Digital healthcare

ELASTICSEARCH ENGINE IN MEDICINE: A TECHNICAL SURVEY

The article is recommended for publication by the Editor-in-chief Tatiana Ershova 02.12.2024.

Prutzkow Alexander

Doctor of engineering, associate professor
Ryazan State Radio Engineering University, Computational and applied mathematics department, professor
Lipetsk State Pedagogical University, computer science, information technologies, and information security department, professor
Ryazan, Russian Federation
mail@prutzkow.com

Abstract

Informatization has affected all spheres of the national life, including medicine. One of the areas of medicine informatization is the introduction of information systems. Elasticsearch (ES) is an search engine. It is used to store and search data in information systems. We analyzed 34 scientific articles on the use cases of ES in medicine, published in 2022-2024. Articles was filtered by type, description of use of ES, the English language, and free access. Our findings are following. ES is used primarily for searching and logging. In 41% of articles ES is used as a database management system (DBMS) that stores only part of the data, in 34% – as a primary DBMS, in 25% – as a secondary DBMS. On average, researchers store almost 2 million documents in a ES index. Redis, Apache Nifi, CogStack, Apache Kafka, Neo4j, PostgreSQL, Python are most often used with ES. We have collected some technical details of using ES (analyzers, data types, query types). ES is used with medical databases and tools such as UMLS, MedCAT, MeSH, SNOMED-CT.

Keywords

Elasticsearch, Elastic Stack, medicine, informatization, search engines, information retrieval, UMLS, MedCAT, MeSH, SNOMED-CT, survey

Introduction

Informatization, Medicine, and Search Engines

The main trend of the last few decades is informatization. Informatization is the process through which the new communication technologies are used as a means for furthering socioeconomic development as a nation becomes more and more an information society [1]. Informatization is a kind of automation. This process has affected all sectors of the national life, including medicine. One of the manifestations of informatization is the introduction of information systems. An almost complete description of information systems in medicine is collected in the encyclopedia [2]. Informatization involves increasing data volumes. Search engines are used for data searching. Search engines in medicine are studied in [3–6]. One of the search engines used in medical informatization is Elasticsearch (ES).

Actually the “informatization” term occurred mainly in articles of authors from xUSSR and China.

Elasticsearch Briefly

ES is a distributed system for storing documents, searching and analyzing data in them [7], namely:

- a document is a collection of fields of various types (text, numeric, compound);
- data can be found by a fragment of a field or by the entire field with conditions (contains/does not contain, logical operations AND, OR, NOT);
- data can be analyzed statistically, grouped, post-processed.

https://doi.org/10.52605/16059921_2025_04_96

ES is part of the Elastic Stack. Elastic Stack includes Logstash, a program for loading data into ES, and Kibana, a program for visualizing data and running queries against ES, as well.

Motivation

We are exploring ES to improve a course on information retrieval and search engines. We've published a textbook [7], an article on choosing books for initial learning of the Elastic Stack [8] and on information retrieval thinking [9], proceeding papers [10–11].

Related Papers

Surveys of the use of ES in medicine are brief. Approaches to extracting data from medical documents are reviewed in [12]. Here you will find some cases of application of ES in medicine. In [13] outlined the prospects of applying graph databases in systems biology with ES. It's a key point of this article.

1 The Purpose of the Study

The purpose of the study is to analyze cases of using ES in medical applications, discovering joint technologies and technical details.

