Framework of a Health Care Information Retrieval System

Pandaram Sathish Kumar, Shashi Mehrotra*

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

**sethshashi11@gmail.com*

Abstract

In these years, digital medical data has increased exponentially, and web search is becoming very common. Information Retrieval (IR) provides users with the needed information. However, it is tedious retrieving relevant information for several reasons. One important reason is more than one meaning of the word. The medical domain is sensitive, and appropriate and timely information retrieval is necessary. To enhance medical information retrieval, we present an information retrieval framework for retrieving articles and medical information according to the query given by the user. We used a vector space model. The health care information retrieval system would be helpful for doctors, patients, and researchers to get information related to their queries. The retrieved related medical records could help in diagnosing and treatment of associated diseases. We used data set TREC-15, for experimental evaluation of the model and obtained promising experimental results.

Keywords- Machine Learning, Information retrieval, Document Term Matrix

1. INTRODUCTION

Medical information searching is becoming very common in present days of digital era and the availability of huge digital data. Health care retrieval systems are very help full for various types of users, such as patients, doctors, and researchers, to access medical information [19]. An information system is understood as a method of searching of information from the web. Users expect relevant retrieval of contents for the query given by them. The query given is matched with the metadata that is about the details in the contents collection. The user needs some sort of information, and the need is formed as a query. Matched documents with the query are returned as a search result. The information is now available on many types of platforms like websites, social media platforms like Twitter, Github, and records of hospital. Searching for health content is very popular these days, and it is the foremost important domain in information retrieval [22]. Physicians want evidence related to the best curing methods for their patients' diseases.

These evidences are available in articles, historical books, and reports which are earlier faced by some expert physicians, researchers. The analysis presented in [22] shows that most commonly search is done for the following information, specific disease related information, treatment, doctors, hospitals, medical insurance, food, drug safety, health or environmental threat [22].

Information retrieval for medical is challenging as medical sciences is very vast and diverse [18, 22]. Publications and organizations are some examples of information resources [22]. The major challenge of IR is the heterogeneity of the information resources. Search engines display retrieved information in the list form. It is very tedious searching for information relevant to the user from the long list [3,4]. Information retrieved and stored in the meaningful cluster would be helpful for efficient searching, where clustering could be used [5,6,7,8]. In [9], the author explored the evolutionary approach of clustering for searching [9].

The remainder of the paper is organised as follows: Section 2 is literature survey discusses related research, and section 3, presents the methodology of our study. Section 4 presents experimental results and analysis, and section 5 is the conclusion and future work.

3. DESIGN AND METHODOLOGY

Medical information retrieval is a sensitive and challenging task. This section discusses some related research. Abadeh MS [1] discussed approaches to dilute fuzzy systems with learning methods made in soft computing. Neural and fuzzy systems will dilute the approximate reasoning method of fuzzy systems with the capabilities of learning neural networks and algorithms. He also said that genetics-based learning algorithm and discussed its usage to detect intrusion in a computer network. Aravind et al.[2], He presented that they will describe a magnificent approach to fetch related medical articles from PubMed collection, based on a required given query. So his Information Retrieval system consists of 3 parts: inverted indexing using Lucene, lexical query expansion to increase recall with Meta Map, and reranking aimed at optimizing the system. He evaluated his system using 30 medical queries. Yang [10], disc ered in his study that the present search engines in web often cannot handle the medical search efficiently because of not considering the special requirements of search. An uncertain procedure for a medical information searcher about questions and is unfamiliar with medical terminology. So, the author sometimes prefers to pose long queries, saying symptoms and the situation in English to receive information relevant from the search results. Eysenbach Gunther [11] identified that e-Health is an essential field in bisecting the medical informatics, business, referring to health services and information enhanced by searching on the Internet or related technologies. More broadly, it characterizes both technical development and state of mind and attributes like attitude, commitment, global thinking, to improve health care logically, and communication technology. Estrela [12] discussed the importance of Medical Image Processing in e-health and telemedicine. This will enable rapid diagnosis with visual, quantitative, and analytical assessment. Remote care could reveal some important changes that indicate a therapy progression. For example, Covid disease evaluation uses behavioural tests, PET scans and MRI of the whole brain. The collection of diverse images offers some chances to improve diagnosis the evidence-based. So there is some need for proper methods to search some of these collections for images with similar images. Wang et al. [13], stated that medical information retrieval set a goal to discover the scientific evidence to support decision making with some knowledge in the medical domain. Knowledge developed by domain experts is able to enhance the understanding of the free text with some knowledge in the domain. Structured knowledge bases represent the information as a group of knowledge graphs consisting of nodes and edges. This representation has a long history in logic and artificial intelligence. Strictly, UMLS is used chiefly knowledge base in the medical domain. UMLS is a classroom of biomedical terms and concepts. Kasban et al. [14], presented a method to retrieve the medical images for searching the required image for the query image in the database. Nishant et al. [16], experimented with Random Forest. Bayesian Classification, Decision Tree, and SVM for medical Data. The SVM algorithm demonstrate highest precision and F-measure. Nishant et al. [17] presented an experiment with sampling methods to improve classifiers' performance. Most of the data in the medical is a class imbalance. Therefore, it is required to apply some sampling methods before any classification to improve the performance. Di Girolamo, Nicola [20], used and evaluated the Bayesian algorithm for information retrieval and clinical decision support.

