Natural Language Processing and Parallel Computing for Information Retrieval from Electronic Health Records

Ali Abu Salimeh [1], *Najah* Al-shanableh[2*], and *Mazen* Alzyoud[2]

[1] Hashemite University, Jordan,
[2] Al al-Bayt University, Jordan,

Abstract. In this paper, we review the literature to find suitable information retrieval techniques for EHealth. Also discussed NLP techniques that have been proved their capability to extract valuable information in unstructured data from EHR. One of the best NLP techniques used for searching free text is LSI, due to its capability of finding semantic terms and in rich the search results by finding the hidden relations between terms. LSI uses a mathematical model called SVD, which is not scalable for large amounts of data due to its complexity and exhausts the memory, and a review for recent applications of LSI was discussed.

1 Introduction

Information retrieval involves searching and extracting metadata from images, texts, and audio files [1]. Within the medical field, information retrieval is involved with the extraction of information related to the patient's records. In the modern-day context, medical information is obtained from the web, social media, hospital records, and journal articles [2]. The amount of data found in Electronic Health Records (EHRs) gives rich information about diseases and historical medical information that could be used to understand better and treat many medical cases [3]. Searching inside these records is a challenge to traditional information retrieval techniques due to the non-standard structure and free text in these records [4].

Many Natural language processing (NLP) techniques have been used to extract valuable information in unstructured data from EHR. One of the best NLP techniques used for searching free text is Latent Semantic Indexing (LSI) due to its ability to find semantic terms and rich the search results by finding the hidden relations between words [5]. Parallel computing is a type of computation where many calculations the execution of processes are carried out simultaneously in a single machine that has multiple processors execute multiple tasks simultaneously and have shared memory communicate with each other using a bus [7]. That will increase the system's accuracy because we do not lose any part of data relations. For Natural language processing from EHR, where massive computations happen, parallel computing can increase the accuracy while minimizing processing time [11].

* Corresponding author: najah2746@aabu.edu.jo

1.1 Electronic Health Records (EHR)

EHR contains information about a patient's medical history, treatment plans, medications, treatment dates, allergies, and laboratory and test results, among others [3]. The digital format of an EHR makes it possible to store and share the patient's information using networked information systems.
There are many challenges to understanding and repurposing EHR information for clinicians and researchers. Since a great part of the substance in the clinical record is in unstructured (free text) clinical documentation and it can collect any type of text character, utilization of natural language processing is frequently required. Regardless of these difficulties, scientists have been fruitful in duplicating known hereditary affiliations and making new revelations utilizing EHR information [4]. So Clinical concepts must be normalized if decision support and analytic applications are to operate reliably on heterogeneous EHR data.

1.1.1 Latent Semantic Indexing (LSI)

Latent semantic indexing is a technique used in information retrieval that helps in the analysis of documents to reveal the hidden relationships (latent) between the words or semantics to improve the understanding of the retrieved information [5]. LSI utilizes a technique known as singular value decomposition (SVD) to scan the unstructured data to understand the relationships between the concepts and terms included in the documents.

1.1.2 Apache Spark for EHR

Apache Spark is an open-source analytics engine that is used in machine learning and big data analytics. The analytics engine is very fast, which makes it suitable in parallel processing frameworks and large-scale big data analytics platforms. Spark is implemented using two main instances, notably a cluster management interface and a distributed storage system [6]. The ability to support data processing from multiple repositories makes Apache Spark an ideal solution for the development of information retrieval systems. Spark has many properties as Speed, Ease of Use, Generality, and Runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud [7].

2 Information Retrieval and Electronic Health Records

Yousefu managed to design and develop a model that aimed at assisting the extraction and retrieval of EHR using scenario-based information. The model works by first creating a framework that defines the relationships between diseases, symptoms, and other related clinical information [8]. Popescu proposed a model that utilizes general word sequencing, which is based on the use of a dynamic algorithm and implemented using the fuzzy version of the Smith-Waterman model. The model uses the domain ontology to calculate and compute the word similarity matrix [9]. Traina and Rosa proposed a system that is used to retrieve the image-based content of electronic health records. The Similarity Retrieval of Images System (SRIS) is designed in such a way that it can perform a similarity query over images containing the patient's data [10].

