Semantic-based Approach for Solving the Heterogeneity of Clinical Data

Basma Elsharkawy, Rashed Salem, and Hatem Abdel Kader
Information Systems Department
Faculty of Computers and Information,
Menoufia University, Shebin Elkom, Egypt.

Abstract: **Clinical records contain massive heterogeneity number of data types, generally written in free-note without a linguistic standard. Other forms of medical data include medical images with/without metadata (*e.g*., CT, MRI, radiology, etc.), audios (*e.g*., transcriptions, ultrasound), videos (*e.g*., surgery recording), and structured data (*e.g*., laboratory test results, age, year, weight, billing, etc.). Consequently, to retrieve the knowledge from these data is not trivial task. Handling the heterogeneity besides largeness and complexity of these data is a challenge. The main purpose of this paper is proposing a framework with two-fold. Firstly, it achieves a semantic-based integration approach, which resolves the heterogeneity issue during the integration process of healthcare data from various data sources. Secondly, it achieves a semantic-based medical retrieval approach with enhanced precision. Our experimental study on medical datasets demonstrates the significant accuracy and speedup of the proposed framework over existing approaches.**

Keywords-Schema data integration, Heterogeneity, Image retrieval, Semantic ontology, OWL, RDF, XML.

I. INTRODUCTION

Big data is a collection of datasets in a large variety of domains [15], healthcare is one of such domains. There are different types of data including structured, semi-structured, and unstructured. Statistically, 80% of medical data are unstructured, which further complicates the management of these data [11, 21]. The major source for healthcare applications is patient records. Data integration is the task of combining different data sources, and providing a unified view of the data. Such integrated data are needed to be standardized and kept in a repository, *i.e*., data warehouses, for ease of retrieval and analytics later [16].

However, integrating data from a variety of sources is not a trivial task, due to the large volumes of heterogeneous data during mapping, ranking, and key matching [9]. Moreover, structural and semantic heterogeneity is another problem that faces data integration [10, 13]. In this paper, we address the problem of resolving structural and semantic heterogeneity for healthcare applications. While structural heterogeneity addresses schema conflicts, semantic heterogeneity addresses meaning conflicts, *e.g*., synonyms and homonyms conflicts. Fortunately, semantic Web can be exploited to resolve semantic heterogeneity issue. Using semantic Web, the same concepts which given by several words, *i.e*., synonyms, as well as the different concepts given by the same word, *i.e*., homonyms, can be defined. Semantic technologies (*e.g*., Ontology) provide major solutions to semantic interoperability in healthcare systems. Moreover, ontologies can deliver solutions for image retrieval. They seek to map the low-level image features with high level ontology concepts.

Compared with content-based and keyword-based image retrieval, ontology-based retrieval concentrates on capturing semantic content. Furthermore, ontology plays an important role to represent the knowledge as a set of concepts within a domain and the relationships between pairs of concepts. Ontologies can be used to support a variety of tasks in different domains including knowledge representation, natural language processing, information retrieval, database integration, digital libraries, *etc*.

This paper proposes a framework in which different medical data types are merged into a unified format. The proposed framework tackles heterogeneity issue during data integration process. By the proposed framework physicians can build a patient's history record, and thus helps physicians in decision making. The proposed framework keeps all patient history without losing any data. Furthermore, this paper proposes a semantic-based framework for medical data retrieval.

This paper is organized as follows; section 2 presents background of semantic schema mapping and integration approaches. Section 3 introduces a literature review of clinical data management. The proposed semantic-based framework for resolving heterogeneity and retrieval of healthcare data is discussed in section 4. Section 5 presents how to handle several cases of medical data. Implementation and discussion of the proposed framework are provided in section 6. Section 7 concludes the work and highlights the future work.

II. Semantic Schema Mapping Approaches

Schema mappings are expressions that specify how an instance of a source database can be transformed into an instance of a target database. In recent years, they have received an increasing attention both from the research community and the commercial tools market.

A schema-mapping system is used to support the process of generating and executing mappings in practical scenarios. It typically allows users to provide an abstract specification of the mapping as a set of correspondences among schema elements, specified through a friendly user-interface. Based on such specification, the mapping system will first generate a number of mappings – usually under the form of tuple generating dependencies (tgds) that correlate source tables with target tables; then, based on these mappings, an executable transformation, *i.e.*, a runtime script in SQL or XQuery, can be practically used to run the mappings and generate solutions.

Christian Bizeret al. [29] introduced the mappings process on the Web and a composition method for chaining partial mappings from different sources based on a mapping quality assessment heuristically. By introducing R2R mapping language that designed to fulfill each of the vocabulary cherry picking and interlinking and discovery, whereas every term must be identified with its own dereferenceable URI in order to enable mappings to be interlinked with RDFS or OWL vocabulary term definitions and mappings by RDF links.

A. Semantic Schema Matching Approaches

The semantics of schema concepts acts a critical role in the determining mappings/ matching process between different data sources. Identifying both the implicit and explicit meaning of schema label, the semantic correspondences among the elements of different schemas have been defined. This identification requires the development of a method for lexical annotation, *i.e.*, finding the meanings of a schema label in a reference lexical database. Several methods are connected with this problem by using lexical knowledge in different ways.

In healthcare environments, Lee, C. Y., et al. [28] proposed an attribute matching algorithm to resolve semantic conflicts and interoperability problems, which does the semantic matching over two steps; first step is checking the attribute similarity with domain knowledge. The second step is checking word relatedness through overlapped phrases, hyponyms and hyponyms.

Partyka, J., et al. [24, 25] addressed semantic heterogeneity challenge between different data sources. One of traditional methods is N-gram method that often fails. Fundamentally, it depends on discovering the similarity among shared instances, that results in an overestimation of semantic matching between independent attributes. They proposed an approach initially depends on choosing similarity among value attributes, then examining the instances between them which is known as an entropy-based distribution (EBD). Then, they compared the N-gram method and the new T-Sim method for calculating EBD.

