ISSN 2454-8596
www.vidhyayanaejournal.org
An International Multidisciplinary Research E-Journal

Approaches for Text Mining using Ontology

Atish M Shah

Department of Computer Science

S.S.Agrawal College of Arts,Commarce & Managment ,Navsari

Veer Narmad South Gujarat University, Surat ,India

ISSN 2454-8596
www.vidhyayanaejournal.org
VIDHYAYANA
An International Multidisciplinary Research E-Journal

Abstract—Extraction of information from the unstructured document depending on an ontology application describes domain of interest which is presented as a new approach. To start with such ontology, we formulate rules to extract constants and context keywords from unstructured documents. For every unstructured document of interest, constants and keywords are extracted and a recognizer is applied to organize constants which are extracted as attribute values of tuples in a database schema generated. To make approach general, all the process is fixed and only ontological description is changed according to different application domain. In this paper, we are describing on two different types of unstructured document: firstly as offline which is based on specific PDF document and secondly as online which is Web-based and our approach attained recall scale in 80 percent and 90 percent range and accuracy near 98 percent.

Keywords—unstructured document; information structuring; information extraction; ontology.

I. Introduction

During the last few years, the amount of data is available on the Web has been grown explosively. Users can retrieve data by browsing or keyword searching, which is useful or logical but limitation are there for access. Browsing is not suitable for locating particular items of data and also not cost effective as users to read the document to find the desired data. Keyword searching is better than browsing but in return it provides more amount of data which is not handled by user. To retrieve data efficiently from the web, researchers have taken ideas from database techniques where databases means structured data is required. Until now data available on the web is unstructured data. Various approaches have been suggested for querying the Web which falls into one of the two categories: generating wrappers for web pages and querying the web with web query languages.

A relation in a structured database can be expressed by set of n-tuples and each n-tuples associates n attribute-value pairs in a relationship. This relationship establish the information supposed by the relation. A well chosen n-place predicate for the relation can make this information easily understandable to humans. An unstructured document does not contain this structuring characteristic. There are no relations with associated predicates, no attribute value pairs and no n-tuples. Similarly, there is no information supposed by any relation about the contents of an unstructured document. It is possible and useful to set structure by establishing relations over the information contents of the document. In such situation, establishing relation automatic is more beneficial. This paper presents an automatic approach to extract information from unstructured documents and reformulating information as relations in a database.

Our approach is being based on ontology. Ontology is a branch of philosophy that attempts to model things as they exist in the world; it is particularly appropriate for modeling objects including their relationships and properties. Using an augmented semantic data model gets an ontology which will describe the view as per domain of interest. The semantic data model allows creating ontological model instance which is a set of objects, of relationships among these objects and constraints over these objects. As augmented, it defines data representation and expected contextual keywords for each object set within the ontology. Application ontology based with this characteristics we apply a parser, a constant keyword recognizer and a structured text generator to filter unstructured document with respect to the ontology and populate a generated database schema with attribute-value pairs associated as relations. Thus, the interested information is extracted from an unstructured document and reformulates it as a structured document.

For all unstructured documents this approach is not expected to work well. However, expected the approach to work well for unstructured documents if data rich and narrow in ontological breadth and containing information of multiple records for the ontology. A document is *data rich set* if it has a number of identifiable constants such as dates, names, ID numbers, currency values, and so on. A document is *narrow in ontological breadth* which describes its application domain with a relatively small ontological model. A document contains multiple records for ontology if there is a sequence of chunks of information about the main entity in ontology. Not all of these definitions are exact, but they express the idea that the kinds of Web documents considered have many constant values, are narrow in the domain they cover, and contain descriptions for several object instances that satisfy the ontology.

To test these ideas as case studies for this paper, we consider newspaper job listings for computer-related jobs and advertisements for automobiles. Both automobile ads and job listings are data rich and narrow in ontological breadth and contain multiple records . Automobile ads typically include constants for and information about year, make, model, price, mileage, features, and contact phone numbers. Computer job listings include degree required, needed skills, and contact information, knowledge related field and general also. Other application areas whose documents have similar characteristics include travel agency, financial transactions, scheduling for meetings, stocks, sports information, genealogy, medical research, product information, and many others.

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.

As the textual information available in electronic form, a large research effort has switched in the database community for finding ways to make Web querying more powerful than browsing and keyword searching. Nowadays efforts have taken several directions including virtual database technology, semi-structured data, Web data modeling, wrapper generation, natural-language processing (NLP) and Web queries.

Our research reported here relates to recent efforts in several areas including Web data modeling, wrapper generation, natural-language processing, semi-structured data, and Web queries. Others have used semantic data models to describe Web documents and populate databases. 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.

