Semantic transformation model for clinical documents in big data to support healthcare analytics

Shujaat Hussain, Sungyoung Lee
Ubiquitous Computing Laboratory
Dept. of Computer Engineering,
Kyung Hee University
Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do,
446-701, Korea

***Abstract*—The standardized healthcare documents are being adopted at an exponential rate all around the world which poses several challenges about its large scale analysis and comprehension. The healthcare standards are complex and difficult to understand for a health analytics expert due to its comprehensive nature. This paper proposes a semantic transformation model of the healthcare documents in a distributed environment to tackle the voluminous data and its variety. In this paper Hadoop is used for the semantic transformation model and clinical document architecture (CDA) standard for the case study. The case study shows that the health analytics can be well supported by the transformation model as it is simple and tailor made for the situation.**

I. Introduction

All the stakeholders in the healthcare industry i.e. the physicians, specialists and health insurance companies generate healthcare data in different forms and standards. These standards usually have a very complex structure which makes the healthcare data very challenging. New standards record comprehensive data for every patient. From patient's admission to discharge, many documents are generated comprising of tests, reports, previous history and doctor recommendations etc. During this process a lot of data is generated as patients today are monitored constantly through different devices [1]. With a lot of data becoming a norm in healthcare, extracting analytics from the data is a big challenge due to performance and scalability issues.

The introduction of 'Big Data' in healthcare has the potential to change the way the data is processed and also the operational structure of IT in healthcare. The complex nature of data sets also result in inefficient and difficult management of conventional databases. Big data has three important traits i.e. volume, variety and velocity [2]. It does not mean that big data can play an important role only where size matters but it also helps in the complex nature of the data sets.

Big data in healthcare refers to the health data sets which are complex, large and cannot be handled with conventional infrastructure and databases. It not only refers to the volume of the data but also the variety and diversity of the data. Health data includes electronic documents, patient scans and images, prescriptions, discharge summaries, insurance claims, patient sensory data and social data from facebook and twitter.

It is a huge opportunity to transform and associate the healthcare data that could identify patterns and assist in health analytics [3]. This could lead to cost effective solutions and influence patient life due to better understanding of the data and its characteristics. The potential benefits can also be long term as many complex situations can be better understood like patient readmission, surgeries, risk analysis etc. 1.2 billion clinical documents are produced in the U.S. each year, and about 60 percent of these contain valuable patient care information in an unstructured/semi-structured format [4].The health analytics extracted from the transformed clinical documents can lead to better decisions and much quicker processing for future use. An estimate of 80% hospitals are using EHRs now [5].

Hadoop [6] is used for the big data implementation i.e. model transformation and metadata parsing. More details will come in late sections. The rest of the paper is organized as follows: Section II presents the related work. Section III, details health standard clinical document architecture, section IV describes the a distributed semantic transformation model and section V explains the overall framework and infrastructure. Case study is explained in section VI. Finally, we conclude our work in section V.

II. Related Work

The authors in [7] use Hadoop [6] and discharge summaries. They use simple log files for discharge summaries and group them. The data set is synthetically generated and conformed to a structure with respect to drugs prescribed, patient, date and quantity. They transform the data by grouping patients against all the drugs they have been prescribed. The data does not follow any data standard and the approach is very simple.

Another Approach is cloud based as they use the Hadoop infrastructure for data aggregation management [8]. They use Hive and HBase for querying and storing relations. They did not do any analytics on the medical data. Mapreduce jobs are deployed just to retrieve stored data.

The Hospital data is used to build a hospital-specific Predixion model [9]. The prediction model is then used to risk stratify patients upon admission. Risk scores are updated throughout the patient's stay. Readmission risk scores are used by care givers to target appropriate patient care paths.

978-1-4673-9152-8/15/$31.00 ©2015 IEEE

TABLE I: CDA Discharge Summary Sections and Details

Section	Contains
Allergies, Adverse, Alerts	Substance. Reaction, Status
Hospital Discharge Medications	Medication and Instructions
Plan of Care	Planned Activity and Planned Date
Family history	Parents Diagnosis and Age At Onset
Functional Status	Functional Condition, Effective Dates, Condition
Immunizations	Vaccine, Date, Status
Procedures	Procedure and Date
Problems	Condition, Effective Dates and Condition Status
Vital Signs	date/Time, Height, Weight, BP etc
Social History	Social History Element, Description and Dates

One technique is a Clinical Data Managing and Analyzing System [10], which uses hybrid XML database and HBase/Hadoop infrastructure to handle big amount of heart disease clinical data analysis online. They have an RDBMS and HBase system to store XML documents and using the system for fault tolerance and replication of the data. They are using a hybrid solution of big data and RDBMS which is known to be inefficient due to different data access techniques.

