Integration of Healthcare Ontologies at Schema Level using Customized Metadata

Monika P., G. T. Raju

Abstract: ***In today's fast growing competitive world, Data mining has become a research area of great interest as the problem of handling data in many circumstances toss lot of opportunities for research discoveries. Data being generated every second particularly in healthcare sector need to be managed efficiently so that further perusal when needed will be easier for medical professionals and researchers as an aid of decision support. Heterogeneity in the structure of data rather than the semantic discovery is the key of open challenge remained yet unaddressed. Structural construct deals at schema level of data depiction. Ontologies as means of data representation in the form of knowledge graphs are serving the field of Machine Learning (ML) from decades supporting automated knowledge extraction. Lot of research contributions are found to handle general formats to certain extent, but handling images and Portable Document Format (PDF) remain open as a major problem statement to be addressed in-order to enjoy successful information retrieval benefits. However not all relevant data is being retrieved during semantic queries due to non-homogeneity in data representation at the schema level resulting in ruling out of the document matches. In order to address the stated issue, an approach has been presented in the paper which aims at extracting metadata about the documents facing problem of heterogeneity, constructing ontologies based on the customized metadata tags followed with integration of ontologies for enhancing the prediction accuracy by increasing the relativity of documents in the semantic context. The proposed methodology can be evaluated using any of the classification techniques and solutions proved worth can be retained for daily access of semantic information thereby achieving good prediction accuracy in the process of efficient knowledge recovery.***

Keywords: Semantic web, Ontologies, Ontology agents, Ontologies Integration, Health care, Schema.

I. INTRODUCTION

In the present competitive world of machine intelligence where machines perform equivalent to humans realizing the task of automated decisions to certain extent; representing and understanding the under lying information plays a major role. The web which is being updated consistently in the exponential rate is loaded with heterogeneous data pertaining to various domains. As per the experts' review, Healthcare is the highly rated domain generating data every second. The phase of disease diagnosis created diverse data representations including handwritten records, laboratory reports, scan summaries & images, detailed history documents etc., in various formats like Joint Photographic Experts Group (JPEG), PDF, .excel, .doc etc.,. Presenting all of these data formats in machine readable format is the greatest challenge remaining unanswered till date.

Several frameworks exist in literature to convert the current information available in the web to machine understandable presentation. Ontological representation is one such well known format, which makes use of triple store (*subject, predicate, object)* pairs to build Knowledge Graphs (KG). Ontologies can be constructed manually using the editors like protégé [1] with the features extracted from the data as components of triple stores though automated construction options are also freely accessible. Lot of Machine Learning (ML) algorithms including supervised, semi-supervised and un-supervised techniques are available to explore the semantics present in the graphical structures with good prediction accuracies depicting the environment of Semantic Web (SW). However it's practically very difficult to represent data of different formats into a single representation like onto-graphs.

Accessing data present in heterogeneous representations by machines for automated conclusions poses complicated or NP hard solutions which are very much difficult to implement effectively to satisfy the application needs. In the field of data mining with ML concepts, the information incompatible with ontological depictions or metadata tagging remain unused resulting in incomplete relevant data extraction particularly at the schema level. Integration of heterogeneous information repositories serves as solution for information sharing and reuse, resulting in relevant data retrieval as results through semantic query searches. Broad categories of data incompatibility levels include semantic, structural and syntactic discrepancies, which has to be addressed to realize good prediction accuracies during knowledge mining.

To achieve interoperability amongst ontologies, integration should be carried out at both data and schema levels. Data level integration involves information abstraction, Meta tagging, aggregation and grouping which helps achieve compatibility between various data contents. Schema level integration deals with the basic representation of the data particularly upon the structure of the data rather than its semantics. Ontology usage suites well in semantic web for information retrieval using SPARQL queries and representing information about the data in the form of schema descriptions in terms of ontological relationships.

Revised Manuscript Received on December 05, 2019.
***** Correspondence Author

Monika P.*, Research Scholar, Department of CSE., RNS Institute of Technology, Assistant Professor, Department of CSE, Dayananda Sagar College of Engineering, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India. Email: monikamanjunath@gmail.com

G. T. Raju, Professor, Department of of CSE, RNS Institute of Technology, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka India. Email: gtraju1990@yahoo.com

In order to reach good percentage of data reuse at the schema level, a procedural approach has been presented which follows the process of constructing ontologies based on the metadata extracted from the data analysis technique employed. Resultant ontologies can be further integrated to form knowledge graphs supporting enhanced accuracy of the semantic search results with reduced redundancy.