2.1 Information Retrieval, Electronic Health Records and Natural Language Processing

Jain aimed at developing a framework designed to improve the paper-based information retrieval technique for accessing EHR. The proposed methodology utilizes a technique known as semantic query expansion and is aimed at supporting the retrieval of information related to the patient's current health condition [11].
MEDREADFAST is a hybrid browser with search capabilities that was proposed and described by Gubanov. The browser is supposed to simplify the process of accessing and retrieving medical information from big clinical data. MEDREADFAST was developed using a combination of several information retrieval techniques, including natural language processing, data management, and machine learning [12]. Pruski developed a framework that helps to address issues of encryption in the retrieval of EHR. The approach aimed at overcoming the barriers of encryption by combining the use of knowledge organizing systems and standard XDS metadata [13].
Yadav proposed the development of a personalized search engine that helps to refine the process of obtaining an internet-based patient's medical information [14]. Zhu managed to describe the development of a model aimed at improving information retrieval of EHR using discharge models. The framework involved the combination of various models that utilized natural language processing to assess the suitability of the information retrieval system [15].

2.2 Information Retrieval, Electronic Health Records, Natural Language Processing and Parallel Computing

Al-Qahtani proposed a novel approach that aimed at improving the process of information retrieval from e-health systems. The model uses Latent semantic indexing (LSI) algorithm and is implemented within The Health Improvement Network dataset (THIN). The main aim of the model is to improve the accuracy of the information retrieval process whereby the Term Document Matrix is utilized [16]. Hammad developed a model for the sole purpose of enhancing the storage and retrieval of XML-based electronic health records. The framework is aimed at improving the querying and indexing operations in an environment characterized by cluster computing [17]. Ufuk Parali also proposed a new algorithm reduced row echelon form IR method (rrefIR) with higher average similarity precision to Get more relevant and noise-free documents by applying (SVD) on the reduced row echelon form obtained by utilizing Gauss-Jordan method of the covariance of (TDM) [18].

2.3 Information Retrieval, Electronic Health Records, and Apache Spark

Md. Rezaul Karim and others propose a healthcare analytics framework with Apache Spark for real-time healthcare monitoring and related analytics by integrating heterogeneous data sources like the Internet of Things (IoT) enabled medical devices, Electronics Health Records (EHRs), picture archiving and communication system (PACS), wearable sensors, and public open health datasets. The ultimate goal of this research is to provide an effective, scalable, and secure framework for increasingly important public health surveillance [19]. One of the best NLP techniques used in previous research for searching free text is LSI due to its ability to find semantic terms and rich search results by finding the hidden relations between terms. LSI uses a mathematical model called SVD, which is not scalable for large amounts of data due to its complexity and memory exhaustion [18].
Many researchers address this problem and proposed solutions that are based on distributed computing witch it divides one task between several machines to achieve a common goal. Thus, memory systems are divided among the processors. Each machine can communicate

with others via the network, but this way affects the accuracy of the system because separating the data will lose part of its relations. Still, this way affects the system's accuracy because separating the data will lose part of its relations. An enhanced parallel information retrieval methodology for the LSI technique to analyse and processes unstructured data from EHR datasets in less time and with more accuracy is needed. Apache Spark, which applies parallel computing, can solve the LSI problem because it uses a very fast analytics engine that makes it suitable in parallel processing frameworks and large-scale big data analytics platforms.

3 Conclusion

In this research, we reviewed the literature on NLP and Parallel techniques that proved their capability to extract valuable information in unstructured data from EHR. Through previous studies, it appears that various techniques can be used to obtain information related to search operations within the free text. One of the best NLP techniques used for searching free text is LSI due to its capability of finding semantic terms and in rich the search results by finding the hidden relations between terms. Table 1 provide a summary of the related works.

