Biomedical Heterogeneous Data Integration and Rank Retrieval using Data Bridges

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 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 these challenges and developed an Integrated Radiology Image Search (IRIS) framework that could be a step toward aiding data-driven research. We propose building data bridges to support retrieving ranked relevant documents from integrated repository.

My research focuses on biomedical data integration and indexing systems that provides 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.

Research Questions: How to identify and integrate biomedical heterogeneous data sources? How to find diagnostically relevant documents from integrated repository?

Several studies have 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 image interpretation using case file databases. Talanow et al. [6] described a reference radiological system for diagnosis, teaching needs, research, and the resulting need for an advanced reference search engine. We applied unsupervised machine learning techniques that performs coverage analysis and of data sources and ontologies. 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. Our analysis showed that data integration is a continuous, iterative process [2]. To start with the integration of healthcare data sources, We developed IRIS as a pilot study for a healthcare data integration framework [1]. IRIS incorporated medical ontologies to augment search terms with synonyms and definitions [3]. An integrated search may result in thousands of matches; thus, we are designing a search algorithm that ranks results by incorporating context computed through a weighted ontology terms. For text-based search ranking evaluation we used Normalized Discounted Cumulative Gain (NDCG) [1] algorithm to measure the quality of search result ranking. Our analysis showed an improvement in ranked retrieval as compared to other search engines. To generalize our solution to heterogeneous biomedical data sources, we plan to create data adapters to serve as a bridge between data providers and the data integration and search framework [4].

CCS CONCEPTS

• **Data integration**; • **Biomedical information retrieval**; • **Metadata indexing**; • **Data bridges**;

KEYWORDS

Biomedical data integration, Metadata indexing, Information retrieval, Medical ontology, Query expansion

ACM Reference Format:
Priya Deshpande and Supervised by Dr. Alexander Rasin. 2019. Biomedical Heterogeneous Data Integration and Rank Retrieval using Data Bridges. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3331184.3331417

REFERENCES

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

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

[3] Deshpande, P., Rasin, A., Brown, E. T., Furst, J., Montner, S. M., Armato III, S. G., and Raicu, D. S. (2018b). 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.

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

[5] Gutmark, R., Halsted, M. J., Perry, L., and Gold, G. (2007). Use of computer databases to reduce radiograph reading errors. *Journal of the American College of Radiology*, 4(1):65–68.

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

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331417

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