The rest of the paper is organized as follows: section 2 describes the techniques proposed by other researchers towards metadata creation and handling for efficient knowledge extraction from the raw schematic representation along with ontologies usage. In section 3, the proposed procedural approach working on extraction of semantic knowledge from documents through metadata and ontology tagging followed with integration of ontologies for prediction accuracy enhancement supporting efficient information reuse has been briefed in detail. Section 4 elaborates on Ontologies integration approach and various evaluation metrics that can be employed to view the correctness of the proposal. The sample input documents; extracted metadata samples, intermediate ontologies constructed and sample cosine similarity computation procedure are detailed in section 5. Section 6 concludes, followed with references.

II. RELATED WORK

Semantic awareness of the web services is very much popular in the present era of Artificial Intelligence. The concepts of machine understanding and responding with wise decisions as humans, involve the technology of Machine Learning with the supervised, semi-supervised and unsupervised learning algorithms. In the present web with heterogeneous data formats achieving complete automation is far stated task which is most important in the field of medicine. Ontologies as easy knowledge representation source serves automated knowledge discovery task to certain extent. Wikipedia is a big knowledge hub working on semantic automations with the support of schema level at the underlying data structures. DBpedia [2] is a well-known implementation of a knowledge graph tailed to extract famous Wikipedia retrieval boxes. Another similar construct to DBpedia is YAGO [3] – 'Yet Another Great Ontology' which also works on extracting the wiki database through the knowledge graphs operating on metadata representations. Many such good implementations are available in the store for efficient information retrieval as per the user needs.

Couple of data sources fails to ever get revisited for knowledge discovery due to lack of semantics represented in machine understandable form. There is a need to integrate such data with the structured contents to explore knowledge to maximum extent. A biomedical knowledge base called *KnowLife* [4] attempts to integrate the unstructured data sources. Lot of ready services called as look up services working on ontology [5] are also available for the researchers to explore on such automations at the schema level particularly working on meta-tags of the operating documents. Linked Open Vocabularies [6] support exploration across domains operating on ontologies and their interactions. Lot many researchers are contributing towards improving the interoperability of the biomedical ontologies [7] as it is very much important for accurate decisions during critical conditions. In order to achieve the listed drawbacks, there is need for the refinement of the existing knowledge graphs [8].

The other part of survey describes that the exploration of metadata of the data being processed will be beneficial for achieving better accuracy of data retrievals [9]. Metadata with cloud based systems promise fastest data retrieval rates [10]. From the detailed survey it is understood that the ontologies coupled with good metadata knowledge services could perform well during knowledge extraction automations. Hence a customized procedural approach has been proposed to extract the hidden semantics, represent the knowledge explored in the Onto-graph format followed with integration of ontologies at the schema level for results with better prediction accuracies in the domain of healthcare.

III. PROPOSED PROCEDURAL APPROACH

Schema level being the basic level of data representation with elementary formatting in a semantic web plays major role during information extraction in an automated environment. Researchers have contributed on automated tools for representing most general formats in machine readable form. But extracting information from images and PDF formats have not been addressed more in the literature. Ontologies support the process of mining automation through ML algorithms to maximum extent. To leverage these benefits into the healthcare applications addressing the stated issues, a procedural approach designed on metadata contents: Metadata based Ontologies Integration at Schema Level (MOIS) as in fig.1 has been proposed. The algorithmic representation of the same is depicted in fig. 2. The suggested approach promises the semantic interoperability in a heterogeneous environment irrespective of domains. Input is expected to be in the form of JPEG or PDF. General formats of input documents include images, clinical observations which are handwritten and printed diagnostic reports.

The unstructred data should be processed to represent in a structured format for achieving automated decisions. Hence Metadata Extraction Algorithm (MEA) procdure as in fig. 3 has been coined. The corresponding algorithmic steps are presented in fig. 4. The MEA aims at utilizing the python library functions and standard Application Programming Interfaces (APIs) like Python Imaging Library (PIL), OpenCV-python, regular expressions (re), pdf2image, Optical Character Recognition to Portable Document Format (OCRmyPDF) etc., neccesary for the intemediate conversions and representations. Given input formats will be handled for PDF and image input format verification. PDF data formats will be preprocessed to enhance the image quality. Structural extracts named tabular data (D_t) and descriptive data (D_p) are future tempered for exposure of textual data based on Opitical Characteer Recognition (OCR) technique. Google Vision APIs play a major role in extraction of meaningful text from the paragraph isolated from tabular data. By applying regular expression for pattern matching feature, set of data extracted will be formatted into the form of meta tags.