Table 1. Summary of Related Works.

N	Reference	Proposed Framework or Algorithm	Goal	Testing Dataset	Major Results
1.	(Jain, Thao and Zhao, 2012)	Semantic Query Expansion (SQE)	Enhance EMR retrieval through SQE	Sample dataset with nursing notes from the community health center in a Midwestern metropolitan area	The proposed framework can improve the retrieval of EHR in a community health center.
2.	(Gubanov and Pyayt,2012)	MEDREADFAST a hybrid browser with search capabilities for big clinical data	Return the links of the most relevant webpages sorted by decreasing order.	10k natural language sentences composed of patient records from different clinical domains.	The results indicate that the hybrid browser yields more relevant results than keyword search while at the same time improving the user's experience.
3.	(Pruski and Wisniewski, 2012)	The framework utilizes two main approaches; the first one uses XDS metadata to describe the contents of CDA documents while the second approach involves the exploitation of XDS metadata content.	Address the issues of encryption in the retrieval of EHR.	Documents prepared using Clinical Data Architecture	The results indicate that the proposed framework improves the precision and recall of information retrieval and search processes.

4.	(Yadav and Poellabauer, 2012)	A new ranking algorithm combines both user query and health profile	Refine the process of obtaining an internet-based patient's medical information.	Personal Health Record system developed in-house	The search engine uses crawling and ranking algorithms to eliminate irrelevant results and ranks the top priority results respectively.
5.	(Traina, Rosa, and Traina, 2013)	Similarity Retrieval of Images System (SRIS) which uses two types of similarity: range queries and k-nearest neighbor.	access the patient's image data and then perform a similarity search to obtain similar results related to6the patient's data.	Image data sets obtained from a clinical hospital at the University of Sao Paulo at Ribeirao Preto-Brazil	The results indicating a high degree of accuracy in terms of the similarity results.
6.	Md. Rezaul Karim, Ratnesh Sahay, and Dietrich Rebholz-Schuhmann. (2015)	propose a healthcare analytics framework with Apache Spark for real-time healthcare monitoring and related analytics by integrating heterogeneous data sources	provide an effective, scalable, and secure framework for increasingly important public health surveillance	public open health datasets	SparkML is used to handle big static datasets coming from static data sources, Proposed Knowledge bases with Ontology Web Language (OWL2) and Open Biomedical Ontology (OBO) are used to generate rules, ontologies, and semantic annotation.
7.	(Al-Qahtani, Amira, and Ramzan, 2015)	Distributed Latent semantic indexing	Improve the accuracy of the information retrieval process whereby the Term Document Matrix is utilized.	THIN dataset	The proposed system was found to be highly effective, especially because it helped to eliminate invaluable information contained in the free text.
8.	(Reis, Bonacin, and Perciani, 2016)	Intention-based information retrieval for EHR	Investigate the means of considering intention in search engines	Dataset sample from Public hospital of "Aguas de Lindóia", São Paulo State, Brazil	The simulation results indicate that the intention-based technique is successful in a real-world medical environment.

9.	(Hammad and Banikhalaf, 2018)	Parallel Approach for Managing XML-based Electronic Medical Records	Improve the querying and indexing operations in an environment characterized by cluster computing.	Not mentioned	The results indicate that XML-based querying and indexing greatly improves the information storage and retrieval process in a distributed computing environment.
10	(Ufuk Parali, Metin Zontul, and Duygu Celik Ertugrul, 2019)	propose a new algorithm reduced row echelon form IR method (rrefIR) with higher average similarity precision	Get more relevant and noise-free documents by applying (SVD) on the reduced row echelon form obtained by utilizing the Gauss-Jordan method of the covariance of (TDM).	Not mentioned	The linear independence of basis vectors provided by the Gauss-Jordan operation makes the rrefIR algorithm retrieve more noise-free and relevant documents than LSI and COV algorithms.