Chena, N., et al. [26] mentioned that the syntactic schema matching method cannot identify possible semantic mapping relationships; for example, in healthcare domain, element ‘diagnosis’ and element ‘prescription’ have identical semantics, until this time they cannot be identified by the syntactic method. They proposed the Node Semantic Similarity (NSS) method based on conjunctive normal forms and a vector space model. They designed a hybrid algorithm based on label meanings and annotations for computing the relationship between concepts of label. Then, the semantic relationship is translated between nodes into a propositional formula which confirms the validity of this formula to confirm the semantic relationships. Firstly, the algorithm calculates the label and node concepts, secondly it computes the conceptual relationship. The Zhao, C. [30] has proposed a multilayer schema matching approach with many layers. The first layer connected with semantic similarity. The second layer verifies the functional dependency to formulize structural information of schemas. A third layer proposes a probabilistic factor. The last layer confirms the mapping element pairs process with reasonable depending on each layer’s results. In general, the semantic similarity measure works on data preprocessing, then it does the lexicographic similarity and generates the filtered matching sets.

Islam, A. and Inkpen, D. [27] addressed the text similarity challenge to solve semantic heterogeneity as a critical challenge in any data sharing integration system, a distributed database system, a web service, or a one-to-one data management system. The Semantic Text Similarity (STS) method has been recommended, that discovers the similarity of two texts in terms of semantic and syntactic information (by common-word order method). They use three similarity functions in order to extract more general text similarity approach. At the beginning, string similarity and semantic word similarity are considered. Then, they introduce common-word order similarity function to combine syntactic information. Finally, the text similarity is derived by merging string similarity, semantic similarity and common-word order similarity with normalization.

In all mentioned data integration research, the semantics of the transformations are strongly linked to the implementation method. The intention is that the integrated database be implemented as a view of the component databases, and that queries against the integrated database be executed by translating them into queries against the component databases and then combining the results. The semantics of the individual transformations are given by their effects on queries. However, the lack of any independent characterization of their semantics makes it difficult to prove properties of the transformations, or to use any alternative implementation of the methodology.

B. Schema Integration Techniques

The schema integration derives from two tasks: database integration, and integration of user views, which occurs during the design phase of a database when constructing a schema that satisfies the individual needs of each of a set of user groups. However, they fail to note that these two kinds of schema integration are fundamentally different. For database integration, instances of each of the source databases are transformed into instances of the merged schema. Moreover, when integrating multiple user views, instances of the merged schema must be transformed back into instances of the user views. A good schema-integration method should therefore take account of its intended purpose and include semantics for the underlying transformations of instances [24].

III. RELATED WORK

There are many approaches proposed in the literature for managing clinical records including, chunking, data-driven, free-text assignment codes, content-based image retrieval, and semantic-based image retrieval. The chunking approach is proposed to identify non-recursive words and base noun phrases in the text, *i.e.*, a key issue in symptom and disease identification as a term. Thus, it extracts structured data from clinical records easily. Chunking handles data annotated by medical domain using a chunk annotation scheme with extra credibility, which involves symbols as noun phrases (NPs), main verb (MVs), and a common annotation for adjectival and adverbial phrases (APs) [19]. However, there is a very limited amount of annotated text of this kind available for health-care systems [22].

In data-driven approaches, a driving data element that is an independent variable should be selected. The independent variable is used to determine the other linked patient information. Syndrome/sub-syndrome classification and 3-digit ICD-9 final diagnosis code are used to determine the driving data element. The data element that is used to realize the patient’s record would have a clear and easily recognized relationship. The mapping from data element is well defined if the patient has been grouped by a single value of this data element. The challenge is not only in data storage and access, but also in scalability of healthcare sources [4].

Medical image retrieval can help physicians in finding information that assist them in decision making. Medical image retrieval systems extract features as color, texture, shape and spatial relationships. Image features are extracted from the full image and then are indexed. The variety of medical image types makes the process of retrieval is a non-trivial task. For instance, radiology images face many difficulties [5]. Particularly, such radiology images contain rich information and specific features that need to be carefully recognized for medical image analysis.

Image retrieval systems are generally classified into two major approaches. The first approach searches local or global image features such as color or texture. The other approaches add key words to images as an annotation. Content-based Image Retrieval (CBIR) approach is considered a rapidly advancing research area. It depends on searching similarity of image features from a database based on the color, shape and texture [5]. Images are presented as a query against image database. The similarity between image features in the database is retrieved with the help of indexing images [7, 20]. The indexing of images provides a rapid path for searching image databases [3]. However, there is still a “Semantic gap” between what users need and what CBIR systems can achieve. In particular, there are no sensible means by which queries can be presented to CBIR systems [14, 18].

The Semantic-based Image Retrieval (SBIR) systems include several components of information extraction such as a textual description and visual feature, and semantic image retrieval. The extraction process of SBIR is based on low level features of images to identify objects. Open issues are the nature of digital images, as well as descriptions of images, *i.e.*, high-level concepts such as rat and dogs. However, the main problem is the semantic gap discrepancy between low-level features and high level concepts [12]. Moreover, different users at a different time may give different interpretations for the same image [1, 17]. TABLE I provides a summarized comparison among medical retrieval approaches and our proposed framework.