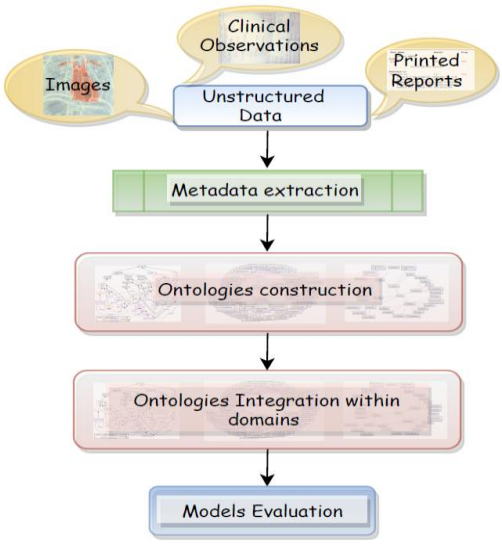


Fig. 1. Architecture of Metadata based Ontologies Integration at Schema Level (MOIS)

Algorithm: Metadata based Ontologies Integration at Schema level (MOIS)

Input: Unstructured data in PDF and image formatted documents
Output: Integrated Ontology with better prediction accuracy achieved at schema level

1. Start processing the documents
2. Let *M* be number of Unstructured data given in the form of documents
3. $\forall m \epsilon M$ ***do***
4. $n = MEA(m)$
5. Ontology construction for every metadata set at *n* using protégé
6. Integrate all ontologies based on cosine similarity index computed between the pair
7. Evaluate the Models for correctness of working
8. Return the integrated ontology with better performance at schema level
9. End

Fig. 2. Metadata based Ontologies Integration at Schema level (MOIS) algorithm

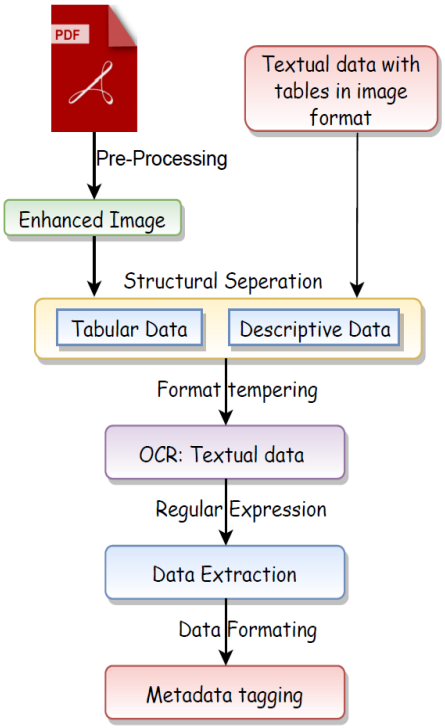


Fig. 3.Metadata Extraction Algorithm (MEA) procedural architecture

Algorithm: Metadata Extraction Algorithm (MEA)

Input: Document (D) to be processed in .pdf or .image format
Output: Metadata tagging (M) of the document texts for semantic processing

1. Start processing the document (D)
2. ***if*** input document *(D)* in image format, **then** proceed to step 3
3. ***if*** input document (D) in PDF format, **then** preprocess it to image format
4. Apply structural separation on D to segregate tabular (D_t) and descriptive (D_p) data
5. Extract textual data (D_x) from D_t and D_p
6. Apply Regular Expression pattern match to extract useful information present in D_x
7. Form key-value pairs and tabulate the meta data tags (M)
8. Return the set of Metadata tags (M) extracted
9. End

Fig. 4. Metadata Extraction Algorithm (MEA)

With the list of metadata extracts of each document, individual ontologies can be created using standard ontolgy editors like protégé [1] from stanford university. Further ontolgy integration using cosine similarity index between the terms of similar ontologies can be experimented inorder to achieve optimal semantic heteroginity resulting in efficient knowledge graph construction. The relationships can be framed with the use of predicates like issubclassof, causes, etc., depending on the tags dicovered during metadata extraction phase. The so far created models can be evaluated for suitabilities using standard ontology evaluation metrics and the results can be observed for future proceedings. The proposed methodolgy is predicted to efficienlty retrieve the semantics of data which are with hidden tabular formats thereby supporting better semantic interoperability compared to exisiting tecniques.