References

1. C. Manningm, P. Raghavan, and H. Schütze, *Book Review: Introduction to Information Retrieval,* Natural Language Engineering, vol. **16**, no. 1, pp.100-103 (2010).
2. C. Smith, *Information retrieval in medicine: The electronic medical record as a new domain,* Proceedings of the American society for information science and technology, vol. **43**, no. 1, pp. 1-30 (2006).
3. K. Häyrinen, K. Saranto, and P. Nykänen, *Definition, structure, content, use and impacts of electronic health records: a review of the research literature,* International journal of medical informatics, vol. **77**, no. 5, pp. 291-304 (2008).
4. J. Denny, H. Xu, *Chapter 12-linking genomic and clinical data for discovery and personalized care,* methods in Biomedical Informatics, Sarkar IN, ed. Oxford: Academic Press, pp. 395-424 (2014).
5. S. Dumais, *Latent semantic analysis*, Annual review of information science and technology, vol. **38**, no. 1, pp.188-230 (2004).
6. J. Shanahan and L. Dai, *Large scale distributed data science using apache spark*, In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 2323-2324 (2015).
7. A. Khumoyun, Y. Cui, and L. Hanku, *Spark based distributed deep learning framework for big data applications*, In 2016 International Conference on Information Science and Communications Technologies (ICISCT), IEEE, pp. 1-5, (2016).

8. A. Yousefi, N. Mastouri, and K. Sartipi, *Scenario-oriented information extraction from electronic health records*, In 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, IEEE, pp. 1-5 (2009).
9. M. Popescu, *An ontological fuzzy Smith-Waterman with applications to patient retrieval in Electronic Medical Records*, In International Conference on Fuzzy Systems, IEEE, pp. 1-6 (2010).
10. A. Traina, N. Rosa, and C. Traina, *Integrating images to patient electronic medical records through content-based retrieval techniques*, In 16th IEEE Symposium Computer-Based Medical Systems, Proceedings., IEEE, pp. 163-168 (2003).
11. H. Jain, C. Thao, and H. Zhao, *Enhancing electronic medical record retrieval through semantic query expansion*, Information systems and e-business management, vol.**10**, no. 2, pp. 165-181 (2012).
12. M. Gubanov and A. Pyayt, *MEDREADFAST: A structural information retrieval engine for big clinical text*, In 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), IEEE, pp. 371-376. (2012).
13. C. Pruski and F.Wisniewski, *fficient medical information retrieval in encrypted electronic health records,* In Quality of Life through Quality of Information, IOS Press, pp. 225-229 (2012).
14. N. Yadav, and C. Poellabauer, *An architecture for personalized health information retrieval*, In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 41-48 (2012).
15. D. Zhu, S.Wu, J. Masanz, B. Carterette, and H. Liu, *Using Discharge Summaries to Improve Information Retrieval in Clinical Domain*, In CLEF (Working Notes), (2013).
16. M. Al-Qahtani, A. Amira, and N. Ramzan, *An efficient information retrieval technique for e-health systems*, In 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, pp. 257-260 (2015).
17. R. Hammad and M. Banikhalaf, *A parallel approach for managing XML-based electronic medical records*, In 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1-5 (2018).
18. U. Parali, M. Zontul, and D. Ertugrul, *Information Retrieval Using the Reduced Row Echelon Form of a Term-Document Matrix*, Journal of Internet Technology, **vol. 20**, no. 4, pp. 1037-1046 (2019).
19. M. Karim, R. Sahay, and D. Rebholz-Schuhmann, *A scalable, secure and realtime healthcare analytics framework with Apache Spark*." In Proc. of the 2nd INSIGHT student conference on Data Analytics, The Insight Centre for Data Analytics, pp. 83-83 (2015).
20. M. Al-Qahtani, A. Amira, and N. Ramzan, *Enhancing the efficiency of information retrieval in e-health systems*, In 2015 British Computer Society Health Informatics Scotland Conference, (2015).