III. Clinical document Architecture

Clinical document Architecture (CDA) [11] is a base standard which provides a common architecture, coding, semantic framework, and markup language for the creation of electronic clinical documents. CDA documents are coded in eXtensible Markup Language (XML). A CDA document is an XML document which comprises of a header which identify the patient, encounter information, contact information, care provider etc. and there is a body part which consists of some mandatory and optional sections. These sections contain diagnosis, plan of care, family history, allergies, vital signs etc. A more detailed view as shown in table 1.

IV. Semantic Distributed Transformation Model

Semantic transformation of complex healthcare documents gives a huge advantage in comprehending and understanding the clinical data. It also assists in the health analytics as the transformed reduced documents can be tailor made for the health analytics scenario. This leads to high speed solutions and less processing for the analytics output. We have taken CDA documents as our case study. There are many sections in the CDA body which comprises of clinical statements such as observations, substance administration etc and their respective codes. These clinical statements have different complex data types like coded value (CV), Concept descriptor (CD) and code with equivalents (CE). Figure 1 shows a sample section specifically the diagnosis which is the required part of a discharge summary gives a glimpse that the codes, naming convention and relationships can be very confusing and difficult to parse or query in a meaningful time. The important information to extract is the diagnosis i.e. the chest pain and the code associated with it.

```
<!--
**********************************************************
Hospital Discharge Diagnosis - Required -
**********************************************************
-->
<component>
    <section>
        <!-- Discharge Summary Hospital Discharge Diagnosis Template Id -->
        <templateId root="2.16.840.1.113883.10.20.16.2.1"/>
        <!-- IHE Hospital Discharge Diagnosis Template Id -->
        <templateId root="1.3.6.1.4.1.19376.1.5.3.1.3.7"/>
        <id extension="9937012" root="2.16.840.1.113883.19"/>
        <code code="11535-2" displayName="HOSPITAL DISCHARGE DX"
            codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
        <title>Hospital Discharge Diagnosis</title>
        <text> Unspecified chest pain </text>
        <entry>
            <act classCode="ACT" moodCode="EVN">
                <!-- Required Problem Concern Entry element -->
                <templateId root="1.3.6.1.4.1.19376.1.5.3.1.4.5.2"/>
                <templateId root="1.3.6.1.4.1.19376.1.5.3.1.4.5.1"/>
                <templateId root="2.16.840.1.113883.10.20.1.27"/>
                <id root="5a784260-6856-4f38-9638-80c751aff2fb"/>
                <code nullFlavor="NA"/>
                <statusCode code="active"/>
                <effectiveTime>
                    <low value="20903003"/>
                </effectiveTime>
                <entryRelationship typeCode="SUBJ" inversionInd="false">
                    <observation classCode="OBS" moodCode="EVN" negationInd="false">
                        <templateId root="2.16.840.1.113883.10.20.1.28"/>
                        <templateId root="1.3.6.1.4.1.19376.1.5.3.1.4.5"/>
                        <code code="409586006" codeSystem="2.16.840.1.113883.6.96"
                            codeSystemName="SNOMED CT" displayName="Complaint"/>
                        <statusCode code="completed"/>
                        <effectiveTime>
                            <low nullFlavor="UNK"/>
                        </effectiveTime>
                        <value xsi:type="CD" code="29857009"
                            codeSystem="2.16.840.1.113883.6.96"
                            codeSystemName="SNOMED CT" displayName="Chest Pain"/>
                    </observation>
                </entryRelationship>
            </act>
        </entry>

    </section>
</component>
```

Fig. 1: CDA Component,section,code,entity relation,observation

The semantic transformation is needed so that the CDA document can be used for health analytics with simple processing. For this purpose we need a model which can maintain the basic relationships but remove the complexity of sections, complex data types and relationships. The data transformation is done with respect to the analytics expert's needs. A generalized object model is needed to create a simple and understandable reduced document. The class diagram in figure 2 shows the reduced document which contains the patient id and encounter information. It also contain the list of reduced sections that the health analytic query may need. The reduced section contains the section code, title and collection of clinical statements. The clinical statement is an atomic object which contains the code and value (description). The remaining classes i.e. driver, XMLInputFormat, mapper and reducer are supporting the distributed environment of the system and will be explained in framework section.