4. Design And Methodology

This section present methodology we used in our study, and the data details used for the experiments.

3.1 Vector Space Model:

It is a very popular framework for applying term weighted. Term frequency inverse (TF-IDF) is used to weight the term, which shows the term's importance [25].

Following formulas for representing the documents as a vector space model. In this model, each document is considered as a vector in the term-space.

Inverse document Frequency = 1/(log(DF)+1) (1)

TF-IDF = TF*(1/(log(DF)+1)) (2)

TF – (count of a required word in query)/ (Total no of words in the document)

DF – (No of documents that contain the required word)/(Total number of documents present in the corpus)

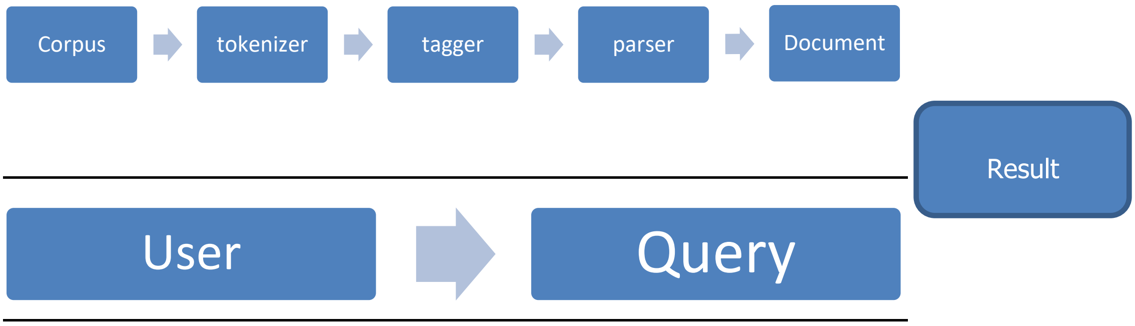


Figure 1. Workflow of our process

3.2 Data Description

We used the TREC-15 dataset for this project. TREC-15 consists of 128491 rows and four columns.

Dataset consists of columns like topic-id, query, question, and narrative.

Topic id refers to the id of the topics. Query refers to the type of query that the user is giving. The second dataset consists of country, latest, long, and dates columns. The country says the country in which the given covid cases are registered. The latest gives the number of cases up to then those are registered. Long gives the largest number of cases registered in a period. Dates represent the covid cases recorded on the specific date.

3.3 Relevance based Evaluation:

The main goal of the IR system is to retrieve relevant information/documents. We considered the relevance and diversity of the search for the evaluation. To measure how the models performed in this regard, we use precision, recall, and f-measure matrices. Information retrieved could be true positive (TP), false positive (FP), true negative (TN), or false negative (FN). TP + FP are the total numbers of documents retrieved, while TP+FN are the relevant documents [23].

Precision is a fraction of the relevant documents retrieved.

Precision = (TP) / (TP + FP) (3)

Recall is fraction of the relevant documents retrieved from the data.

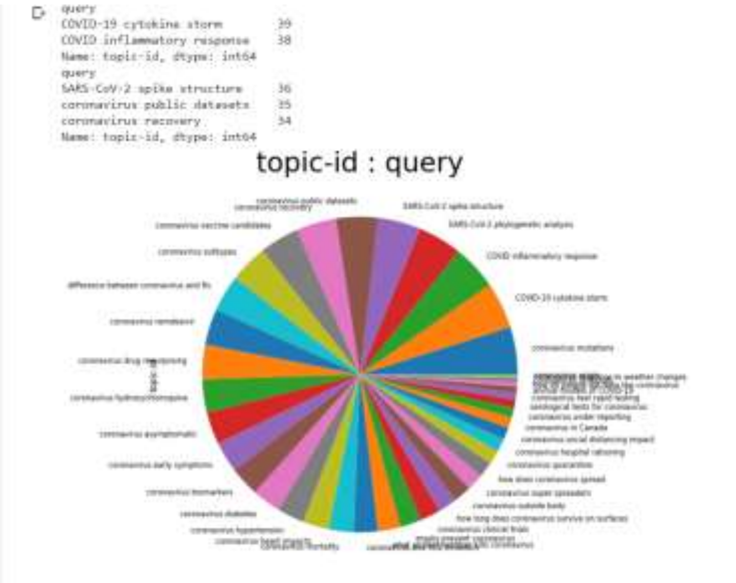
Recall = (TP) / (TP + FN) (4)

