IJCTA, 9(24), 2016, pp. 357-362
ISSN: 0974-5572

Data Extraction and Analysis of Electronic Health Records for prediction of Flu

M. Nirupama Bhat*, N. Veeranjaneyulu and B. Jyostna Devi*****

ABSTRACT

Many applications in the area of HealthCare are generating the patients data in the form of Electronic Health Records(EHR). The data of the patients coming from these electronic gadgets are increasing day by day leading to big data. Generally this data is unstructured or semi-structured. It is converted into a standard form and extraction of knowledge can be done for better understanding and decision making. The data extracted can be stored in Resource Description Framework (RDF) store. In this paper, an algorithm is designed to generate the probability tables using Bayesian approach. This algorithm is applied on the dataset of the patients to identify the favorable chances of getting the flu.

Keywords: Electronic Health Records, Big Data, Decision Making, Resource Description Framework.

1. INTRODUCTION

Electronic Health Records play an important role in the field of health care. The size and structure of the current records are sufficient to extract the knowledge for future use by the doctors. Even many systems are proposed to store clinical records. They are also providing information. Sometimes this information may not be sufficient. The RDF (Resource Description Framework)[1] Store proposed by W3C standards helpful for developing semantic based approach of data store which leads to Linked Data. So clinical records stored in the Linked Data [2] form will help to generate more extraction of knowledge.

Medical records are the key source of information for future predictions and better treatments. The records should be stored a structured way to extract knowledge from it in future. So these data should be stored as EHR (Electronic Health Records). More and more types of disease are identified by means of these records. It is important to the doctor to identify the disease with the previous data.

Even health care related gadgets are also become a source of getting more data and should be properly recorded with the particular patient. The outcome of this data gives big data problems. It is also important to share this data with other datasets to identify the depth of the problem. It can be materialized only when the data should be shared among all the hospitals as Linked Data [2, 5].

Clinical records are stored in traditional RDBMS is not good enough to share and extract the knowledge from it. The transformation of this data into Linked Data is required to find the similarities of the data with other world datasets [3]. This help the doctor to predict health and diseases. The semantic extraction [4] for big data integration has proposed better approaches to extract the knowledge.

Electronic Health Records as Linked Data results better extraction of knowledge. This also helps in analysis and identification of the possible diseases, leading to the prescription of drugs. Improved algorithms

* Associate Professor, Department of IT, Vignan's University, Vadlamudi, India, *Email: nirupamakonda@gmail.com*

** Professor, Department of IT, VFSTR University, Vadlamudi, India, *Email: veeru2006n@gmail.com*

*** Assistant Professor, Department of CSE, VFSTR University, Vadlamudi, India, *Email: Jyostna.bodapati82@gmail.com*

could be developed for better extraction of medical data. So the rest of this paper is as follows: Literature Review is given in Section II. Proposed architecture system is described in Section III. Creation of probability tables for a disease in explained in Section IV and the Conclusion and Future Work in Section V.

2. LITERATURE REVIEW

Extraction of data from different hospitals is not same in structure and they even with homonym is transformed into meaningful format. In addition, security is also required for this data as it connected with heterogonous data[6]. The semantic extraction of data is required to this EHR to integrate as single store for this many frameworks are proposed [2]. The semantic ETL(Extraction-Transformation-Loading) of data is stored in data warehouse. The transformation from original schema into RDF schema is proved in different fields[7,8].A generic ontology is required to developed and clean the historical data from different heterogeneous sources and integration is required to get the new knowledge

The sharing of clinical records via social media is also utilized as a source to get the semantic knowledge extraction with the hybrid data repository [9]. Event based knowledge is essential for better knowledge extraction [10]. This helps to reduce the geographical gap between the people with similar medical records. A Systematized Nomenclature of Medical Clinical Terms (SNOMED CT) is a medical ontology required to extract relevant concepts and relations from given datasets[11]. With the rapid increase of clinical data it must be put in structured format and restructured for good quality of information among all the hormones datasets. There are some other ontology like Unified Medical Language System (UMLS), International Classification of Diseases (ICD) among all the SNOMAT CT is widely accepted by the world wide as it is a concept oriented and machine-readable medical terminology in EHR. The refined algorithms are also proposed for decision making with Bayesian belief approach[11] could be able to give better results in knowledge extraction. The SNOMAT CT is developed by the college of American Pathologists and United Kingdom’s Health Service for better maintains of records.