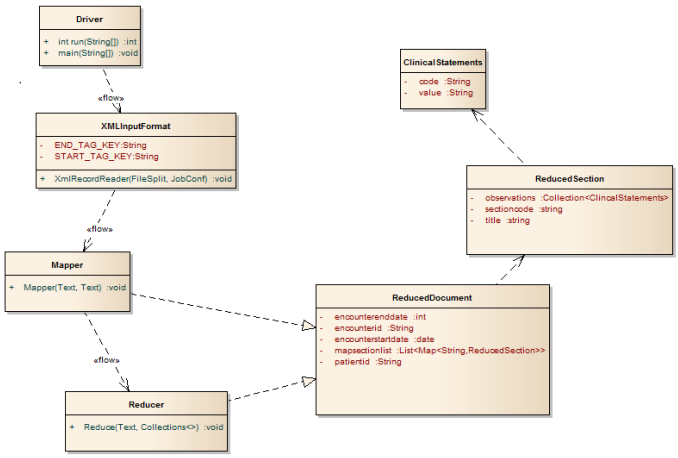

Fig. 2: Class Diagram

V. Proposed Framework

The input of the proposed framework will be the health and clinical standard documents from hospitals, insurance companies, physicians and pharmacies. This framework will have a big data storage and processing layer. This layer caters to the need of volume of the documents and the complexity of the documents.

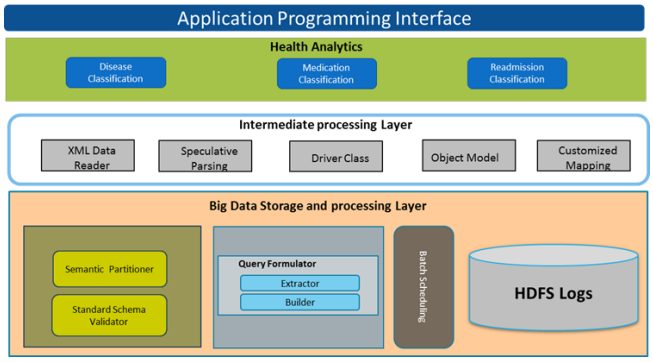

Fig. 3: Big Data Classification Framework

A. Big Data Storage and processing Layer

The big data layer consists of a repository and a job scheduling algorithm inherited from the hadoop implementation. The component explanation is given below

1) Semantic Partitioner : The storage layer is using Hadoop Distributed File System (HDFS) for storing the data. This layer has a semantic partitioner to extract the complete healthcare document. The default behavior of Hadoop is to read all the data line by line which in this case is not practical. If the whole document is not retrieved the semantics of that document is lost as the completeness is a must for extraction.

2) Standard Schema Validator: This module is specific to the healthcare documents that are going to be processed and parsed. For example if the CDA documents are being processed then the CDA schema is the standard schema. This schema is saved in the distributed cache so that all the data nodes in the big data environment have this schema.

3) Query Formulator: The query formulator has two parts i.e. the extractor and the builder. This query is based on the health analytics scenario and it maps the concepts of this query to HDFS repository.

4) Batch Scheduling: Batch scheduling is related to the multiple jobs in the distributed environment. The jobs after processing assigns the intermediate data for processing to the data nodes in queue. This needs to be customized so that similar intermediate data can be broadcasted to one data node. This increases the overall job execution time and performance.

B. Intermediate Processing Layer

The intermediate processing layer comprises of the functions which are passed on the big data storage layer for assistance in processing. This is a supporting functions layer which will pass on different information like partition the data and parse the data.

1) XML Data Reader: This is the custom data reader which assists in retrieving the data from the HDFS. This indicates the starting and ending of the healthcare documents. As most of the healthcare documents are in XML, so a custom reader was designed to increase the efficiency for the semantic partitioner. Different health standards have different header information in their documents, so an additional check in the data reader can assist to identify healthcare information

2) Speculative Parsing: A speculative parser for healthcare documents speculates the section parsing with the help of the schema. This will assist the schema validator in the big data layer. It will increase the efficiency as it will predict the medical document section due to its prior knowledge of document schema. It will automatically extract data from the medical data which are compulsory in the data and in worst case will traverse the whole document.

3) Driver Class: The driver class calls all the functions relevant to the distributed execution. In this instance it calls the mapper and reducer classes. The mapper parses the document and broadcast the extracted intermediate information. The driver then calls the reducer class and converts this information into the object model and customized mapping depending on the standard.