IV. ONTOLOGIES INTEGRATION AND EVALUATION METRICS

A. Evaluation of Metadata extraction

Metadata extraction in the proposed solution operates on the hidden tabular format embedded in the test documents retrieving the semantics successfully. Comparative observations can be done with similar services like *Amazon textract* which has inbuilt functions to process the images read in the text format. However the later method miss to extract the match pairs from the documents with hidden formatting like tables etc. as its services are defined to operate in a key value pair of any tabular structures or textual formats only.

B. Cosine similarity evaluation for integration of ontologies at schema level

The customized ontology constructs for supporting efficient knowledge retrieval follows the path of integration of couple of ontologies at schema level. As schema level deals with structural paradigms more than the semantic constructs, to achieve better prediction results – cosine similarity measure (1) can be applied. The cosine observations nearer to 1 show better match between the documents being compared.

$$\text{sim}(D1, D2) = \frac{D1.D2}{||D1||D2|} \quad (1)$$

Where *D1* and *D2* are feature vectors of 2 different ontologies which are framed in-turn from two different documents under analysis.

$\|D1\|$ is Euclidean norm of vector D1 = $\{D1_1, D1_2, \ldots, D1_p\}$ defined as

$$D1 = \sqrt{D1_1^2 + D1_2^2 + D1_3^2 + \cdots + D1_p^2} \quad (2)$$

Similarly, $\|D2\|$ is Euclidean norm of vector D2 = $\{D2_1, D2_2, \ldots, D2_p\}$ defined as

$$D2 = \sqrt{D2_1^2 + D2_2^2 + D2_3^2 + \cdots + D2_p^2} \quad (3)$$

The pair of documents with close cosine observations will be suitable for integration. Ontologies integration can be realized by deciding upon the features chosen for integration and establishing a triple representation in the Onto-graph representation at the schema level with the new predicate created or reusing the existing once.

C. Resultant ontology evaluation metrics

The integrated ontology, which is ready for semantic usage through ML algorithms, can be evaluated to check its prediction accuracy using any of the classification algorithms like Naïve Bayes theorem or Neural Networks concepts which are well known for handling discrepancies in the field of medicine. The tabulated results of prior and post integration operations justified with the metrics like precision (4), recall (5) and accuracy (6) provides visual clarity on the model correctness.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

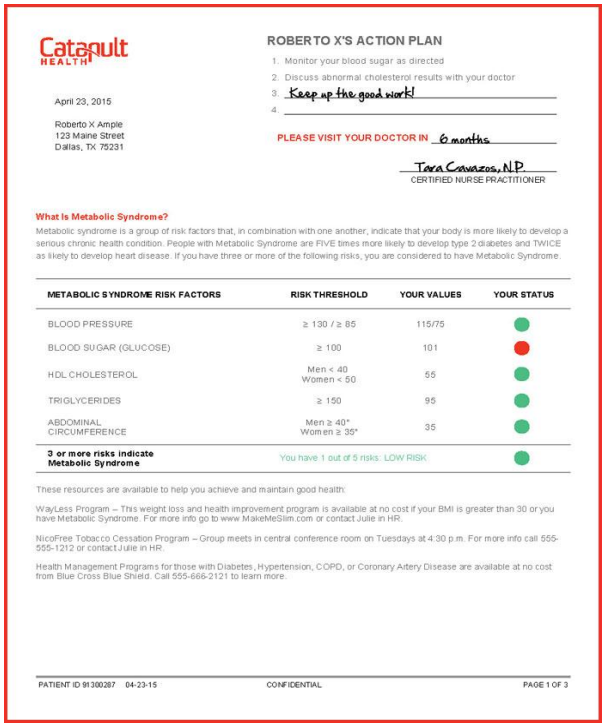
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative measures of the prediction test samples.

V. RESULTS AND DISCUSSIONS

The proposal operates on the input documents of type PDF or JPEG considered as unformatted data sources. The process of metadata extraction results in the tabulated list of key-value pairs describing the contents of the document being processed. Fig. 5 and Fig. 7 illustrate the sample input documents of blood test report in PDF and JPEG formats. Table I and table II documents the corresponding metadata tags generated.