F-measure is harmonic mean of precision and recall.

F-measure = 2*(precision*recal)/(precision + recall) (5)

5. EXPERIMENT AND RESULT ANALYSIS

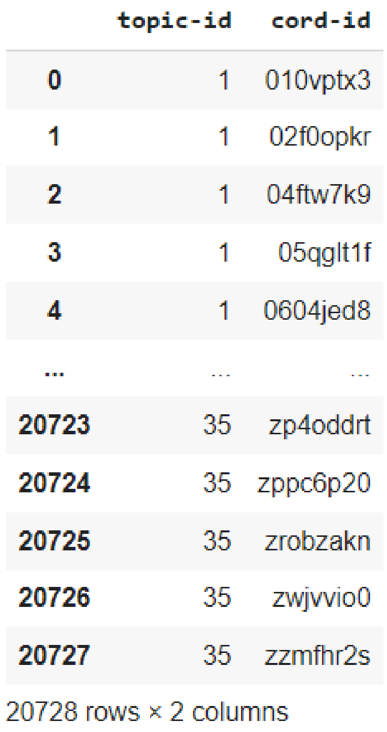
The Figure 2, presents pie chart with various categories of queries in the data.



```
query
COVID-19 cytokine storm        39
COVID inflammatory response    38
Name: topic-id, dtype: int64
query
SARS-CoV-2 spike structure     36
coronavirus public datasets    35
coronavirus recovery           34
Name: topic-id, dtype: int64
```

Figure 2. Number of occurrences queries

We can observe in the Figure 2 queries and their count, and pie chart with classification of queries.



	topic-id	cord-id
0	1	010vptx3
1	1	02f0opkr
2	1	04ftw7k9
3	1	05qglt1f
4	1	0604jed8
...	...	...
20723	35	zp4oddrt
20724	35	zppc6p20
20725	35	zrobzakn
20726	35	zwjvvio0
20727	35	zzmfhr2s

20728 rows × 2 columns

Figure 3. Documents retrieved

6. Conclusion And Future Work

The paper presents a health care information retrieval framework. Searching for various information is becoming very common in our day-to-day life due to the massive usage of internet and the availability of digital data. Various health related people such as doctors, patients, health insurance seekers, researchers etc. could be benefited from a lot of available information on web.

References

[1] Abadeh, M. S., Habibi, J., & Lucas, C. (2007). Intrusion detection using a fuzzy genetics-based learning algorithm. *Journal of Network and Computer Applications*, *30*(1), 414-428.

[2] Aravind, M., Viswanath, S., Mohan, N., Adarsh, R., & Bhaskar, J. (2019, December). A modified medical information retrieval system. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)* (pp. 218-222). IEEE.

[3] Mehrotra, S., & Sharan, A. (2020). Comparative Analysis of K-Means Algorithm and Particle Swarm Optimization for Search Result Clustering. In *Smart Trends in Computing and Communications* (pp. 109-114). Springer, Singapore.

[4] Nishant, P. S., Mehrotra, S., Sree, P. R., & Srikanth, P. (2020, August). Hierarchical clustering based intelligent information retrieval approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 862-866). IEEE.

[5] Mehrotra, S., & Kohli, S. (2016). Application of clustering for improving search result of a website. In *Information Systems Design and Intelligent Applications* (pp. 349-356). Springer, New Delhi.

[6] Mehrotra, S., & Kohli, S. (2015, October). Comparative analysis of K-Means with other clustering algorithms to improve search result. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 309-313). IEEE.

[7] Mehrotra, S., Kohli, S., & Sharan, A. (2018). To identify the usage of clustering techniques for improving search result of a website. *International Journal of Data Mining, Modelling and Management*, *10*(3), 229-249.

[8] Mehrotra, S., Kohli, S., & Sharan, A. (2019). An intelligent clustering approach for improving search result of a website. *International Journal of Advanced Intelligence Paradigms*, *12*(3-4), 295-304.