Combining of EHR datasets as single repository is required to gain knowledge from it. The creation of data jackets(summary of datasets) [12] are proposed by Oshawa[12] to support human creativity for problem solving by utilizing data Innovators Marketplace on Data Jackets (IMDJ) is also proposed[13]. Therefore it is required in the fields of medical records to use the concept of data jackets as solutions and relationship of data in EHR. It is included to match the similar records for analysis purpose which will help the doctor to predict the future disease and try to give better treatment to the patient.

The improved patient record with an ontological approach [14] combine different anthologies and is focused through a single model. It explains the ontological approach in EHR and also supports the medical data interoperability that may prevent repeated tests, potential redundancy [6]. So linked data is better way to store the data and also sharing purpose, gain knowledge for the future benefit. The algorithms also can be developed with a better way and understandable both to humans and computer systems. The EHRs are important to convert as Linked Data representation [15] and an ontology approach is required and it is the secondary usage of EHR data.

The cTAKES[16] is a system which works with specific annotation schema and generates XML-based annotated data. It also connects annotated data stored in RDF and queried with SPARQL. Many different systems are proposed and about EHR and emphasis is to store the data as Linked Data. Still these methods are helpful to store structured data and are not sufficient to share and generate data in form of probabilistic approach.

Various steps in EHR are 1.Extact the clinical data from traditional RDBMS storage 2. Clean the data and store it in RDF Store 3. Collecting datasets from different hospitals and store it in cloud environment 4. Extraction of knowledge by Bayesian probability tables approach.

3. PROPOSED ARCHITECTURE

3.1. Ontology

Ontology is a knowledge representation language used to describe the accurate, related, specific and concept oriented representation of data definition that is understandable for humans and machines[17]. To enable domain specific knowledge and analysis, ontology is used. It is sharing of understandable knowledge between the software agents and provide sound information concept to the users. The language engineering is done based on the domain specification for better understandable about the subject.

3.2. Proposed Architecture of EHR

The architecture shows the complete process from the point of data extraction, transformation and loading. The extraction is done from different sources as shown in the Figure 1 and after that the data is transformed into linked data. Then loading of data is take place in the RDF storage is made. This is known as Knowledge Space.

The cloud storage is a knowledge space where the data is stored in structured format from all different hospitals. The application program need to installed in the client system who are going to access the knowledge space, this paradigm will help to give more secure to the public data as medical records are important and it should not be shared with unauthorized purpose.

3.3. Architecture Flow

The data storage of EHR is stored in heterogeneous systems as shown in the figure 1 and the most of the data is not structured, because the data is coming from the different hospitals. The data transformation stage includes the cleaning of the data with the functions like normalization, elimination of duplicate values, and integrity violation is done. In the next stage the data is now transformed into Linked Data.

Now, data loading process takes place. The RDF Storage is loaded into the cloud with Dataset #ID (Identification) as a Data Jackets. The system will keep on loading the data from different hospitals, store in RDF Storage and assigned a unique Dataset identification for each record.

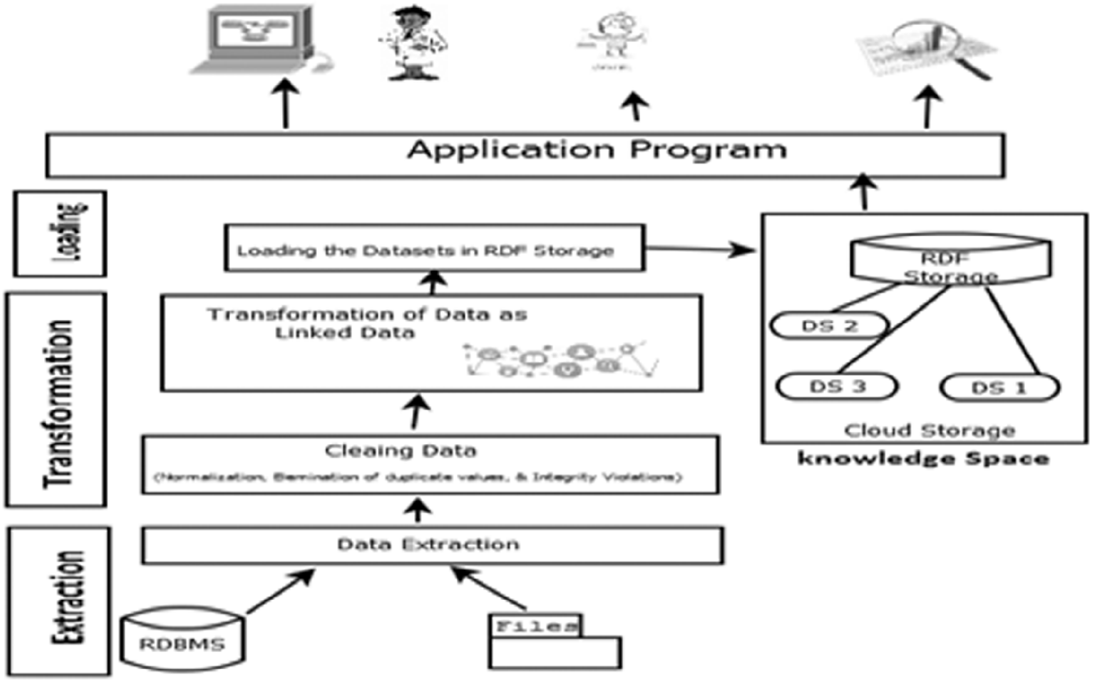
Application Program
Loading
Loading the Datasets in RDF Storage
Transformation
Transformation of Data as Linked Data
Cleaing Data
(Normalization, Elimination of duplicate values, & Integrity Violations)
Extraction
Data Extraction
RDBMS
Files
RDF Storage
DS 2
DS 3
DS 1
Cloud Storage
knowledge Space

