

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to management for their invaluable guidance and encouragement.

REFERENCES

[1] Ali Yassine. 2004. ‘An Introduction to Modeling and Analyzing Complex Product Development Processes using the Design Structure Matrix (DSM) Method’, Product Development Research Laboratory, University of Illinois at Urbana- Champaign.

[2] Ali Yassine and Dan Braha. 2003. ‘Complex Problems in Concurrent Engineering and the Design Structure Matrix Method’, Concurrent Engineering Research and Applications, Vol.11, No. 3.

[3] Apache Software Foundation. 2009. PDFBox. Retrieved from http://pdfbox.apache.org/.

[4] Deepak Agnihotri, Kesari Verma and Priyanka Tripathi. 2014. ‘Pattern and Cluster Mining On Text Data’, Fourth International Conference on Communication Systems and Network Technologies.

[5] Dnyanesh G Rajpathak and Satnam Singh. 2014. ‘An Ontology-based Text Mining Method to Develop D-matrix from Unstructured Text’, IEEE Transactions on Systems, Man and Cybernetics, Vol. 44, No. 7.

[6] Jian Ma, Wei Xu, Yong-hong Sun,Efraim Turban, Shouyang Wang and Ou Liu. 2012. ‘An Ontology-based Text-mining Method to Cluster Proposals for Research Project Selection’, IEEE Transcations on Systems, Man and Cybernetics, Vol. 42, No.3.

[7] Juan C. Rendón-Miranda, Julia Y. Arana-Llanes, Juan G. González-Serna and Nimrod González-Franco. 2014. ‘Automatic Classification of Scientific Papers in PDF for Populating Ontologies’, International Conference on Computational Science and Computational Intelligence.

[8] Kent Lofgren. 2013, Feb 15. What is Ontology? Introduction to the word and the concept [Video file]. Retrieved from https://www.youtube.com/watch?v=XTsaZWzVJ4c.

[9] Lindemann. 2009. The Design Structure Matrix (DSM). Retrieved from http://www.dsmweb.org.

[10] Maryam Hazman, Samhaa R El Beltagy and Ahmed Rafea. 2011. A Survey of Ontology Learning Approaches’, International Journal of Computer Applications, Vol.22, No. 9.

[11] Milos Radovanovic and Mirijana Ivanovic. 2008. ‘Text Mining: Approaches and Applications’, Novi Sad J. Math., Vol. 38, No. 3.

[12] Natalya F Noy and Deborah L McGuinness. 2001. ‘Ontology Development 101: A Guide to Creating your First Ontology’, Standford University Knowledge Systems Laboratory Technical Report.

[13] Satnam Singh, Steven W Holland and Pulak Bandyopadhyay. 2010. ‘Trends in the Development of System Level Fault Dependency Matrices’, IEEE Conference.