TABLE I MEDICAL DATA TYPE RETRIEVAL APPROACHES

	CBIR	**SBIR**	**Chunking**	**Data-driven**	**Proposed Framework**
Data type	Image	Image	Free text	Free text	Image, Free text, audio, and video
Data missing	Lossy	Lossy	Lossy	Lossy	Loss-less
Challenges	1)Semantic Understanding of media is visual 2)Integrating, Searching, Selecting	1)The different Forms of images 2)Lack of relation between objects and the meaning	1)Identification Of medical concepts 2)Clarification of medical concepts relations	1)Data scalability 2)Data access	1)Grew up of Global ontology
Precision	In accurate with medical images	In accurate with medical images	In accurate in medical concepts	Inaccurate due to data missing	Accurate
Performance	Degrading with large database	Degrading with general concepts	Degrading with medical concepts	Degrading with large database	High Performance
Scalability	Low scale	Low scale	Low scale	Low scale	High scale

IV.PROPOSED SEMANTIC-BASED FRAMEWORK FOR INTEGRATING MEDICAL DATA

Clinical data are represented in structured and unstructured form. Surgical producers, treatment and drugs data are examples of structured data. Structured data can be computerized and allow performing analysis of data, queries and aggregation for patient records. They are organized in a mightily mechanized and manageable structure. Structured data are prepared for seamless integration into a database or well-structured file format. Structured data needs to stay comparatively simplistic and uncomplicated [11]. Moreover, structured data depend on a data model. The data model specifies how data will be generated, stored, processed and accessed [6]. Structured data are generated through constrained choices in the form of data entry, which overall drop-down menus, check boxes, and pre-filled templates. This type of data is easily searchable and aggregated, can be analyzed and reported, and is linked to other information resources. The high cost and performance limitations of storage, memory and processing allows relational databases and spreadsheets using structured data are the only way to effectively manage data [6].

Unlike, unstructured clinical data may contain free clinical notes and multimedia contents such as medical images and voice. These data may have an interior structure, nevertheless they are still considered as an unstructured form because the data which they contain don't care appropriate sorted into a database. The concept of "big data" is widely associated with unstructured data. Big data denote to extremely large datasets that are difficult to analyze with traditional tools [2, 6].

There is a variety of challenges for handling clinical notes including ungrammatical, short phrases and abbreviations. The proposed framework helps in solving these challenges of clinical notes, and the heterogeneity of clinical record. In addition to structured medical data, the proposed framework merges medical images, clinical notes and audio data type into a unified framework. Then, physicians can perform a "DL language" or "SPARQL" query to access the different data types. The proposed medical image retrieval framework handles medical images as *url* and gives a label for each medical image. In the case of medical reports, which consist of both medical images and text, the process of text extraction is performed firstly before performing medical image process.

In this paper, the Semantic Web ontology are used to solve data integration challenges in healthcare system. Ontology makes the relationships between patient data clear, the data can be derived from various resources. For example, if two different patient infected with the same diagnosis, whereas they are represented by different names or different identifiers, ontologies are used to map these names to the same diagnosis descriptor. Due to heterogeneity and complexity of data integration in healthcare system, the hybrid ontology approaches will be used [13, 32]. The semantic ontology of each patient is described by local ontology. In order to build the global ontology, all local ontologies are merged to one with shared all vocabulary and physician's concepts.

A. Technical process of integrating medical record

The data type in patient record is not unified, it is consisting of various media types. Therefore, the mapping document can be generated using local ontologies that defines the semantics of the source data then merging them with the global one. The free text is transformed to XML. The structure of an XML node is basic building part of mapping ontology and the relationship between the elements. Each XML node has a specific element in the OWL ontology. For example, the following XML node maps the general part of the ontology.

```
<XMLNode ="globalOWL:patient1">
</XMLNode>
```

The OWL class element determines the class to be mapped into output RDF file. The OWL class mapping element may contain multiple OWL class elements that are arranged in the ontology. Thus, the OWL class element consists of defined element containing fixed data, which specifies the class names to be constructed in the RDF document. The OWL property element contains specific relations between classes.

The RDF document is created from the input XML document, OWL ontology and the medical record. The main class in the OWL ontology is owl:Thing and all classes are subclasses of it. The semantic relationship has been defined between patient's data in the output RDF file. The ontology also defines the meaning of each element. The RDF is based on the idea of identifying things using Web identifiers (URIs), and describes resources by giving them labels. The RDF statement defines patient's label and asserts that some relationships, indicated by the predicate, holds between other patient records. As an example of a resource on the Web, we can have the following statement. The web page whose URI is "http://www.patient1.org/boons.html"is referred by an URI as "http://wiki.hip.fi/xml/ontology/BonesDEPT.owl". Generally, an ontology is a description of concepts and relationships that can be found in medical records. In the context of this paper, the language in which these statements are written is Description Logics "DL" and queries are performed using SPARQL [22].

B. System Architecture

The architecture of the proposal approach according to previous methodology is shown in Fig. 1. It contains three layers: physical layer, semantic service layer, and application layer.

1. The physical layer consists of data sources in healthcare system.
2. The semantic layer involves the ontology base, semantic query and reason service.
3. The application layer is designed to access ontology from remote locations and at different platforms.

V. Multiform Medical Data Handling

The proposed framework focuses on integrating data from heterogeneous medical, data sources. Here in, we introduce how the proposed framework handle unstructured medical data such as clinical notes, medical images, physician's reports and audio data type. In Fig. 2, patient record has been handled either as clinical note, medical image, audio type or patient report. In the proposed framework a local ontology will be merged with the global one. The physicians or administrator allow doing any query on the global ontology.