Figure 1: Architecture of Electronic Health Record System

3.4. Application Program

The application program is user interface agent developed in Java used to extract the information, because the use of knowledge is different with respect to users as shown in the figure 1. The doctor needs a knowledge that the favorable chances of getting disease and unfavorable chances. In the other hand a chemist need to track the information about the combination of the drug and results. In the same way public or private agency require the data for analysis purpose to figure the particular disease.

To examine the data different SPARQL are executed in RDF storage. This application will give sound knowledge to the different users in the health care sector. SPARQL is RDF query language used to extract the data from RDF storage. It is like relational SQL (Structured Query Language) language that is linked with subject, predicate and object paradigm. SPARQL uses the OPTIONAL instead of LEFT OUTER JOIN for better results.

4. Creation of Probability Tables for a Disease

As part of application program we use the Bayesian probability method to extract the knowledge. A sample of 6 patients' data, who are admitted with fever is taken and these data is stored in RDF Storage . This data is compared with all datasets available in Knowledge Space by using Bayesian belief approach. The application program will give the final results to the doctor giving the favorable changes of patients getting flu infection.

Based on this table values the doctor can predict the future. He will be able to give the required treatment to save from the flu. The algorithms to generate probability tables are shown below.

4.1. Algorithm: To generate probability tables

```
Step 1: Load the data in the knowledge space
Step 2: Assign Id to the data
Step 3: Submitting the required data
Step 4: Comparison of data
  i. Collect the similar data from all the datasets
 ii. If found
         Generate table with Bayesian approach values
iii. Not found
       Results as original no change
Step 5: Reporting the data in required format.
Step 6: End
```

The following Table 1 values are shown to the doctor console of a sample of 6 patient's data of getting flu with favorable chances. Based on this, the doctor can sense the rate of percentage of getting flu and able to give better treatment.

Another example of this application program with respect to the chemist is also done and the result graph is shown in Figure 2. The results given to the chemist are related from five Data Sets with different hospitals.

Table 1
FLU Favorable Result

Patient ID	*Favorable Chances*
PID_101	0.6
PID_102	0.8
PID_103	-
PID_104	0.2
PID_105	-
PID_106	0.3

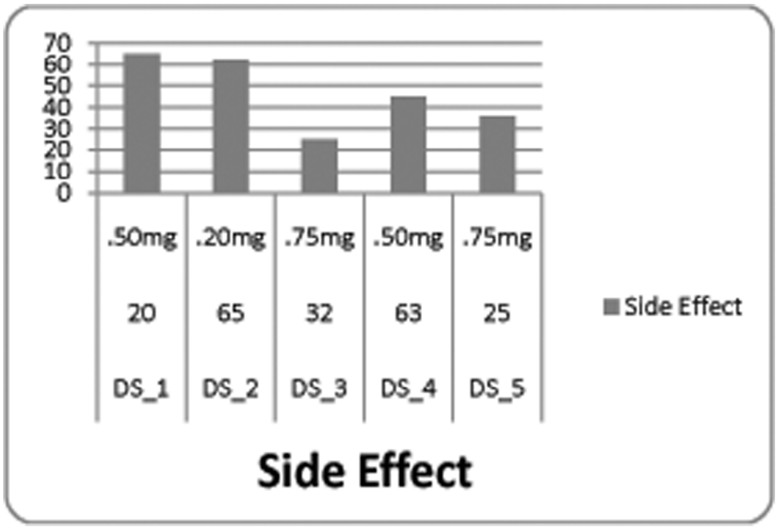

Figure 2: Side Effect of Drug

Apart from this the application is useful for some other researches related purposes to generate reports and also help to gain more knowledge about subject related. The multidimensional view about subject will give more information from all the accepts.