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 column names and rows 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 or web page 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.

III. System Architecture

A. Extraction and Structuring Document Framework

The extraction and structuring data from an unstructured document is as shown in Figure 1. Boxes are represented as a files and ovals as process. According to Figure 1, the input to approach is application ontology and an unstructured document, and the filtered and structured document whose data is in a database is provided as output. In advance all the processes and intermediate file formats are fixed, Figure 1 depicts a general process that takes as input any declared ontology for an application domain of interest and an unstructured document within the application's domain and produces as output structured data, filtered with respect to the ontology. For human interfaces the important step required is the initial creation of application ontology.

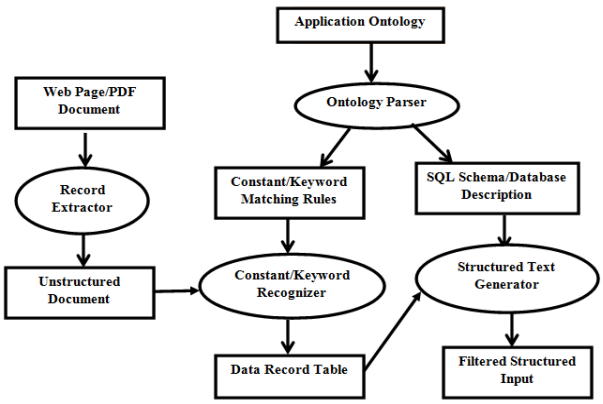


Figure 1: Extraction and Structuring Document Framework

As shown in Figure 1, there are four main processes in framework: an ontology parser, a constant/keyword recognizer, a structured text generator and a record extractor. The input is application ontology and unstructured document which is extracted through web page/PDF document and the output Is filtered

structured document. The main program invokes parser recognizer and generator sequentially.The ontology parser is invoked only once at the beginning of execution, whenever the recognizer and generator are invoked repeatedly in sequence for each unstructured document which to be processed. Constants are potential values i.e. lexical object sets while context keywords are associated with any object set either lexical or non lexical so it is possible to present ontology textually.

Recently we are passing specific document or website manually to the record extractor which automatically removes the HTML tags and seperates the input document into unstructured document. Further ontology parser is invoked which creates an SQL schema as a sequence of create table statements for a given application ontology. All information is not needed to the structured-text generator, so parser can extracts only the relevant information like list of objects, constraints and relationships to be used by the generator. A mapping is provided between the table declarations in the SQL schema and the relationships in the ontology. It also provides the cardinality relationship constraint which can be one-one, one-many, and many-many. And the parser also creates a file of constant/keyword matching rules which further passed to constant/keyword recognizer. Then data record table creates a data according to table and provided to structured-text generator. Finally, the structured-text generator process creates a structured file output.

B. Mathematical model

Let S be a system which extracts information from the unstructured documents depending on an ontology application.

Such That S= {I, F, O} where,

I represent the set of inputs:

I= {D, W}

D= Set of Input Pdf document i.e. Unstructured document.

W= Total Methods for retrieving Structured Data.

F is the set of functions:

F= {T, F, M}

T= Pdf document validation

ISSN 2454-8596
www.vidhyayanaejournal.org
An International Multidisciplinary Research E-Journal

F= Parsing the document into Xml

M= Threshold Comparison

O is the set of outputs:

O= {C}

C= Retrieved Structured Data.

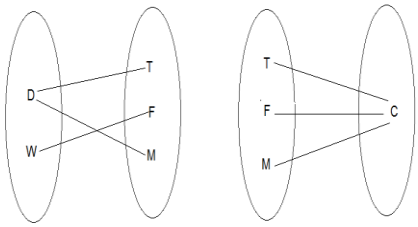


Figure 2: Venn diagram

IV. Implementation Details

A. Modules

There are two modules in our project as follows:

1) Structuring of data from unstructured specific PDF file.

2) Get structured data from Webpage like Flip-kart, Amazon, and eBay and so on.

1) **Structuring of data from unstructured PDF file.**

First user 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 like sorting, searching etc.