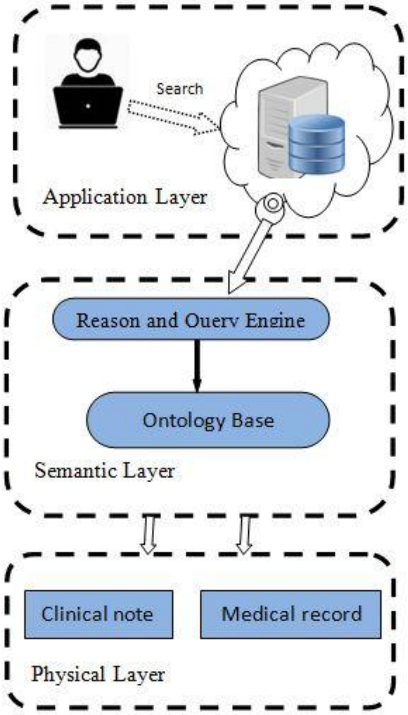


Fig. 1 The architecture of proposal approach

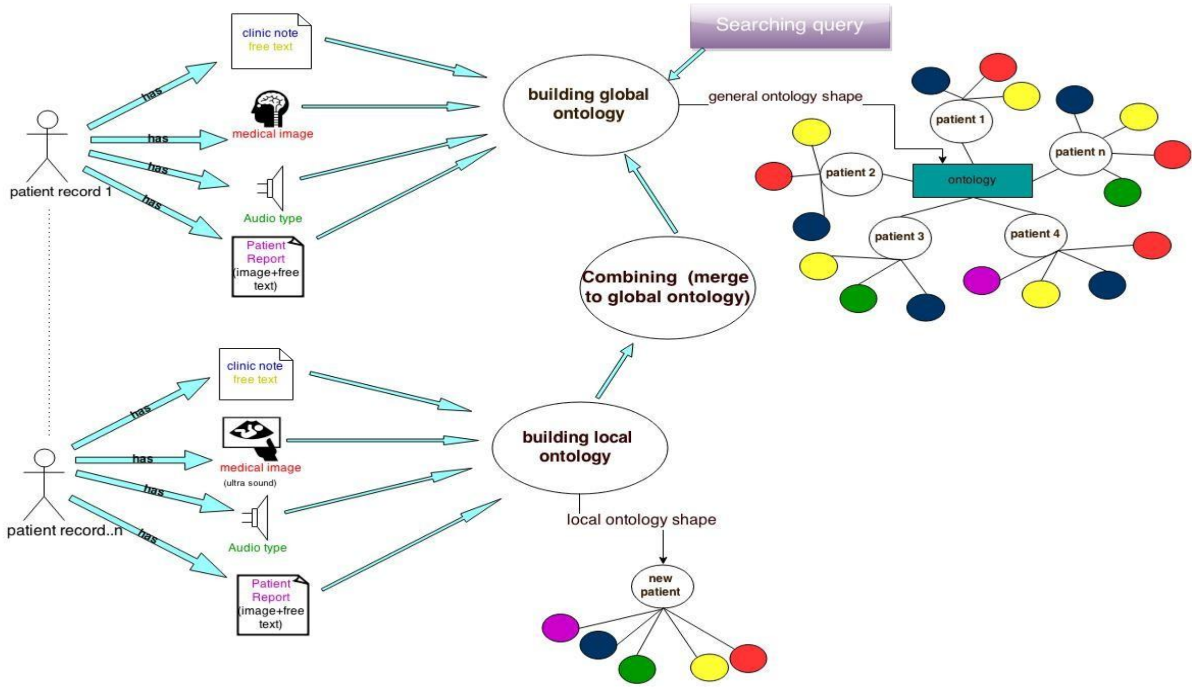


Fig. 2 The proposed framework

A. *Clinical text - free form*

Clinical text has wealthy detailed information of great possibility usage to scientists and health service researchers. Text schema describes the structure of a text file and how a text document is read or written in a raw format. The structure of text stream defines either fixed column widths or columns which are separated by delimiters. To convert a text schema to XML Schema, a specified separator, field separator and the field names of the text file should be determined. In other side, XML Schemas contain annotations for providing additional information, such as medical information. The conversion process between XML and CSV has been done automatically by detecting all repeated elements in XML that are used for splitting data to rows.

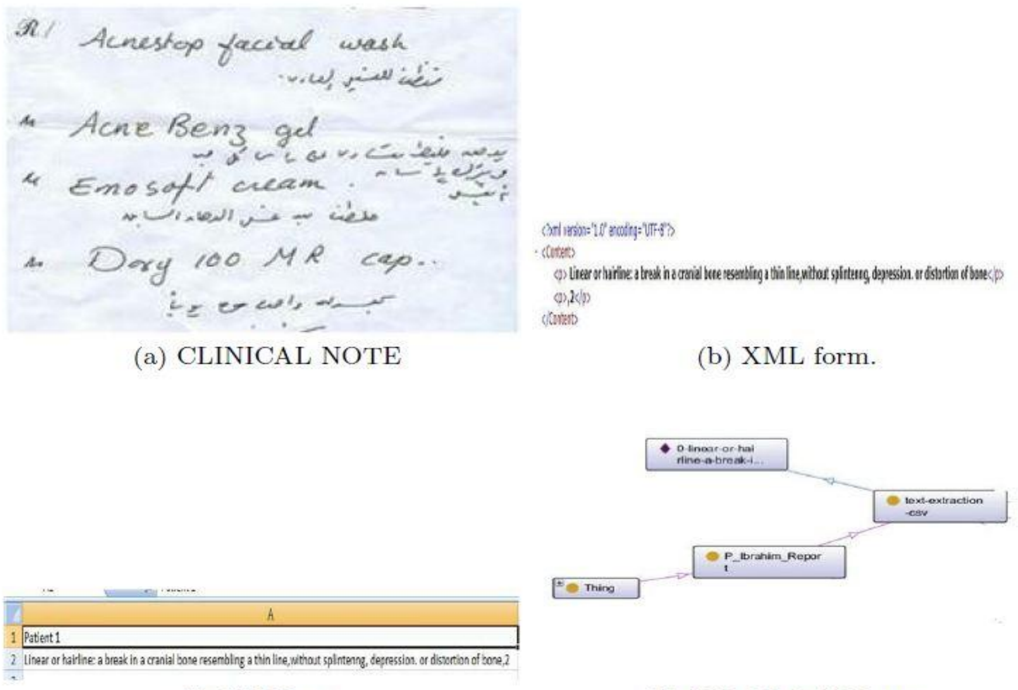
The proposed framework can handle free text as a block. Clinical note is extracted from clinical notes of physician’s prescriptions. While the input is free text, the expected output merges all patient data and performing query to get all data merged for the patient. There are three processes in managing clinical notes. Firstly, extracting clinical notes process has done from physician’s prescriptions and the output of this process is in unstructured form. Secondly, transforming process for unstructured form has done to get semi-structured form as to facility the dealing when building ontology. Finally, transforming XML form to get a structured form, as in "CSV file" which allow building RDF "ontology" file where queries can be carried out to access the medical data, see Fig. 3.