A total of 14 and 18 important features were extracted from blood test reports provided in the form of PDF and JPEG which were fed as input to the metadata extractor. Following the list of features being considered as the attributes, Ontology construction process triggers generating a customized ontologies as depicted in fig. 6 and fig. 8.


Catapult HEALTH

ROBERTO X'S ACTION PLAN

1. Monitor your blood sugar as directed
2. Discuss abnormal cholesterol results with your doctor
3. Keep up the good work!
4.

April 23, 2015

Roberto X Ample
123 Maine Street
Dallas, TX 75231

PLEASE VISIT YOUR DOCTOR IN 6 months

Tara Cavazos, N.P.
CERTIFIED NURSE PRACTITIONER

What Is Metabolic Syndrome?

Metabolic syndrome is a group of risk factors that, in combination with one another, indicate that your body is more likely to develop a serious chronic health condition. People with Metabolic Syndrome are FIVE times more likely to develop type 2 diabetes and TWICE as likely to develop heart disease. If you have three or more of the following risks, you are considered to have Metabolic Syndrome.

METABOLIC SYNDROME RISK FACTORS	RISK THRESHOLD	YOUR VALUES	YOUR STATUS
BLOOD PRESSURE	≥ 130 / ≥ 85	115/75	
BLOOD SUGAR (GLUCOSE)	≥ 100	101	
HDL CHOLESTEROL	Men < 40 Women < 50	55	
TRIGLYCERIDES	≥ 150	95	
ABDOMINAL CIRCUMFERENCE	Men ≥ 40" Women ≥ 35"	35	
3 or more risks indicate Metabolic Syndrome	You have 1 out of 5 risks: LOW RISK		

These resources are available to help you achieve and maintain good health:

WayLess Program – This weight loss and health improvement program is available at no cost if your BMI is greater than 30 or you have Metabolic Syndrome. For more info go to www.MakeMeSlim.com or contact Julie in HR.

NicoFree Tobacco Cessation Program – Group meets in central conference room on Tuesdays at 4:30 p.m. For more info call 555-555-1212 or contact Julie in HR.

Health Management Programs for those with Diabetes, Hypertension, COPD, or Coronary Artery Disease are available at no cost from Blue Cross Blue Shield. Call 555-666-2121 to learn more.

PATIENT ID 91300287 04-23-15 CONFIDENTIAL PAGE 1 OF 3

Fig. 5. Sample blood test report in PDF form

Table- I: Sample Metadata tags extracted from PDF

#	Key	Value
1	Patient Name	Roberto X Ample
2	Doctor Name	Tara Cavazos N P
3	Address	123 Maine Street
4	City	Dallas
5	State	TX
6	Code	75231
7	BLOOD PRESSURE	115/75
8	BLOOD SUGAR (GLUCOSE)	101
9	HDL CHOLESTROL	55
10	TRIGLYCERIDES	95
11	ABDOMINAL CIRCUMFERENCE	35
12	Risk status	BLOOD SUGAR
13	Risk Indicator	you have 1 out of 5 risks: LOW RISK
14	PATIENT ID	91300287

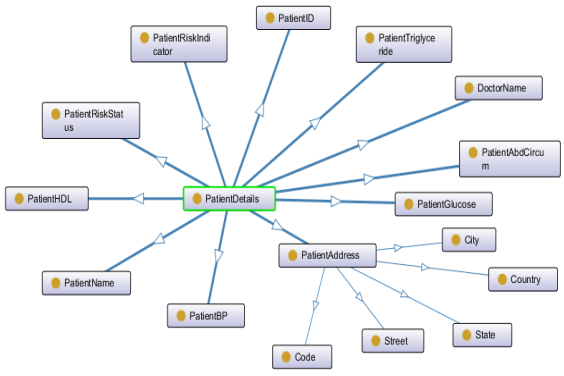


Fig. 6. Patient Ontology Skeleton of the analysed PDF form

The class structure of the patient ontologies consist of the extracted metadata tags as class labels. Address class has subclasses detailing the address splitup. Object property assertions may be includes as hasName, hasHDL, hasGlucose, hasRisk, riskLeadsTo etc., to connect the classes in the tripple store. Defining the object assertions with the properties like sameAs, subPropertyOf, equivalentTo etc., help discribing the similarities among rest of the ontologies.