C. Health Analytics

The health analytics layer consists of the different scenarios. It is layer which is customizable and depends on the need of the health expert. In the framework the mentioned classifications are disease, medication and readmission. As these classifications are related to the case study, so it is mentioned.

VI. Case Study

The case study of readmitted patients and how likely is the risk of readmission again for these patients. Around 20% of all hospital admissions occur within 30 days of a previous discharge 30% of the 2 trillion annual cost of healthcare in the United States [12]. A discharge summary reports on a patient's

episode of care. Identifying patients at risk of readmission can guide efficient resource utilization and can potentially save millions of healthcare dollars each year. Effectively making predictions from such complex standard data and volume in coming years requires the development of effective analytical models.

We have implemented naive bayes in mapreduce which helps predicting by training huge volume of data and we have labeled data from the parallel parser.

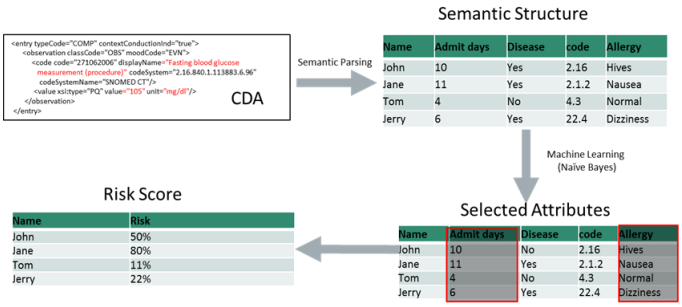

Fig. 4: Readmission Case Study

Figure 4 starts from a Clinical Document Architecture (CDA) [5] and information is extracted into a table which breaks it down to admission days, disease, codes and allergies. This is done by the transformation module which is explained in section 4. This object model is then converted into a table for applying the algorithm After applying naive bayes on selected attributes i.e. the admit days and allergies. By this process the risk attached to readmission of respective patients are found. The last table shows the risk score of the users in terms of their likelihood of readmission. The machine learning technique is not explained as it is out of scope of this paper. This case study proves that converting unstructured/standard clinical documents can support the health care analytics.

VII. Conclusion

Health standards nowadays are very comprehensive and complete as it has all the infromation about patient demographics and medical history. This in most cases leads to a complex medical document which is hard to parse and process. This creates Our approach takes advantage of parallel processing in big data and make machine learning and parsing more efficient. Readmissions problem is selected as our application as it is ideal due to such complex standard data and volume

Acknowledgment

This work was supported by the Industrial Core Technology Development Program (10049079 , Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)" and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) NRF-2014R1A2A2A01003914.

References

[1] A. Bell, P. Rogers, C. Farnell, B. Sparkman, and S. C. Smith, "Wireless patient monitoring system," in *Healthcare Innovation Conference (HIC), 2014 IEEE*. IEEE, 2014, pp. 149–152.

[2] P. Zikopoulos, C. Eaton *et al.*, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[3] R. H. Ellaway, M. V. Pusic, R. M. Galbraith, and T. Cameron, "Developing the role of big data and analytics in health professional education," *Medical teacher*, vol. 36, no. 3, pp. 216–222, 2014.

[4] datamark.net/uploads/files/unstructured_ehr_data_white_paper.pdfq/, [Online; accessed 15-June-2015].

[5] http://hadapt.com/blog/2014/06/17/how-big-data-analytics-is-changing-healthcare.com/ , [Online; accessed 15-June-2015].

[6] http://hadoop.apache.org/, [Online; accessed 1-June-2015].

[7] H. Horiguchi, H. Yasunaga, H. Hashimoto, and K. Ohe, "A user-friendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 151, 2012.

[8] A. Bahga and V. K. Madisetti, "A cloud-based approach for interoperable electronic health records (ehrs)," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 5, pp. 894–906, 2013.

[9] http://predixionsoftware.com/, [Online; accessed 15-June-2015].

[10] Y. Wang, L. Wang, H. Liu, and C. Lei, "Large-scale clinical data management and analysis system based on cloud computing," in *Frontier and Future Development of Information Technology in Medicine and Education*. Springer, 2014, pp. 1575–1583.

[11] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "Hl7 clinical document architecture, release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 30–39, 2006.

[12] http://siam.org/meetings/sdm13/sun.pdf.com/, [Online; accessed 15-June-2015].