B. *Medical image data type*

The second form is medical images in healthcare application. The major challenge in this case is how to handle these large numbers of images with their various formats and then merging them with other medical data types. The proposed framework helps in solving this challenge by building ontology for these images and accesses any of images by its label through the global ontology. Physicians can get medical image in ontology by its label or URL, see Fig. 4(a). DICOM images as example consist of two parts, i.e., text header and binary image. Medical images metadata as well as field names or image’s URL are indexed and transformed to XML elements.

C. *Audio media data type*

Audio media type is represented in health care as medical diagnosis from physicians a broad or encounters between physicians and patients. There are five processes to complete this stage. The first process is performing speech to text process. The second process is acting as the first process in free text data type. Output from the first process should be converted to semi-structured form as XML form. Then a structured form represented in CSV file has been created and global ontologies have been built, which we can perform query simply to access all media types.



(a) CLINICAL NOTE

```
<?xml version="1.0" encoding="UTF-8"?>
- <Content>
    <p> Linear or hairline: a break in a cranial bone resembling a thin line,without splintenng, depression. or distortion of bone</p>
    <p>,2</p>
  </Content>
```

(b) XML form.

	A
1	Patient 1
2	Linear or hairline: a break in a cranial bone resembling a thin line,without splintenng, depression. or distortion of bone,2

(c) CSV form.

(d) ONTOLOGY form.

Fig. 3 Clinical note processing cycle

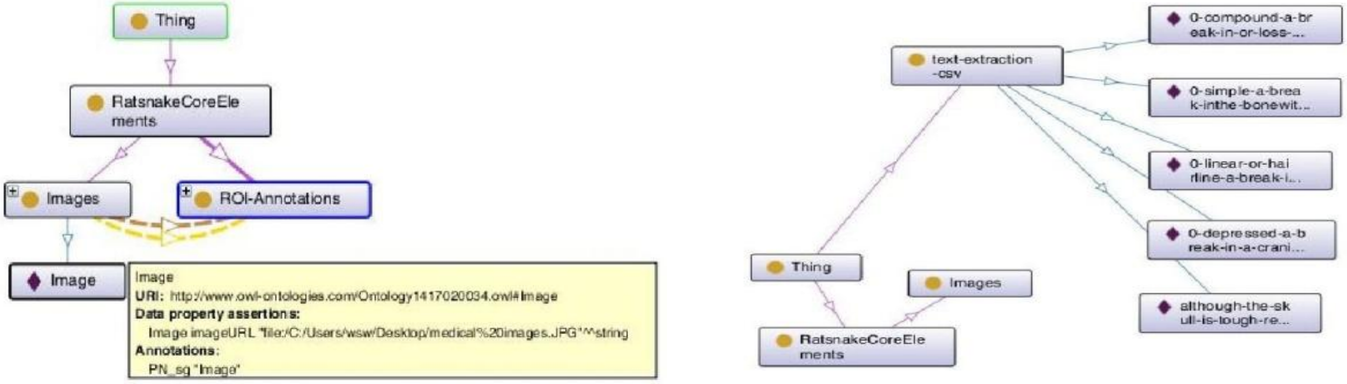


Image
URI: http://www.owl-ontologies.com/Ontology1417020034.owl#Image
Data property assertions:
Image imageURL "file:/C:/Users/wsw/Desktop/medical%20images.JPG"^^string
Annotations:
PN_sg "Image"

(a) Ontology of Medical image (b) Ontology of patient report

Fig. 4 Both ontology of medical images and patient report

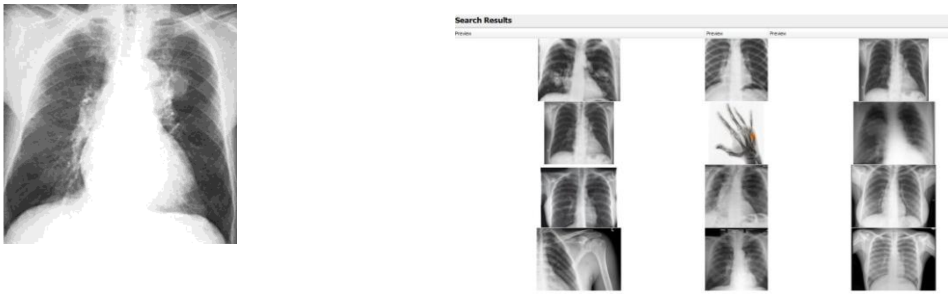
D. *The medical patient report*

Output report may contain free text and medical images. Thus, images are extracted firstly from reports and then ontology is created for both image and extracted free text. This case involves handling both free text and medical images as discussed previously. The proposal framework deals with patient report depending on the extracted free text of the report and creates ontology for free text see Fig. 4(b). Medical images have been handled as a normal image, but also we can get pure image and build its ontology.