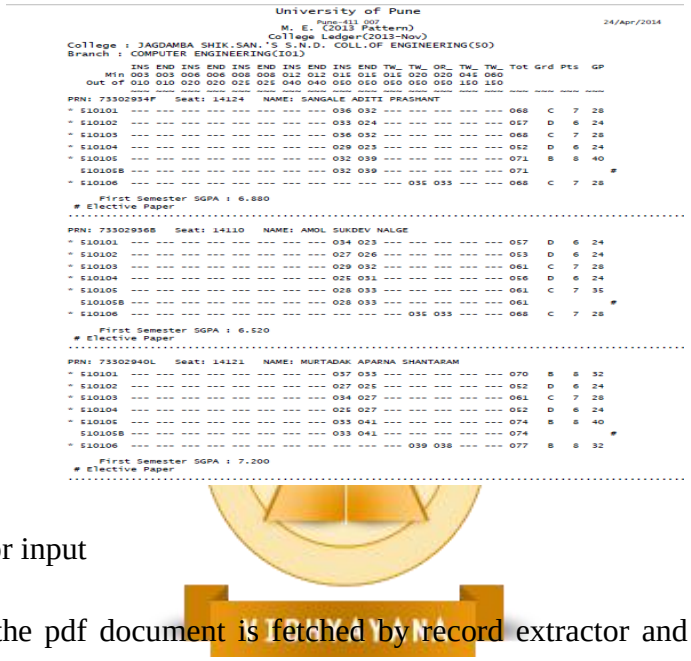
2) **Get structured data from Webpage**

User passing web URL in text box as an input parameter. After getting URL we are validating URL

with null URL or URL name which is passed. 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 like sorting, searching etc.

V. Results

A PDF document is passed as an input which contains a mark sheet of M.E students.



University of Pune

Pune-411 007 24/Apr/2014

M. E. (2013 Pattern)

College Ledger(2013-Nov)

College : JAGDAMBA SHIK.SAN.'S S.N.D. COLL.OF ENGINEERING(SO)

Branch : COMPUTER ENGINEERING(IO1)

	INS	END	INS	END	INS	END	INS	END	TW_	TW_	OR_	TW_	Tot	Grd	Pts	GP
Min	003	003	006	006	008	008	012	012	015	015	020	020	045	060		
Out of	010	010	020	020	025	025	040	040	050	050	010	010	050	150	150	

PRN: 73302934F Seat: 14124 NAME: SANGALE ADITI PRASHANT

*	510101	---	---	---	---	---	---	---	---	036	032	---	---	---	068	C	7	28
*	510102	---	---	---	---	---	---	---	---	033	024	---	---	---	057	D	6	24
*	510103	---	---	---	---	---	---	---	---	036	032	---	---	---	068	C	7	28
*	510104	---	---	---	---	---	---	---	---	029	023	---	---	---	052	D	6	24
*	510105	---	---	---	---	---	---	---	---	032	039	---	---	---	071	B	8	40
	510105B	---	---	---	---	---	---	---	---	032	039	---	---	---	071			#
*	510106	---	---	---	---	---	---	---	---	---	---	035	033	---	068	C	7	28

First Semester SGPA : 6.880

Elective Paper

PRN: 73302936B Seat: 14110 NAME: AMOL SUKDEV NALGE

*	510101	---	---	---	---	---	---	---	---	034	023	---	---	---	057	D	6	24
*	510102	---	---	---	---	---	---	---	---	027	026	---	---	---	053	D	6	24
*	510103	---	---	---	---	---	---	---	---	029	032	---	---	---	061	C	7	28
*	510104	---	---	---	---	---	---	---	---	025	031	---	---	---	056	D	6	24
*	510105	---	---	---	---	---	---	---	---	028	033	---	---	---	061	C	7	35
	510105B	---	---	---	---	---	---	---	---	028	033	---	---	---	061			#
*	510106	---	---	---	---	---	---	---	---	---	---	035	033	---	068	C	7	28

First Semester SGPA : 6.520

Elective Paper

PRN: 73302940L Seat: 14121 NAME: MURTADAK APARNA SHANTARAM

*	510101	---	---	---	---	---	---	---	---	037	033	---	---	---	070	B	8	32
*	510102	---	---	---	---	---	---	---	---	027	025	---	---	---	052	D	6	24
*	510103	---	---	---	---	---	---	---	---	034	027	---	---	---	061	C	7	28
*	510104	---	---	---	---	---	---	---	---	025	027	---	---	---	052	D	6	24
*	510105	---	---	---	---	---	---	---	---	033	041	---	---	---	074	B	8	40
	510105B	---	---	---	---	---	---	---	---	033	041	---	---	---	074			#
*	510106	---	---	---	---	---	---	---	---	---	---	039	038	---	077	B	8	32

First Semester SGPA : 7.200

Elective Paper

Figure 3: Sample data for input

After passing an input the pdf document is fetched by record extractor and gets converted into XML. Then after extracting data is in unstructured format which is passed to constant/keyword recognizer. Then according to recognizer data record table is formed and forwarded to structured-text generator and finally the structured output is generated in the form of D-matrix.

