Biomedical Data Categorization and Integration using Human-in-the-loop Approach

Priya Deshpande
Supervised by Dr. Alexander Rasin
DePaul University
Chicago, IL, USA
pdeshpa1@depaul.edu

ABSTRACT

Digitized world demands data integration systems that combine data repositories from multiple data sources. Vast amounts of existing clinical and biomedical research data are considered a primary force enabling data-driven research toward advancing health research and for introducing efficiencies in healthcare delivery. Data-driven research may have many goals, including but not limited to improved diagnostics processes, novel biomedical discoveries, epidemiology, and education. However, finding and gaining access to relevant data remains an elusive goal. We identified different data integration challenges and developed an Integrated Radiology Image Search (IRIS) framework that could be a step toward aiding data-driven research. We propose building a biomedical data categorization and integration framework using human-in-the-loop and developing data bridges to support search and retrieval of relevant documents from the integrated repository.

My research focuses on biomedical data integration, indexing systems, and providing relevance-ranked document retrieval from an integrated repository. Although we currently focus on integrating biomedical data sources (for medical professionals), we believe that our proposed framework and methodologies can be used in other domains as well.

1. INTRODUCTION

A growing amount of available biomedical data poses new challenges in data management. Data re-usability is a highly desirable goal, both for advancing science as well as for replicating or validating results of previous studies. Recognizing this need, publishers and funding bodies may require researchers to submit data generated in their work and make it available to the research community. For example, National Institutes of Health (NIH) is encouraging funded investigators to use cloud computing to conduct research and make their work accessible to larger audiences[1].

[1]https://commonfund.nih.gov/strides/

Proceedings of the VLDB 2019 PhD Workshop, August 26th, 2019. Los Angeles, California.

However, in the healthcare domain, datasets are often not shared because of security concerns, lack of integration, or limitations of retrieval engines. A data integration framework should make data available, accessible, and support fine-grained access control for different users [6]. It would also greatly reduce the need for manual curation of data sources and data repositories. Data integration alone is insufficient without associated information retrieval mechanisms that would rank retrieved results based on relevancy. From our discussions with University of Chicago (UofC) radiologists, even the internal UofC commercial system lacks some of the Natural Language Processing (NLP) features (e.g., detecting synonyms and negation) and multimodal (text and image) search capabilities. We studied publicly available radiology data sources MyPacs.net[2], EURORAD[3], and RSNA Medical Imaging Resource Community (MIRC)[4], that provide a collection of clinical reports and associated images, which are known as *teaching files*. Teaching files contain information such as patient history, findings, diagnosis, differential diagnosis, or discussion notes. While all of these public data sources are available, most of them provide only basic search capabilities – not offering NLP support or ranked retrieval mechanisms. Several studies highlighted the need to integrate clinical reports and images into databases with advanced search capabilities. Gutmark et al. [5] argued for building a system that reduces errors in radiological images interpretation using teaching file databases. Talanow et al. [12] described reference radiological image use for diagnosis, teaching needs, research, and the resulting need for an advanced reference search engine.

An integrated repository of teaching files can retrieve thousands of results for a text search. A search can thus become effectively useless without being able to show the most relevant results first. Publicly available radiology teaching file search engines do not provide text relevance ranking or combined text-and-image search. Lack of such systems motivated us to build Integrated Radiology Image Search (IRIS) and develop the ranking algorithm presented here. We presented IRIS at the annual Society for Imaging Informatics in Medicine (SIIM 2018) meeting (two posters: one focusing on search and another on data integration) and received feedback from doctors indicating that this work would be useful for the medical domain practitioners.