VI. Implementation and Discussion

The proposed framework is implemented with freely and open-source tools, *i.e.*, semantic Web technologies. The proposed ontology is configured as follows: number of classes, number of individuals and number of properties are 21,58 and 68, respectively. Moreover, maximum depth, maximum number of children and average number of children are 5,75 and 4, respectively. Indeed, a local ontology for each media type of medical information is created, and then such local ontologies are merged into global one. The global ontology being built is efficiently scalable; new patient records can be added and merged easily. Thus, all patients' medical information is integrated into their history without loss of any data. The global ontology merges other healthcare domain ontology such as accounts, hospital budget, geographical places analogies, etc. By merging all these information, a background for patient history is complete, which can help physicians in decision making. Furthermore, the technician can apply queries against the global ontology using either URI, labels, DL language or SPARQLE. Searching inside the ontology is tested several times, the searching process is simple, and the results are returned quickly and accurate as shown in Fig. 5. The main challenge in retrieval system was the correlation among medical concepts. Fortunately, the proposal framework tackles this challenge by adding rules and relations among concepts. In Fig. 6, three patient have a problem in lung, but two of them have cancer in lung. Therefore, physicians get alarm that patient 3 may has cancer in lung. Accordingly, the proposal framework can help physicians in decision making.

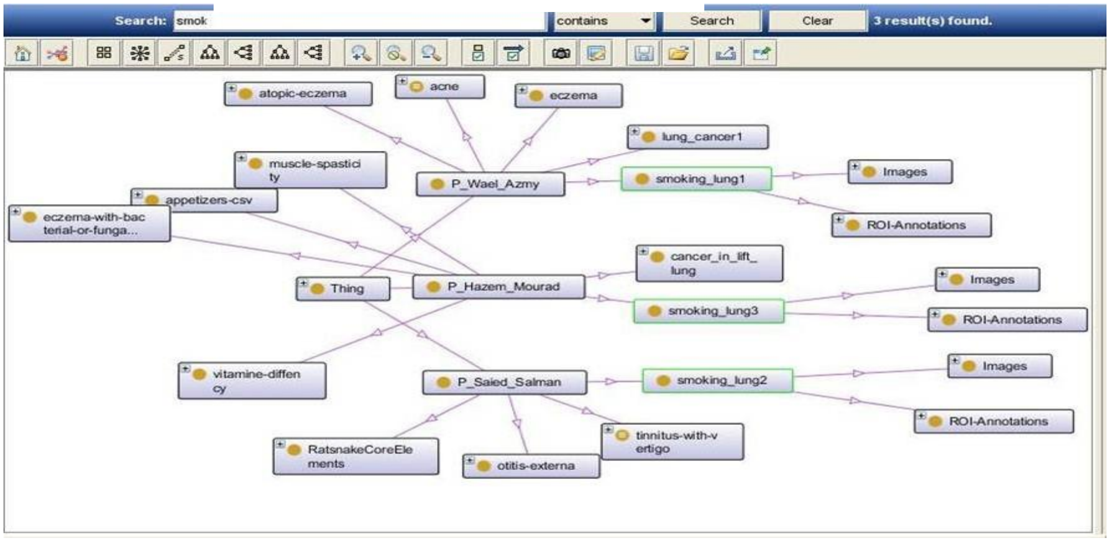
To implement CBIR approach for comparison with the proposed framework, dataset of around 90 medical images is used. Three different categories are used including several datasets. The size of the first dataset is 4.6M including "20 image" and the size for each one is about 232 K. The size of the second dataset is 6.1M including "23 image" with size for each one is 266 K. The third dataset is 12.9M including "25 image" and the size for each one is 516 K. The size of the fourth dataset is 13.4M including "25 image" with size for each one is 548K. Note that, the main problem facing retrieval systems is semantic gap and relationships among medical concepts [8]. The proposed framework tackles this issue including semantic gap and related concepts. CBIR systems retrieve medical images according to the distance similarity between the query image and the dataset as shown in Fig. 5. By the experiments, CBIR query is an image showed in Fig. 5 (a). The results of CBIR system depend on the distance similarity between the query image and the dataset. Fig. 5(b) shows the result from CBIR. However, it can be noticed easily that query is a chest medical image has cancer and the results show different images including broken hand, thus it leads to low accuracy.



(a) Query of CBIR

(b) Result of CBIR

Fig. 5: CBIR-based retrieval


Search: smok

contains

3 result(s) found.

Fig. 6 Relationships of medical concepts

However, the proposed framework retrieves all medical image according to their labels. For example, if we apply searching against the ontology using label or annotator "smoking lung", physicians retrieve medical images of three patient have a smoking lung. Moreover, physicians can observe that patient 1and 2 have a ray cancer from the retrieved results. Thus, they conclude that patient 3 may be infected by cancer as shown in Fig. 3. Therefore, the proposed framework exploits related concepts and helps physicians in decision making. To calculate the accuracy of the proposed framework against CBIR, Fig. 7 shows the recall, precision and consumed time for transforming images into XML format. Generally, results indicate that the proposed framework outperforms CBIR. Precision is the fraction of retrieved images that are relevant to physician’s information need. The precision of the proposed frame work is better than CBIR, although the proposal framework takes into account the human errors. Moreover, there call of the proposed framework outperforms the CBIR due to challenges of CBIR which tackled by the proposed framework. Generally, results in TABLE II indicate that the proposed framework outperforms CBIR. Precision is the fraction of retrieved images that are relevant to physician’s information need.

Annotation approach implementation - Now, we highlight some implementation observations and differences between the proposed framework and other literature approaches. For instance, NLP approaches apply POS tagging to get annotation for all tokens such as NN, MV, etc. However, when applying POS tagging, massive data are lost particularly the medical concepts in addition to the lack of relationship information among such concepts. Compared with the proposed framework, there is no data loss. The annotation approach had to find syntactic structure as possible. The main units of annotation are adopting chunk. The Harvey corpus is used to deal with free text. The Harvey corpus is a chunk-annotated corpus. Chunking tends to solve this challenge using part of speech (POS) tagging. The main problems are caused by unknown tokens which caused obscurity due to neglected words or phrases. Generally, the main challenge in the annotation approach is to get the similarity between encoding enough information to improve the research, and realizing clarity, accuracy, and conciseness [18]. The two typical examples (without chunk annotation for clarity) and example (3) below illustrates the chunking annotation process.