4.2. Technologies used

RDOTE is used to extract the historical data in to RDF data format[18]. RDOTE is tool to extract relational data in to semantic data with better user interface. RDF is a better way to store the data and traditional data need to convert to RDF format. To store RDF storage Jena is used and eclipse environment is used to develop the application and SPARQL is used as a query language.

5. CONCLUSION AND FUTURE SCOPE

Extraction of knowledge is very important in fields of health care using electronic health records every time. The traditional relational data store is not useful and it also leads to get big data issues. Sharing of knowledge is also not much effective. Thus our proposed system will able to extract the historical data into semantic web data and the application gives better user interface to extract and gain sound knowledge about the related fields of medical records. Sharing of these knowledge is also helpful to the all the peoples in the medical fields.

The proposed system need to develop with different modules with respect to different fields of medical industry. Improved programs need to develop for effective work of the application. The cloud store of data is needed to be protected with some security from unauthorized use of clinical records.

REFERENCES

[1] http://www.w3.org/RDF/

[2] *Linked Data* - The Story So Far Christian Bizer, Freie Universität Berlin, Germany Tom Heath, Talis Information Ltd, United Kingdom Tim Berners-Lee, Massachusetts Institute of Technology, USA

[3] Hsiao-Hsien Rau Chien-Yeh Hsu,Yen-Liang Lee, Wei Chen,Wen-Shan Jian, "*Developing Electronic Health Records in Taiwan*", IT Professional, Vol. 12, No. 2, 2010, pp. 17-25.

[4] Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration by Srividya K Bansal Dept. of Engineering & Computing Systems Arizona State University in 2014 IEEE International Congress on Big Data.

[5] C. Bizer, T. Heath, and T. Berners-Lee, "Linked datathe story so far," *International journal on semantic web and information systems*, vol. 5, no. 3, pp. 1–22, 2009.

[6] Personal Healthcare Record Integration Method based on Linked Data Model by Boyi Xu,Yan You *College of Economics &Management, Shanghai Jiao Tong University in* 2014 IEEE 11th International Conference on e-Business Engineering.

[7] "IBM InfoSphere Platform–big data, information integration, data warehousing, master data management, lifecycle management and data security." [Online]. Available: http://www-01.ibm.com/software/data/infosphere/. [Accessed: 28-Feb-2014].

[8] "Warehouse Builder 11gR2: Home Page on OTN." [Online]. Available: http://www.oracle.com/technetwork/developertools/arehouse/overview/introduction/index.html. [Accessed: 28-Feb-2014].

[9] Healthcare-Event Driven Semantic Knowledge Extraction with Hybrid Data Repository by Hong Qing Yu, Xia Zhao, Xin Zhen, Feng Dong, Enjie Liu, Gordon Clapworthy Department of Computer Science and Technology University of Bedfordshire Luton UK in 978-1-4799-4233-6/14/$31.00 ©2014 IEEE.

[10] Chen, H.L., A socio-technology perspective of museum practitioners' image-using behaviors, *The Electronic Library*, vol. 25 no.1 (2007), 18-35.

[11] Ontology Based EMR for Decision Making in Health Care Using SNOMED CT by J. Kulandai Josephine Julina , D. Thenmozhi Department of Computer Science and Engineering SSN College of Engineering Chennai, India in ISBN: 978-1-4673-1601-9/12/$31.00 ©2012 IEEE ICRTIT-2012.

[12] Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data by Teruaki Hayashi , Yukio Ohsawa in 2015 2nd International Conference on Signal Processing and Integrated Network (SPIN) @ 2015 IEEE.

[13] Data Jackets Site, [online] available from https://sites.google.com/datajacekts/[Accessed 5th January]

[14] Improving Patient Care in Transport Medicine through an Ontological Approch by Phillip DEPalo, Kyungeun Park, Yeong-Tae Song in IMCOM(ICUMC) 14, January 9-11,2014 Siem Reap, Cambodia.ACM 978-1-4503-2644-5.

[15] Towards Semantic Web based Knowledge Representation and Extraction from Electronic Health Records by Cui Tao, Jyotishman Pathak and Susan Rea Welch in October 28,2011 Glasgow, Scotland, UK , ACM 978-1-4503-0954-7/11/10.

[16] Strategic Health IT Advanced Research Projects (SHARP) , area 4 on secondary usage of EHRs, http://sharpn.org.

[17] Validation of an Ontology Based Search Engine for the Electronic Medical Record: Application in the Emergency Department Setting. Krishnaraj,Arun, et al, Orlando: American Medical Information Association 2012.