[9] Mehrotra, S., & Kohli, S. (2016, March). Identifying evolutionary approach for search result clustering. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 3778-3782). IEEE.

[10] Change to [2.2]Yang, L., Mei, Q., Zheng, K., & Hanauer, D. A. (2011). Query log analysis of an electronic health record search engine. In *AMIA annual symposium proceedings* (Vol. 2011, p. 915). American Medical Informatics Association.

[11] Eysenbach, G. (2001). What is e-health?. *Journal of medical Internet research*, *3*(2), e20.

[12] Estrela, V. V., & Herrmann, A. E. (2016). Content-based image retrieval (CBIR) in remote clinical diagnosis and healthcare. In *Encyclopedia of E-Health and Telemedicine* (pp. 495-520). IGI Global.

[13] Wang, H., Zhang, Q., & Yuan, J. (2017). Semantically enhanced medical information retrieval system: a tensor factorization based approach. *Ieee Access*, *5*, 7584-7593.

[14] Kasban, H., & Salama, D. H. (2019). A robust medical image retrieval system based on wavelet optimization and adaptive block truncation coding. *Multimedia Tools and Applications*, *78*(24), 35211-35236.

[15] Mourão, A., Martins, F., & Magalhaes, J. (2015). Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, *39*, 35-45.

[16] Nishant, P. S., Mehrotra, S., Mohan, B. G. K., & Devaraju, G. (2020). Identifying Classification Technique for Medical Diagnosis. In *ICT Analysis and Applications* (pp. 95-104). Springer, Singapore.

[17] Nishant, P. S., Rohit, B., Chandra, B. S., & Mehrotra, S. (2021). HOUSEN: Hybrid Over–Undersampling and Ensemble Approach for Imbalance Classification. In *Inventive Systems and Control* (pp. 93-108). Springer, Singapore.

[18] Sheikhshoaei, Fatemeh, Gholamreza Roshandel, Marzieh Zarinbal, MolukoSadat Hosseini Beheshti, Mohammadhiwa Abdekhoda, Auwal Abdullahi Abubakar, and Masoud Mohammadi. "Mapping Global Knowledge Domain, Research in Information Retrieval in Medical Sciences: A Scientometric and Evaluative Study." (2021).

[19] Di Girolamo, Nicola. "Advances in Retrieval and Dissemination of Medical Information." *Veterinary Clinics: Exotic Animal Practice* 22, no. 3 (2019): 539-548.

[20] Balaneshinkordan, Saeid, and Alexander Kotov. "Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine." *Journal of Biomedical Informatics* 98 (2019): 103238.

[21] Gudivada, Akhil, and Nasseh Tabrizi. "A literature review on machine learning based medical information retrieval systems." In *2018 IEEE symposium series on computational intelligence (SSCI)*, pp. 250-257. IEEE, 2018.

[22] Hersh, William, Hersh, and Weston. *Information retrieval: A biomedical and health perspective*. New York: Springer, 2020.

[23] Hersh, William. "Evaluation of biomedical text-mining systems: lessons learned from information retrieval." *Briefings in bioinformatics* 6, no. 4 (2005): 344-356.

[24] Xu, Bo, Hongfei Lin, Yuan Lin, Yunlong Ma, Liang Yang, Jian Wang, and Zhihao Yang. "Improve biomedical information retrieval using modified learning to rank methods." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15, no. 6 (2016): 1797-1809.

[25] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. *Modern information retrieval*. Vol. 463. New York: ACM press, 1999.

Biographies

Pandaram Sathish Kumar received the bachelor's degree in computer science and engineering from K L University in 2020, pursing the master's degree in computer science and engineering from KL University, respectively. He contributed research papers in international conferences. He participated workshops/seminars related to machine learning. His area of interest is Natural Language Processing, Artificial Intelligence, Machine Learning, and Deep Learning.

Shashi Mehrotra completed her Ph. D from Birla Institute of Technology, Mesra, Ranchi, in the area of Information retrieval. M.Tech in Computer Engineering from ITM Gurgaon, MCA, and M.Phil from Madurai Kamaraj University. She is actively engaged in teaching and research in areas of Computer Science. Currently working as an Associate Professor in the Department of Computer Science and Engineering at KL University, Vaddeswaram, India, since 1/12/2017. She contributed research papers in national/international journals/conferences/symposiums of repute, participated in keynote speech delivery, and invited talk. She has organized many professional activities like special sessions in conferences/FDPs, workshops, and expert lectures as an editorial board member or as a reviewer for several international journals and conferences. Her research interest includes Text mining, Machine learning, and Deep learning.