Big Data Integration Case Study for Radiology Data Sources

Priya Deshpande, Alexander Rasin, Eli Brown,
Jacob Furst, Daniela S. Raicu
DePaul University
Chicago, USA
{*pdeshpa1, arasin, ebrown80, jfurst, dstan*}*@depaul.edu*

Steven M. Montner, Samuel G. Armato III
University of Chicago, Department of Radiology
Chicago, USA
smontner@radiology.bsd.uchicago.edu,
s-armato@uchicago.edu

***Abstract*—Today's digitized world urgently needs Big Data integration and analysis. Healthcare records are responsible for generating petabytes of data in a single day. Such data is heterogeneous in nature, captured in different files and formats, and varies from hospital to hospital. By integrating data from different sources and extracting meaningful information for the medical community, we can improve the overall quality of patient care. Our research targets the problem of integration for health records. To start, we already developed the Integrated Radiology Image search (IRIS) engine, which could represent a data integration framework for the healthcare domain. IRIS provided support for multiple public data sources and incorporated medical ontologies which would help radiologists and improve search interpretation by considering the meaning of the search query terms. In this paper, we describe a case study of data integration for radiology data sources. While the need for data integration is self-evident, we learned that rather than being a single step, data integration is an iterative process that requires continuous integration of metadata and additional supporting data sources. Our results show that an each step of data integration further improved IRIS engine results.**

1. Introduction

In today's digital world, we are generating large amount of data on daily basis. Social media, Healthcare, Government, Internet of Things and many more systems generate tremendous amount of structured and unstructured data. While large amounts of data are available, we are lagging in performing meaningful analysis of this data. To extract meaningful information from diverse data sources, one of most important requirements is developing a data integration framework and combining data content into a common repository. Data preparation is a mandatory first step before starting any analysis, which consists of: 1) finding and collecting relevant data, 2) cleaning and integrating data, and 3) managing data for analysis. Research groups and labs talk about data integration and today's need of data integration. However, are we really addressing these issues? There are many challenges which vary among applications that need to be addressed by the database community to help identify and use information extracted from these data sources. If we consider the healthcare domain alone, electronic health records constitute petabytes of data throughout the world in a single day. This data has all of the Big Data characteristics – high volume, variety and velocity. In terms of volume, each patient record might have images, clinical reports, and pathology reports which all result in a high volume of data. The data is heterogeneous in nature, captured in different files and formats, and varies from hospital to hospital despite standard data storage (e.g., PACS system for image storage). Considering the characteristics of medical data and developing a custom-made analytic tools will ultimately help medical society improve the diagnostic process and the overall quality of patient care. There is a huge demand for efficient integration of medical health records. Currently available electronic health records systems are lagging to incorporate pathology reports, radiology reports with a high volume of images and clinical finding reports [1].

To start with integration of medical records we initiated our research work by focusing on radiology teaching files. In addition to pedagogy, teaching files are widely used by radiologists as a resource in the diagnostic process. Teaching files contain images, recorded discussion and notes, external references, augmenting annotations, patient history, and associated images. In our study of the many publicly available data sources and in-house teaching files repositories, we found that these sources are highly heterogeneous and difficult to access in practice. Integration of these heterogeneous data sources is thus of great use to radiologists. We developed IRIS engine, which is an integrated repository of radiology teaching cases supporting natural language query interpretation and an image-based search.

In this paper we include a full overview of publicly available data sources for radiology domain, including both their advantages and limitations. We discuss our integration of several sources into our repository, our search engine design and a preliminary evaluation of its search capability. In our system, radiologists will be able to easily contribute new cases or augment existing cases by supplying additional comments and annotating images in the single shared repository. We integrated two teaching file data sources Radiology Society of North America Medical Imaging Resource Community (RSNA MIRC) and MyPacs along with two

978-1-5386-6709-5/18/$31.00 ©2018 IEEE

ontologies Radiological Lexicon (RadLex) [2] and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [3]. These ontologies provides a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information. In our system we integrated radiology teaching file data sources and, while integrating these data sources, we performed experiments to evaluate the accuracy of search results. This evaluation concluded that integration of medical ontologies is necessary to improve the search quality (the results are discussed in Section 4). Successful ongoing integration of medical ontologies demonstrates that data integration is a continuous process – integrating data sources (with teaching files) is not sufficient without also integrating the related metadata sources. We further incorporated support for content based image retrieval (CBIR) and observed that searching medical image data sources enabled us to get better results. In this paper, we discuss the need of data integration for healthcare domain and how metadata further supports this process.