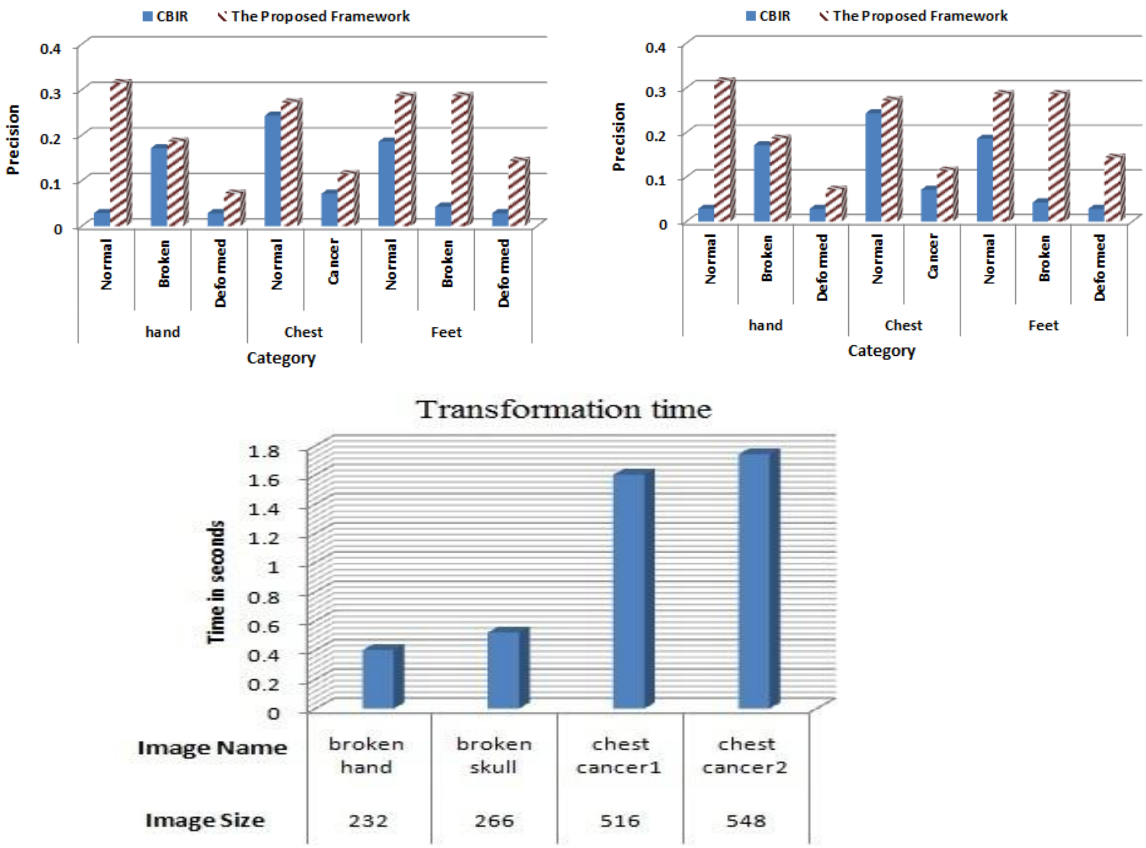
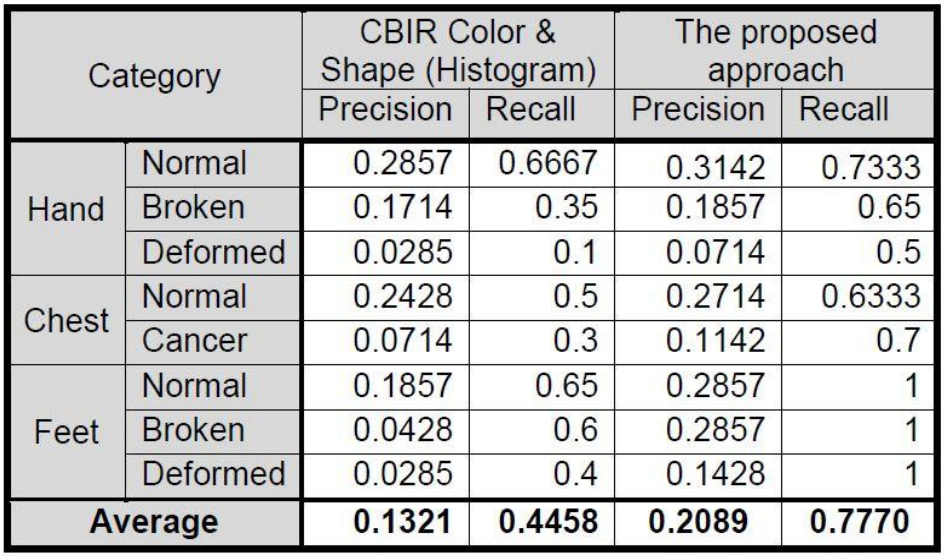


Fig. 7 Precision and Recall of the proposed framework vs. CBIR system

TABLE II RESULTS OF PRECISION AND RECALL FROM THE PROPOSAL APPROACH



Category		CBIR Color & Shape (Histogram)		The proposed approach	
		Precision	Recall	Precision	Recall
Hand	Normal	0.2857	0.6667	0.3142	0.7333
	Broken	0.1714	0.35	0.1857	0.65
	Deformed	0.0285	0.1	0.0714	0.5
Chest	Normal	0.2428	0.5	0.2714	0.6333
	Cancer	0.0714	0.3	0.1142	0.7
Feet	Normal	0.1857	0.65	0.2857	1
	Broken	0.0428	0.6	0.2857	1
	Deformed	0.0285	0.4	0.1428	1
Average		**0.1321**	**0.4458**	**0.2089**	**0.7770**

VII. CONCLUSION

Managing unstructured clinical data is one of the major problems in healthcare systems. The heterogeneity of clinical data is considered as a critical roadblock to achieving integration and interoperability between systems. In this paper, we proposed a semantic-based framework for managing heterogeneous medical data including free clinical notes, audio data, and medical images. It tackles the heterogeneity by unifying different medical data types into a unified form and building ontology. The ontology keeps all patient history and enables queries for patient and diseases history. The manipulation of the proposed framework is demonstrated by the different medical cases.