ID	Name	PRN Number	Seat Number	SGPA	Branch Name
77	AVHAD SWATI SARANGDHAR	73303010G	14111	8.04	COMPUTER ENGINEERING(I01)
55	GUNDAP SANTOSH NARAYAN	73303002F	11972	8	ELECTRICAL (POWER SYSTEMS)(D01)
92	REVGADE RATNA BHANUDAS	73302992C	12653	8	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
91	RANDHAWANE PRAJAKTA DILIP	73302985L	12652	7.84	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
60	BABAR SWAPNIL DATTATRAY	73303009C	11965	7.8	ELECTRICAL (POWER SYSTEMS)(D01)
89	NIKAM BHARAT VIKRAM	73302971L	12648	7.68	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
90	PATIL POOJA PRAKASH	73302981H	12651	7.64	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
80	LACHAKE VIKRAM GAJANAN	73303018B	14120	7.52	COMPUTER ENGINEERING(I01)
88	PATIL MANISHA ASHOK	73302965F	12650	7.52	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
96	VIKHE PRATIMA SOPAN	73303019L	12654	7.48	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
85	BAKLE BHAKTI BANDOPANT	73302943E	12641	7.48	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)
79	VAIDYA VIJAYSHRI DATTATRAY	73303014K	14126	7.36	COMPUTER ENGINEERING(I01)
56	GAMBHIRE SUPRIYA MADHUKAR	73303004B	11971	7.36	ELECTRICAL (POWER SYSTEMS)(D01)
94	KULKARNI TRUPTI DEEPAK	73303011E	12646	7.32	ELECTRONICS & TELECOMMUNICATION (VLSI & EMBEDDED S(F04)

VI. Conclusion

In our paper, a novel 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. In actual life, the manual construction of a D-matrix diagnostic model corresponding to the complex systems is not practical as it would involve significant effort to integrate the knowledge and represent it in a D-matrix.

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. However, much remains to be done. Three particular tasks lie ahead: (1) improve and fine-tune the implemented procedures, (2) add front-end page processors, and (3) diversify back-end display generators.

ACKNOWLEDGEMENT

It is with the greatest pleasure and pride that I present this paper before you. At this moment of triumph, it would be unfair to neglect all those who helped me in the successful completion of this paper presentation. I am very much thankful to my respected project guide Prof. I.R. Shaikh, Computer Engineering Department, for his ideas and help proved to be valuable and helpful during the creation of paper presentation and set me in the right path. I am thankful to my friends who shared their knowledge in this field with me.

REFERENCES

[1] Dnyanesh G. Rajpathak, Satnam Singh, Member “An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text,” IEEE Transactions on System, Man and Cybernetics System, Vol.44,No.7, July 2013.

[2] M.Schuh, J.Sheppard and C.Izurieta, “Ontology-guided knowledge discovery of event sequences in maintenance data," IEEE AUTOTESTCON Conf., vol. 7, no. 5, Mar. 2011.

[3] M.Gaeta, F. Orciuoli 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, 2011.

[4]AM. Schuh, J. Sheppard 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, 2013.

[5] S. Singh and C. Pinion, “Data-driven framework for detecting anomalies in eld failure data," IEEE Aerosp. Conf., vol. 7, no. 5, Apr. 2011.

[6]W.Zhang, T. Yoshida and Q. Wang, “Text clustering using frequent item sets,” Knowledge.-Based System, vol. 23, no. 5, pp. 379-388, 2010.

[7] J. Sheppard, M. Kaufman, and T. Wilmering, Model based standards for diagnostic and maintenance information integration, in Proc. IEEE

AUTOTESTCON Conf., 2012, pp. 304310.

[8] Berners-Lee, T, Hendler, J, Lassila, O.: The Semantic Web, Scientific American ; 2001.

[9] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction:

An introduction and a survey of current approaches. Journal of Information Science, 2010, 36(3): 306.

[10] P. Cimiano. Ontology Learning and Population from Text: Algorithms,

Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus,

NJ, USA, 2006. ISBN 0387306323. 15, 66, 67, 72

[11] B. Popov, A. Kiryakov, D. Ognyano, D. Manov, and A. Kirilov. KIM

A semantic platform for information extraction and retrieval. Natural

Language Engineering, 10 (3-4):375-392, 2004. 65, 68

[12] Alexander Maedche2, Gunter Neumann1, Steffen Staab Bootstrapping

an Ontology-Based Information Extraction System. In: Studies in Fuzziness and Soft Computing, IntelligentExploration of the Web, Springer, 2002