In Section 2 we describe RSNA MIRC, MyPacs, RadLex and SNOMED CT ontologies, along with prior data integration work in the radiology domain. In Section 3 we focused on data integration methods used in this research work. In Section 4 we discussed our current results to show how integration of data sources and medical ontologies can help radiologists in the diagnostic process. We expect to speed up reference search for radiologists by providing them with an integrated teaching file database solution. Otherwise they may have to refer to different heterogeneous sources, making it difficult to find and retain information. Overall, this case study shows that data (and metadata) integration improves the search accuracy and performance. In Section 5 we summarize the conclusions of this work and describe planned future work.

2. Related work

Our literature survey is based on articles from Journal of Radiology, Radiographics, Digital Imaging, IEEE and other established medical publication venues. We reviewed papers that discussed the need for big data integration of health care systems. There are many papers that argue the need for big data utilization and disparate source integration to better serve the medical field, which greatly inspired us to proceed with building IRIS engine. Ron Gutmark [4] argued for building a system that reduces errors in radiological images using teaching file database. Easy-to-use computer teaching files are useful for training physicians, serve as a reference tool for experienced physicians and help them improve diagnostic accuracy. The work in [5] discussed how critical radiologic images are for diagnosis, teaching needs and research. They were particularly interested in using case-based radiologic teaching files for radiology teaching. Their proposed architecture was meant to be integrated with existing medical image databases (featured by MIRC interoperability), but it is not publicly available. Availability of a large and diverse set of clinical cases need the integration of profiles published by Integration Healthcare Enterprise (IHE) [6]. Having a repository of pathology-proven cases in a dashboard also has the potential to enhance and encourage the formation of accurate teaching files, as well as educational publications in the form of case series or "case of the day" submissions [7]. As the use of positron emission tomography computed tomography (PET-CT) has increased rapidly, there is a need to retrieve relevant medical images that can assist an image interpretation. Building a database which may provide integrated repository with images to improve diagnosis accuracy [8]. Larger clinical reference datasets that are relevant to a larger number of patients may help to retrieve complex query results. (e.g., "retrieve the PET-CT study containing the lymph node lesion, which showed no interval change for more than 2 years"). Data integration and a centralized data repository for clinical data, patient history, physical exam findings, laboratory data, imaging data is important as a reference during the diagnostic process. Authors of [9] discussed how big data analysis could be helpful for radiologist daily work. From our survey we can conclude that in radiology there is a need to integrate clinical reports and images and develop a unified reference database. The following is the list of sources and repositories we have evaluated in determining what databases are currently available to radiologists. We intend to integrate these sources as we have integrated RSNA MIRC, MyPacs, RadLex and SNOMED CT.

RSNA MIRC: It is a large repository with 2,500 teaching files including the information about the history of patients, diagnosis, differential diagnosis, findings, discussion as well as external references (journal articles). Radiological terms are highlighted and linked to RadLex browser (see discussion about RadLex below). However, search is done verbatim with no processing to interpret the goals (e.g,. synonyms, negation). No image-based search is possible.

Mypacs.net [10] Publicly available teaching file resource. In total more than 37,000 cases are available with 200,000 images (18,000 public cases). User can search records based on anatomy, pathology, modality, age, gender, etc. Limitations of this search engine include lack of consideration for synonyms, negation, or image-based search.

RadLex [2] Radiology Lexicon term browser. RadLex is an ontological system that provides a comprehensive lexicon vocabulary for radiologists. RadLex browser was developed by RSNA and includes more than 45,000 unique terms.

SNOMED CT [3] ontology provides a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other healthcare providers for the electronic exchange of clinical health information. The SNOMED CT ontology follows the National Library of Medicine (NLM) Unified Medical Language System format; it has a hierarchical structure and includes clinical findings, anatomy, test findings, and morphological connections. This ontology covers more than 300,000 terms with preferred name, synonyms, definition, and semantic meaning, making SNOMED CT a comprehensive, computerized healthcare terminology.

Open-i [11] Open Access Biomedical Image Search En-

gine of the National Library of Medicine enables search and retrieval of abstracts and images (e.g., charts, graphs, clinical images) from the open source literature and biomedical image collections. Searching may be done using text queries as well as query images. Open-i provides access to over 3.7 million images from about 1.2 million PubMed Central® articles. Open-i is great source of image collection, however this data source does not include categories such as history or diagnosis information for the patient case.

EURORAD (European Society of Radiology) [12] is a peer-reviewed educational tool based on teaching cases. There are a more than 7,000 teaching cases – similar to other teaching file sources there is no support for negations,synonyms, or image-based search. There are many other radiology data sources available publicly such as Goldminer, Yottalook, Radiopaedia.org that need to be integrated to provide an integrated search engine for radiology society. In our previous work we discussed these data sources in detail [13]. IRIS integrated with two major data sources RSNA MIRC and MyPacs.net and two medical ontologies RadLex and SNOMED CT.