2. BACKGROUND AND RELATED WORK

In this section we discuss papers that addressed the need for data integration and retrieval systems along with an overview of existing medical data retrieval systems. Several studies have highlighted

[2]https://www.mypacs.net/
[3]https://www.myesr.org/eurorad
[4]http://mirc.rsna.org/query

the need for integration of healthcare data [10]. Holzinger et al. [7] talked about knowledge discovery and interactive data mining techniques in bio-informatics, the challenges to integrating biomedical data, and open research directions. Li et al. [8] proposed a hybrid human-machine data integration approach that integrates records from databases with similar data types (e.g., iphone users data). However, healthcare domain data integration needs to combine heterogeneous data sources with different categories of data types. Simpson et al. [11] proposed a multimodal image retrieval system that retrieves biomedical articles used in Open-i[5]. Ling et al. [9] designed GEMINI, an integrative healthcare analytics system, and studied problems related to healthcare data heterogeneity and data integration in that context. From this literature survey, we concluded that healthcare needs are not met by the current search engines. The limitations of existing systems motivated us to design and develop a radiology multimodal search engine. IRIS integrates two well-known public data sources MIRC and MyPacs and two medical ontologies RadLex[6] and The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[7]. RSNA MIRC: Publicly available large repository with more than 2,500 teaching files and more than 12,000 images.
Mypacs.net: Publicly available teaching file resource with more than 35,000 cases and 200,000 images.
RadLex: RadLex is an ontological system that provides a comprehensive lexicon vocabulary for radiologists.
SNOMED CT: 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.

3. METHODOLOGY AND RESEARCH STEPS

In this section, we discuss major biomedical data sources and significant goals that we identified as a part of my PhD proposal.

3.1 Datasets

We currently focus on three types of data a) Electronic health records; b) Radiology teaching files or teaching files used by doctors and radiologists; c) Research datasets.
Electronic Health Records (EHRs): An electronic health record is a digital version of a patient's record. EHRs are maintained at hospitals and provide patient information such as history of patient, medical test results, allergies, immunization details, radiology images, and clinical reports.
Medical Teaching Files: A radiology teaching files system is a collection of important cases for teaching and clinical follow-up. Teaching files share a similar overall structure but significant variations exist even within the same data sources and can include information such as patient history, findings, diagnosis, discussion, comments, references, and images related to clinical reports.
Research datasets: From our survey with different research institute datasets, we observed that most of the data in healthcare domain are images (e.g., CT, X-ray, MRI). Those images are most typically stored in formats such as JPEG, DICOM, or PNG and include associated text data describing patient and case information.

3.2 Data integration and rank retrieval

We have organized this project into three phases (I finished the first two phases and working on the last phase of my PhD work).

[5]https://openi.nlm.nih.gov/
[6]http://www.radlex.org/
[7]https://www.nlm.nih.gov/healthit/snomedct/

.

Table 1: Research work summary

ID	Summary
1	IRIS 1.0 Teaching file text pre-processing and indexing. Smart search through substitution of synonyms and interpreting negation. Query expansion using RadLex through an exact term match. [1]
2	IRIS 1.1 Query synonym expansion. SNOMED CT ontology integration, shown improved results compared with other search engines [3].
3	Data integration as an iterative process, showing how each integration step improved IRIS results [2].
4	Cluster analysis and coverage analysis for both ontologies and radiology data sources. Unsupervised machine learning to identify data source properties – to identify best data sources and ontologies for integration (Journal paper – under review).
5	IRIS 1.2 Multimodal ranked retrieval for integrated radiology data sources using context of search term by considering weighted ontology and category terms (Conference paper – under review).
6	Toward using FAIR Principles for Fine-Grained Access to aid Biomedical Data Driven Research [4].

For each phase we have identified a research question. Publications related to this work are briefly summarized in Table 1

3.2.1 Design an integrated smart database with heterogeneous data sources

Research question #1: How to determine which data sources and ontologies need to be integrated?

Most hospitals maintain a collection of teaching files, but many public teaching file collections are also available through curated online sources (e.g., RSNA MIRC, MyPacs, and EURORAD). We developed IRIS engine as a pilot for a data integration system for the healthcare domain [1]. In IRIS, we captured heterogeneous data from MIRC and MyPacs data sources, loading data into an integrated data repository. Using medical ontologies, we built our own dictionary which maps terms to their synonyms from the datasets and medical ontologies [3]. We designed an unsupervised machine learning technique that performs coverage analysis of data sources and medical ontologies to learn properties of the data (e.g., topic coverage). By learning data repositories contents, one can decide which data sources need to be integrated or what repository content is lacking. Thus, this coverage analysis algorithm benefits data integration process by extracting knowledge about the repositories (addressing research question #1). Our analysis also confirmed that data integration is a continuous, iterative process [2].