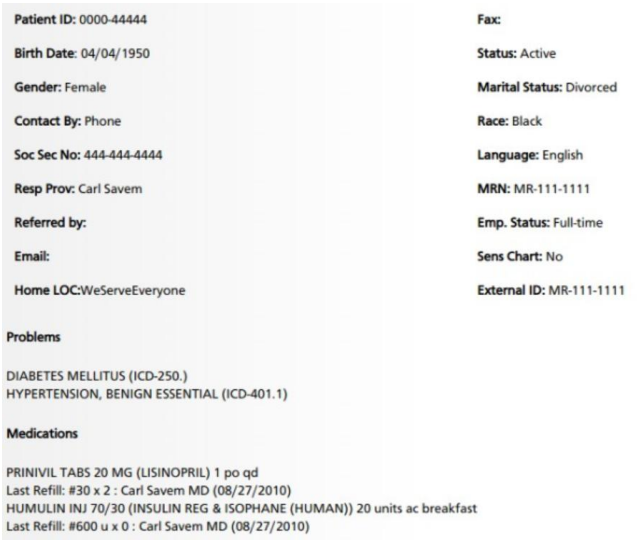


Fig. 7. Sample blood test report in JPEG form

Table-II: Sample Metadata tags extracted from image

#	Key	Value
1	Patient ID	0000-44444
2	Birth Date	4/4/1950
3	Gender	Female
4	Soc Sec No	444-444-4444
5	Resp Prov	Carl Savem
6	Status	Active
7	Marital Status	Divorced
8	Race	Black
9	Language	English
10	MRN	MR-111-1111
11	Emp. Status	Full-Time
12	Sens Chart	No
13	External ID	MR-111-1111
14	Problem	DIABETES MELLITUS
15	Problem	HYPERTENSION
16	Medications	PRINIVIL TABS 20 MG (LISINOPRIL) 1 po qd
17	Medications	HUMULIN INJ 70/30 (INSULIN REG & ISOPHANE (HUMAN)) 20 unit ac breakfast
18	Contact By	Phone

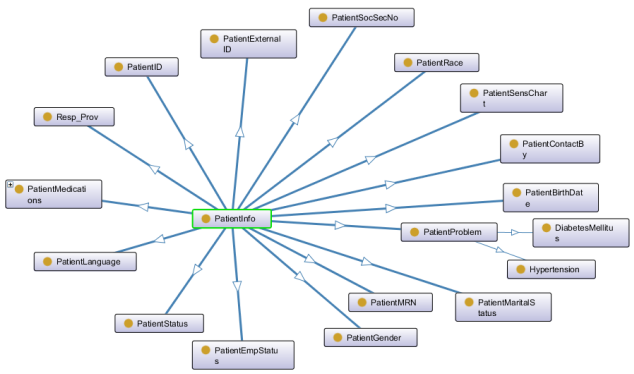


Fig. 8. Patient Ontology Skeleton of the analysed JPEG form

By observing the common features set tabulated in the form of metadata key-value pairs represented in tables I and II, the following object assertions can be made:
(Doctorname, equivalentTo, Resp_Prov)
(PatientProblem, sameAs, PatientRiskStatus)
(DIABETESMELLITUS, sameAs, BLOODSUGAR)
(DIABETESMELLITUS, mayCause, BLOODPRESSURE)
(DIABETESMELLITUS, mayCause, HYPERTENSION)
(DIABETESMELLITUS, mayCause, TRIGYCERIDE)
(TRIGYCERIDE, mayLeadTo, DIABETESMELLITUS)
(HYPERTENSION, mayLeadTo, DIABETESMELLITUS)

Assertions conclude that diabetes may cause changes in trigyceride and hypertension, hence patient should undergo consequent observational tests and vice versa. Cosine similarity measure based on euclidean norms can be applied on degree of occurance of the selected features { PatientID, DoctorName, BloodPressure, BloodSugar, RiskStatus} from both the sets following the assertions. Two document vectors, also known as term frequency vectors (tfv) include D1 = {1, 1, 1, 1, 1, 0} from table I and D2 ={1, 1, 1, 1, 1, 2} from table II consisting of degree of occurance of most important features chosen based on experts' opinion. The similarity value of 0.77 points is computed by applying (1) on D1 and D2. Upon setting the threshold of considertion based on cosine results, the pairs passing the threshold can be combined. Further, by applying any of the rated classification algorithms, the evaluation metrics reading can be tabulated for observation and drawing conclusions on the quality of the integrated ontologies.