3. Proposed system

Our research started with the integration of radiology teaching files, including all the different categories they contain (e.g., findings, diagnosis). Along with teaching files, there is a need to integrate pathology reports as well. Pathology reports are the laboratory test results and information about the size, shape, and appearance of tissue sample analyzed by pathologists. Studies have shown that within the same hospital, radiology and pathology reports are not currently integrated [14] and thus not leveraged to improve diagnosis process. Because both radiologist's and pathologist's data are essential to better diagnosis and patient treatment decisions, this isolation of radiology and pathology workflow can be detrimental to the quality and outcomes of patient care. This underscores the need for pathology and radiology workflow integration and for building systems that facilitate the synthesis of all data produced by both specialties. With the enormous technological advances currently occurring in both fields, the opportunity has emerged to develop an integrated diagnostic reporting system that supports both specialties and, therefore, improves the overall quality of patient care. To start with integration of healthcare data sources, we developed IRIS [13] engine which could serve as a template of a data integration system in healthcare domain. We integrated support for medical ontologies – which would help radiologists interpret clinical reports by considering context behind the radiology terms. IRIS integration can be further applied to other healthcare domains such as surgery, where doctors can refer to radiology reports or pathology reports when performing surgery and for follow up treatment decisions. We are aiming to develop a data warehouse system, which can provide doctors with access to important clinical information across heterogeneous data sources, and supplement it with ontology powered search. Our current radiology search engine supports natural language queries, image-based queries, and hybrid (image and text based) search. Our goal is to integrate all of the available public sources and let users retrieve results by augmenting searches with synonyms and correctly interpreting negation and adjectives. In our database system we captured the data from these publicly available sources and cleaned the data to a normalized schema before loading it. As shown in Figure 1, our logical schema supports integration of heterogeneous data sources. Central entity is based on radiology teaching file information along with image features that stores vector of image features. We also include pathology, patient, doctors, and diagnosis information, so we can expand our study for a wider healthcare domain. Integration of these teaching files is a challenging task, as all the data sources available are in heterogeneous form with different files and formats. Data cleaning is an important step we implemented, including handling of missing values, fixing invalid values (e.g., 200000-10-03 to 2000-10-03), and finding common patterns to enforce the HIPAA constraints on personal patient information. After integration

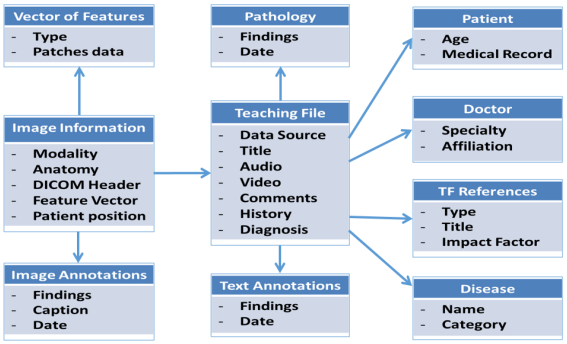

Figure 1: IRIS Logical Schema

of MIRC and MyPacs datasets, to improve search results we first integrated RadLex ontology with definitions for over 45,000 terms. However, our analysis after RadLex integration forced us to consider adding more ontologies, as our searches were missing important results. To evaluate the need for additional ontologies and to understand integrated data sources, we performed coverage and cluster analysis of the data; i.e. how many ontology terms were covered in each datasets and how well ontology terms cover dataset unique terms. Coverage analysis showed only a 5% overlap is present between different ontologies. Based on this analysis we integrated an additional SNOMED CT ontology which improved the number of relevant search query results by more than 150% [15]. IRIS system also supports performing an image-based search. CBIR is performed by using an image feature extractor to extract latent features from our images along with a mechanism to quantify the similarity or dissimilarity between the features from the query image and the images in our database. We constructed an image feature extractor using a convolutional autoencoder. For supervised machine learning algorithm we needed labelled data to train the model. However, MIRC and MyPacs dataset have only 1,100 Digital Imaging and Communications in

Medicine (DICOM) images that provides image modality as a label, the remaining 90,000 images are JPEG or PNG format. The lack of labelled data encouraged us to integrate labelled modality images. We used ImageCLEF [16] dataset that provided 5,000 of modality labelled images, further demonstrating that data integration is an iterative process.