3.2.2 Ranked retrieval search engine with multimodal text and image-based search capabilities

Research question #2: How to find relevant documents given a keyword query or hybrid (text+image) query? Figure 1 shows the architecture of IRIS engine. When a user enters a text query, IRIS performs query expansion using relevant ontologies, and retrieves relevant results to the query term. Our database also stores accuracy feedback from users which is then used to evaluate and iteratively improve IRIS results.

An integrated search may result in thousands of matches; thus, we are designing a search algorithm that ranks results by incorpo-

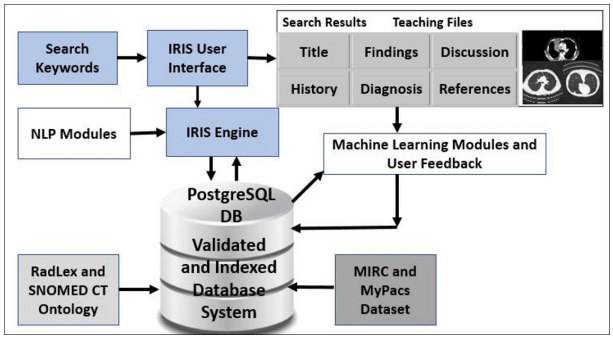

Figure 1: IRIS Architecture

rating context computed through a weighted ontology terms. For text-based search ranking evaluation we used Normalized Discounted Cumulative Gain (NDCG)[8] algorithm to measure the quality of search result ranking. Our analysis showed an improvement in ranked retrieval as compared to other search engines (addressing research question #2).

3.2.3 *Data bridges and indexing mechanism to integrate biomedical data sources*

Research question #3: How data integration performance (time) and scalability (adding variety of data sources) can be improved using data bridges? In order to make our integration solution applicable to other biomedical data sources (e.g., EHR's, clinical reports), we plan to create data adapters that will serve as a bridge between data providers and data integration systems (this work was a part of my internship at NIH). Data providers can share their data in any file format and bridges will interpret that data in a uniform manner. As shown in Figure 2, our data clustering indexing approach starts

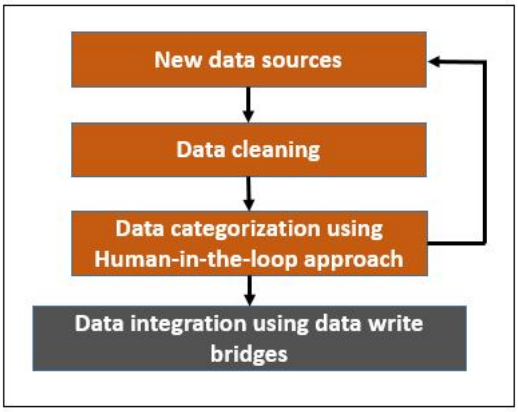

Figure 2: Data Categorization with human-in-the-loop

with collecting different biomedical data sources. From our literature survey we observed that data preparation accounts 80% of data scientist work. Data preparation includes finding relevant data sources, extracting data from those data sources, data cleaning, and data integration. Our proposed data integration system would help data scientists and researchers optimize and streamline data preparation. We collected different biomedical data sources and working on defining standard data cleaning technique that would be applicable to the most of the similar data sources that we proposed in this work. Our data categorization module categorizes data items into different sets based on the usage of those data elements in search operation. We need support from a human to check the accuracy of data categorization, to set similarity thresholds between different data items, and apply additional domain knowledge to categorize these data items based on relevance between data objects. Our data categorization algorithm will differentiate data items based on diagnostic relevance. For example, teaching cases with title, findings, and diagnosis would be treated as one sub-category in teaching cases (that would also integrate clinical reports) while another subcategory could integrate fields those are medically less relevant e.g., discussion, history, or comments. Based on data categorization we will be designing database schema and would also evaluate schema based on standard database schema benchmark techniques. Data write bridges would be responsible for the extracting data from different data categories and loading data to the respective database schema. This data categorization work is ongoing and we do not have any experimental results yet. We will address research question #3 by implementing this module.