VI. CONCLUSION

Semantic Web of data requires effective Machine Learning algorithms to retrieve the available data efficiently and accurately marking the knowledge discovery with reusability to maximum extent. Though lots of technological advances are foreseen, uniformity of data representation remains as drawback due to heterogeneity of data formats observed particularly at schema level. Healthcare industry requires accurate information retrieval without any compromise in the knowledge extraction quality due to criticality of application of knowledge being retrieved in real life applications. As a solution to the stated issue, a procedural approach has been proposed which involves the idea of customized ontology construction according to the metadata tags generated upon analyzing the documents keenly. The so created ontologies can be integrated to achieve better accuracies of knowledge extraction and sharing based on cosine similarity measures computed using Euclidean distance concept. The presented technique can be evaluated with any of the ML classification algorithms for conclusion of prediction correctness. Future work progresses towards extending the proposal on all other types of documents rather than only PDF and images formats.

REFERENCES

1. Musen, M.A, "The Protégé project: A look back and a look forward", AI Matters, Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.
2. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, S¨oren Auer, Christian Bizer, "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, 2015,Vol. 6(2), PP. 167-195.
3. Suchanek F M, Kasneci G, Weikum G, "Yago: A Core of Semantic Knowledge", 16th International Conference on World Wide Web, ACM, 2007, PP. 697-706.
4. Earnt P, Siu A, Weikum G, "KnowLife: A Versatile Approach for constructing a large Knowledge Graph for Biomedical Sciences", BMC Bioinformatics journal, 2015, Vol. 16(1).
5. Jupp S, Burdett T, Leroy C, Parkinson H E, "A New Ontology Lookup Service at EMBL-EBI", SWAT4LS International Conference, 2015, PP. 118-119.
6. Vandenbussche P Y, Atemezing G A, Poveda Villaln M, Vatant B, "Linked Open Vocabularies (LOV): A Gateway to Reuse Semantic Vocabularies on the Web", Semantic Web Journal, 2016, Vol. 8(3), PP. 437-452.
7. Oliveira D, Pesquita C, "Improving the Interoperability of Biomedical Ontologies with Compound Alignments", Journal of Biomedical Semantics, 2018, Vol. 9(1).
8. Paulheim H, "Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods", Semantic Web journal, 2017, Vol. 8(3), PP. 489-508.
9. Pierson J. M, Seitz L, Duque H, Montagnat J, "Metadata for efficient, secure and extensible access to data in a medical grid", Proceedings of 15th International Workshop on Database and Expert Systems Applications, 2004, DOI:10.1109/dexa.2004.1333534.
10. Anitha R, Mukherjee S, "MaaS: Fast Retrieval of Data in Cloud Using Metadata as a Service", Arabian Journal for Science and Engineering, Vol. 40(8), 2015, PP. 2323–2343, DOI: 10.1007/s13369-015-1652-7.

AUTHORS PROFILE

Monika P. has received M. Tech. (CSE), Degree from Visvesvaraya Technological University (VTU), Belagavi, Karnataka in 2011. Currently pursuing research at Department of Computer Science & Engineering, RNS Institute of Technology, Bengaluru, Karnataka – 560 098, affiliated to VTU and working as Assistant Professor in Department of Computer Science & Engineering at Dayananda Sagar College of Engineering, Bengaluru, Karnataka – 560 078. She has published research papers in reputed International Journals and conferences. She has 11 years of teaching experience and 1.6 years of Industry experience. Her areas of research interests include Web Mining, Semantic Web, Artificial Intelligence and Machine Learning.

Dr. G. T. Raju has received M.E. (CSE), Degree from Bangalore University in 1995 and Ph. D (CSE) from Visvesvaraya Technological University (VTU), Belagavi, Karnataka in 2008. Currently working as Vice-Principal, Professor & Head in the Department of Computer Science & Engineering , RNS Institute of Technology, Bengaluru, Karnataka – 560 098. He has 25 years of teaching and research experience. His areas of research interests include Web Mining, Semantic Web, Artificial Intelligence, Machine Learning, Knowledge Data Discovery, Internet of Things, Image Processing and Pattern Recognition. He has published 100+ research papers in reputed International Journals and conferences. He has authored 5 technical text books. He has completed two funded projects. 10+ Research Scholars have been awarded Ph. D degree under his supervision.