4. Results

In this section we present results from integration of additional data sources. Initially, we used a naive method for data integration (without our proposed logical schema or explicit integration). This method involved comparing the query term in each teaching file body (text) in the database; it was not only time consuming but also error prone. For example, for a "renal artery" query the naive approach would have to match both words in the text exactly; we could also search for individual words ("renal" and "artery"), but that would generate too many false-positives. Results discussed here use 5 sample queries to illustrate how integration of an additional ontology improved IRIS results. We compared our initial IRIS search (IRIS 1.0 with RadLex ontology) with new IRIS 1.1 (with RadLex and SNOMED CT ontologies). Table 1 shows that adding another ontology greatly improved search results. Search for "chiari" produced 153 results in IRIS 1.0; however, adding a second ontology improved results by 39 matches. After query expansion with "hindbrain hernia" and "arnold–chiari malformation" synonyms, the search resulted in 192 relevant teaching files. This search was able to find so many matches by applying both ontologies. "Hindbrain hernia" is a synonym from RadLex ontology and which is not present in SNOMED CT ontology. Similarly "arnold-chiari malformation" is from SNOMED CT, this synonym is not present in RadLex ontology. Benefits of integrating radiology reports, pathology reports, medical ontologies, and support of search engine will improve clinical decision making and reduce innate human memory errors. Creation of a single consolidated health record database will save doctor's time in diagnosis and interpretation of medical reports.

Table 1: IRIS results before and after ontology integration

Query	IRIS 1.1	IRIS 1.0
Chiari	192	153
Cardiomegaly	169	158
Bronchus Intermedius	5	3
Tracheal dilation	986	758
Angiosarcoma	126	27

5. Conclusion

Big data integration is one of the most demanding requirements in healthcare domain. Radiology integrated system approach discussed here can be further applied to other healthcare domains such as surgery, which will help doctors to offer more accurate and timely care by providing a reliable reference database. We believe that our integrated data warehouse system search engine would also help in education and research in healthcare domain. We presented IRIS project at the annual SIIM 2018 meeting (as posters) and received feedback from doctors indicating that this work is highly useful for practitioners in the medical domain. Future work on this project would allow domain experts to integrate their own teaching cases and annotate images with further metadata. We would like to expand this study to integrate electronic health records and continue integrating other publicly available medical sources into our database repository.

References

[1] E. R. J. Walton and S. Gandhi, "Record keeping in radiology: are we doing enough?" *The British Journal of Radiology*, 2016.

[2] RSNA, "Radlex ontology," http://www.radlex.org/, February 11, 2018.

[3] S. I. I. H. T. S. D. Organization, "Snomedct ontology," http://www.snomed.org/, February 11, 2018.

[4] R. Gutmark, M. J. Halsted, L. Perry, and G. Gold, "Use of computer databases to reduce radiograph reading errors," *Journal of the American College of Radiology*, vol. 4, no. 1, pp. 65–68, 2007.

[5] R. Talanow, "Radiology teacher: a free, internet-based radiology teaching file server," *Journal of the American College of Radiology*, vol. 6, no. 12, pp. 871–875, 2009.

[6] M. Dos-Santos and A. Fujino, "Interactive radiology teaching file system: the development of a mirc-compliant and user-centered e-learning resource," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 5871–5874.

[7] L. R. Margolies, G. Pandey, E. R. Horowitz, and D. S. Mendelson, "Breast imaging in the era of big data: structured reporting and data mining," *American Journal of Roentgenology*, vol. 206, no. 2, pp. 259–264, 2016.

[8] K. H. Hwang, H. Lee, G. Koh, D. Willrett, and D. L. Rubin, "Building and querying rdf/owl database of semantically annotated nuclear medicine images," *Journal of Digital Imaging*, pp. 1–7, 2016.

[9] A. P. Kansagra, J. Y. John-Paul, A. R. Chatterjee, L. Lenchik, D. S. Chow, A. B. Prater, J. Yeh, A. M. Doshi, C. M. Hawkins, M. E. Heilbrun *et al.*, "Big data and the future of radiology informatics," *Academic radiology*, vol. 23, no. 1, pp. 30–42, 2016.

[10] M. M. I. Group, "Mypacs tfs," https://www.mypacs.net/, February 31, 2018.

[11] NIH, "Openi," https://openi.nlm.nih.gov/, February, 11, 2018.

[12] E. S. of Radiology Neutorgasse, "Eurorad," http://www.eurorad.org/, May 21, 2018.

[13] P. Deshpande, A. Rasin, E. Brown, J. Furst, D. Raicu, S. Montner, and S. A. III, "An integrated database and smart search tool for medical knowledge extraction from radiology teaching files," vol. 69, pp. 10–18, 14 Aug 2017. [Online]. Available: http://proceedings.mlr.press/v69/deshpande17a.html

[14] J. Sorace, D. R. Aberle, D. Elimam, S. Lawvere, O. Tawfik, and W. D. Wallace, "Integrating pathology and radiology disciplines: an emerging opportunity?" *BMC Medicine*, vol. 10, no. 1, p. 100, Sep 2012. [Online]. Available: https://doi.org/10.1186/1741-7015-10-100

[15] P. Deshpande, A. Rasin, E. Brown, J. Furst, D. Raicu, S. Montner, and S. A. III, "Augmenting medical decision making with text-based search of teaching file repositories and medical ontologies: Text-based search of radiology teaching files," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, yet to appear.

[16] H. Mller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 2012.