[8] https://en.wikipedia.org/wiki/Discounted_cumulative_gain

4. EXPERIMENTAL RESULTS

In this section we briefly discuss the current results from proposed system.

4.0.1 *Text-based results*

We evaluated IRIS search ranking using a combination of queries received from radiologists at a well-known hospital and other queries chosen from an extensive literature survey. We have initially tested a total of 28 text queries, out of which we picked a subset of 10 queries (Q1:Cardiomegaly, Q2: ACL Tear, Q3: Annular Pancreas, Q4: Pseudocoxalgia, Q5: Varicocele, Q6: Angiosarcoma, Q7: Tracheal dilation, Q8: Appendicitis, Q9: Bronchus intermedius, Q10: Cystitis glandularis) to perform an in depth evaluation. Due to space constraints we briefly discuss text based results. We evaluated text-based results on a scale from 0 ("not relevant") to 2 ("very relevant"). We defined five categories to score text search results: "not relevant" = 0 (when term and synonyms do not appear anywhere in the results), "relevant" = 0.5 (if term or synonyms appear in any category of teaching file), "more relevant" = 1 (if term or synonyms appear in discussion category), "most relevant" = 1.5 (if term or synonyms appears in history or ddx category), and "very relevant" = 2 (if term or synonyms appears in title, findings, or diagnosis categories).

Comparison of IRIS and MIRC relevance rank algorithm using same datasets:
We compared IRIS relevance rank algorithm with MIRC using the same dataset. We considered top four teaching file results from IRIS, MIRC, and Google site search. We calculated relevance score by scoring top four teaching files from each engine, using weighted ontology ranking algorithm . Figure 3 shows an overall analysis of results from these 3 search engines. score for each search engine shows that IRIS relevance rank algorithm performs better than other two engines.

Ranking evaluation of other medical search engines:
We also considered how other public medical radiology teaching file search engines rank their search results. We used the same query set and performed a search using MIRC, MyPacs, EURORAD, and Open-i search engines. We discuss only two queries (Q1:"cardiomegaly" and Q8:"appedicitus") in detail and reporting scores for the top 10 search results. Figure 4 shows a comparative