The future step is to handle medical videos, which can be transformed into medical images, and integrate them with different medical data types. Moreover, we plan to enrich the framework with spatial and temporal information of patients to discover new insights from analytics.

REFERENCES

[1] M.Alkhawlani, M.Elmogy, and H.El Bakry. Text-based, content-based, and semantic-based image retrievals: A survey. International Journal of Computer and Information Technology (ISSN: 2279,0764) Volume04?Issue 01, January 2015.

[2] Belle, R. Thiagarajan, S. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian. Big data analytics in healthcare. Biomed research international, 2015.

[3] D. P. Bhamare and S. A. Abhang. Content based image retrieval: A review. International Journal Of Computer Science And Applications, 8(2), 2015.

[4] A. L. Buczak, S. Babin, and L. Moniz. Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making, 10(1):59, 2010.

[5] R. Chaudhari and A. Patil. Content based image retrieval using color and shape features. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 1(5), 2012.

[6] N. Grover. 'big data'-architecture, issues, opportunities and challenges. IJCER,3(1):26–31, 2014.

[7] L. Haldurai and V. Vinodhini. A study on content based image retrieval systems. International Journal of Innovative Research in Computer and Communication Engineering , Vol. 3, Issue 3., March 2015.

[8] R. Jobay and A. Sleit. Quantum inspired shape representation for content based image retrieval. Journal of Signal and Information Processing, 5(02):54, 2014.

[9] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq. Challenges of data integration and interoperability in big data. In Big Data (Big Data), 2014 IEEE International Conference on, pages 38–40. IEEE, 2014.

[10] L. Kang, L. Yi, and L. Dong. Research on construction methods of big data semantic model. In Proceedings of the World Congress on Engineering, volume 1,2014.

[11] A. Katal, M. Wazid, and R. Goudar. Big data: Issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404–409. IEEE, 2013.

[12] H. Kaur and K. Jyoti. Survey of techniques of high level semantic based image retrieval. International Journal of Research in Computer and Communication technology, IJRCCT, ISSN 2278-5841, Vol 2,, 2(1):015–019, Issue 1, January 2013.

[13] R. Kienast and C. Baumgartner. Semantic data integration on biomedical data using semantic web technologies. INTECH Open Access Publisher, 2011.

[14] P. Kulkarni, S. Kulkarni, and A. Stranieri. A novel architecture and analysis of challenges for combining text and image for medical image retrieval. International Journal for Infonomics (IJI), 2014.

[15] S. J. Pooja, Reema Gupta. Big data: Advancement in data analytics. International Journal of Computer technology and applications, 2014.

[16] K. Priyanka and N. Kulennavar. A survey on big data analytics in health care. International Journal of Computer Science and Information Technologies 5(4):5685–5688, 2014.

[17] R. Rahimzadeh, A. Farzan, and Y. F. Fathabad. A survey on semantic content based image retrieval and CBIR systems. International Journal on "Technical and Physical Problems of Engineering" (IJTPE) Published by International Organization of IOTPE, March 2014.

[18] S. Sasikala and R. S. Gandhi. Efficient content based image retrieval system with metadata processing. International Journal for Innovative Research in Science and Technology, 1(10):72–77, 2015.

[19] A. Savkov,J. Carroll, and J. Cassell. Chunking clinical text containing non- canonical language. ACL 2014, page 77, 2014.

[20] D. S. Seema H. Jadhav1, Dr.Sunita Singh. Content based image retrieval system with semantic indexing and recently retrieved image library. International Journal of Advanced Computer Technology (IJACT), 2012.

[21] J. Sun and C. K. Reddy. Big data analytics for healthcare. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1525–1525. ACM, 2013.

[22] O. Uzuner, M. Yetisgen, and A. Stubbs. Biomedical/clinical NLP. COLING 2014, pages 1–2, 2014.

[23] C. Y. Lee, H. Ibrahim, M. Othman, and R. Yaakob. Reconciling semantic conflicts in electronic patient data exchange. In Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, pages 390–394. ACM, 2009.

[24] J. Partyka, L. Khan, and B. Thuraisingham. Semantic schema matching without shared instances. In Semantic Computing, 2009. ICSC'09. IEEE International Conference on, pages 297–302. IEEE, 2009.

[25] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Qing Yu, D. Giordano, I. Marenzi, and B. Pereira Nunes. Interlinking educational resources and the web of data: A survey of challenges and approaches. Program, 47(1):60–91, 2013.

[26] N. Chen, J. He, C. Yang, and C. Wang. A node semantic similarity schema matching method for multi-version web coverage service retrieval. International Journal of Geographical Information Science, 26(6):1051–1072, 2012.

[27] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data(TKDD), 2(2):10, 2008.

[28] D. Ramesh and C. Kumar. Schema integration based merging and matching algorithm for agricultural HDDBs. Arabian Journal for Science and Engineering, 40(9):2555–2569, 2015.

[29] C. Bizer and A. Schultz. The R2R framework: Publishing and discovering mappings on the web. In Proceedings of the First International Conference on Consuming Linked Data-Volume 665, pages 97–108. CEUR-WS. org, 2010.

[30] L. C. Keung, S. Niukyun, J.-F. Ethier, L. Zhao, V. Curcin, and T. N. Arvanitis. The integration challenges in bridging patient care and clinical research in a learning healthcare system. 2014