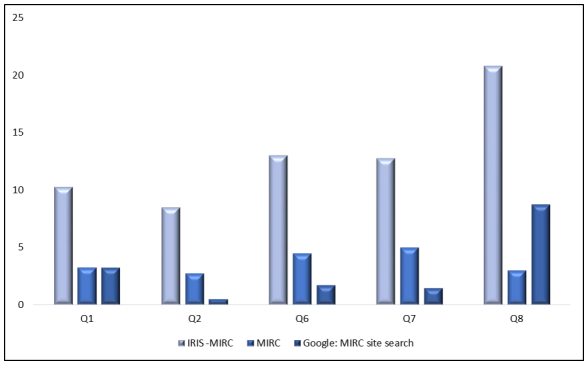

Figure 3: IRIS relevance rank results comparison with MIRC

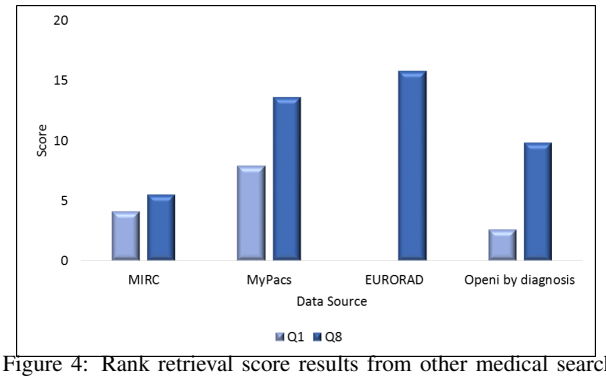

Figure 4: Rank retrieval score results from other medical search engines

analysis of ranked results from these four engines using the relevance scores based on our metric described above. Open-i can rank search results based on different categories (e.g., based on diagnosis or based on teaching file date) – we used a diagnosis based search in Open-i. MIRC ranks results based on the date of modification with no other option available. Our analysis shows that none of the search engines return the most relevant results first. Interestingly, top results are often less relevant than the subsequent search results. For example for "cardiomegaly" MyPacs fourth result is more relevant than the top three results. EURORAD does not retrieve any results for "cardiomegaly" but we checked "appendicits" results – and those were also not ranked based on the relevance of the search term.

4.0.2 Hybrid Text and Image based results

IRIS hybrid algorithm augments the text search with image search and re-ranks results based on the relevance to the query. Due to space constraints we briefly discuss hybrid search result. IRIS text-based and hybrid search results scored an score of 0.83 out of 1. Image search scored only about 0.53 out of 1, validating our use of the image search as an enhancement to the text search (rather than a standalone search). Hybrid search scored 0.84 out of 1 because of text results were augmented with image-based results. For hybrid search Some of the results were noticeable better than text-based search.By combining text search with image results, we are striving to get a text-based match that also includes a similar image.

5. CONCLUSIONS

The ranking approach presented in this paper is significant because it enables IRIS to present the user with top relevant reference cases first. Through integrating term frequency, adding more weight to ontology terms we show that teaching files can be better ranked in order of their relevance to a search query. Currently I am working on data write bridges and categorization algorithm to improve biomedical data integration process.

6. ACKNOWLEDGMENTS

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

7. REFERENCES

[1] P. Deshpande, A. Rasin, E. Brown, J. Furst, D. Raicu, S. Montner, and S. Armato III. An integrated database and smart search tool for medical knowledge extraction from radiology teaching files. In *Medical Informatics and Healthcare*, pages 10–18, 2017.

[2] P. Deshpande, A. Rasin, E. Brown, J. Furst, D. S. Raicu, S. M. Montner, and S. G. Armato. Big data integration case study for radiology data sources. In *2018 IEEE Life Sciences Conference (LSC)*, pages 195–198. IEEE, 2018.

[3] P. Deshpande, A. Rasin, E. T. Brown, J. Furst, S. M. Montner, S. G. Armato III, and D. S. Raicu. 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)*, 8(2):18–43, 2018.

[4] P. Deshpande, A. Rasin, J. Furst, D. Raicu, and S. Antani. Diis: A biomedical data access framework for aiding data driven research supporting fair principles. *Data*, 4(2):54, 2019.

[5] 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*, 4(1):65–68, 2007.

[6] J. R. Hemler, J. D. Hall, R. A. Cholan, B. F. Crabtree, L. J. Damschroder, L. I. Solberg, S. S. Ono, and D. J. Cohen. Practice facilitator strategies for addressing electronic health record data challenges for quality improvement: Evidencenow. *The Journal of the American Board of Family Medicine*, 31(3):398–409, 2018.

[7] A. Holzinger, M. Dehmer, and I. Jurisica. Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, 15(6):I1, 2014.

[8] G. Li. Human-in-the-loop data integration. *Proceedings of the VLDB Endowment*, 10(12):2006–2017, 2017.

[9] Z. J. Ling, Q. T. Tran, J. Fan, G. C. Koh, T. Nguyen, C. S. Tan, J. W. Yip, and M. Zhang. Gemini: an integrative healthcare analytics system. *Proceedings of the VLDB Endowment*, 7(13):1766–1771, 2014.

[10] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. DAgostino. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *BioMed research international*, 2014, 2014.

[11] M. S. Simpson, D. Demner-Fushman, S. K. Antani, and G. R. Thoma. Multimodal biomedical image indexing and retrieval using descriptive text and global feature mapping. *Information retrieval*, 17(3):229–264, 2014.

[12] R. Talanow. Radiology teacher: a free, internet-based radiology teaching file server. *JACR*, 6(12):871–875, 2009.