2 Materials and Methods

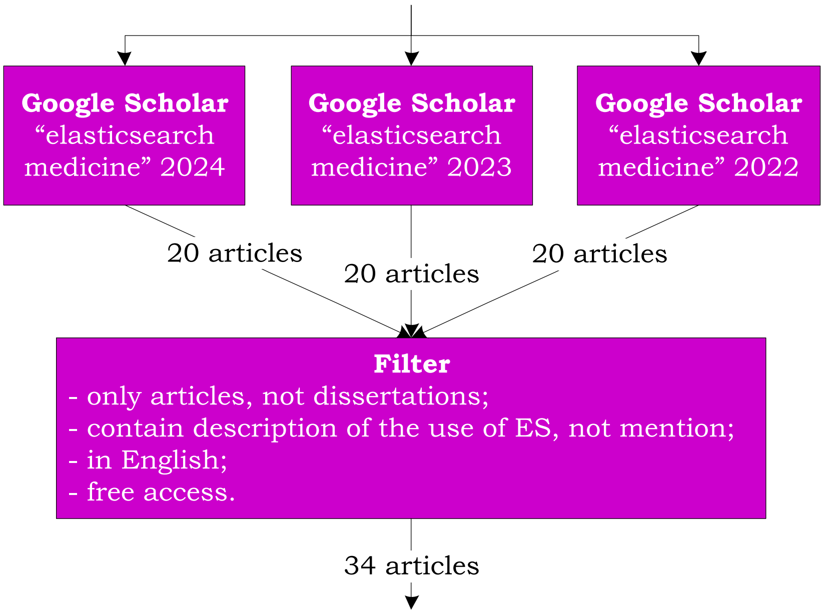
We searched for articles in Google Scholar. For each year of publication from 2022 to 2024, we selected the first 20 articles at the search query "elasticsearch medicine" (see fig. 1). Articles were filtered by the following criteria:

- only articles, not dissertations;
- contain description of the use of ES, not mention;
- in English;
- free access.

It's interesting 3 articles [14–16] in our survey are from one issue of the Nucleic Acids Research journal.

Articles for 2023-2024 were downloaded on March 25, 2024, for 2022 – April 01, 2024.

Methods are data aggregations, statistical and technical analysis.



Google Scholar "elasticsearch medicine" 2024

Google Scholar "elasticsearch medicine" 2023

Google Scholar "elasticsearch medicine" 2022

Filter
- only articles, not dissertations;
- contain description of the use of ES, not mention;
- in English;
- free access.

Figure 1. Article retrieval pipeline

3 Results

3.1 Result Organization

We revealed use cases of ES in medicine. We've grouped use cases to searching, logging, and other use cases. In each group, the articles are sorted by the name of the first author.

3.2 Searching

Alskaf E. et al. [17] studied machine learning models to predict all-cause mortality. The source data are the Electronic Health Records (EHRs) data and Stress Perfusion Cardiac Magnetic Resonance (SP-CMR) imaging. EHRs stored in ES. The best model is the support vector machine (SVM).

Cao C. et al. [14] created the PGS-Depot web site [18]. It provides comprehensive polygenic scores (PGSs), computed summary statistics. PGS-Depot utilizes ES to enable searching by summary statistics ID in PGS-Depot, trait name, and the PubMed Identifier (PMID).

Dilmaç F. and Alpkoçak A. [19] noted the performance of ES decreases when data preprocessing. ES is used for searching medical research articles in Automated Keyword Assignment System (AKAS).

Debauche O. et al. [20] designed an architecture of a system for elderly patient monitoring. Data produce by medical sensors and devices. One of technologies in this system is ES. ES is utilized as a secondary database management system (DBMS) for searching data.

Guo Q., Cao S., and Yi Z. [21] retrieved candidate question answering (QA) pairs from ES, trained entity recognition model, and built a knowledge graph. A QA system fetches data from two Chinese medical community websites.

Knox C. et al. [16] used ES to search data in Drugbank [22]. Drugbank is knowledgebase of drugs, approved by the U.S. Food and Drug Administration (FDA), investigational drugs, drug–drug and drug–food interactions, other data of drugs.

Lee H.Y. et al. [23] included ES in the MEDBIZ platform. The platform connects home and wearable medical devices of patients with the main part for storage, analysis and presentation of data. Metadata are analyzed in ES.

Lee W. et al. [24] stored in an ES index generated descriptions of electroencephalography data (EEG) during sleep. The index is used to look for EEG according to their text description and automate the generating of the report.

Noor K. et al. [25] used the Cogstack platform as a tool to solve some problems. The problems are clinical trial recruitment, identifying reasons for why patients had an allergic reactions based on reports in the National Reporting and Learning System, improving the clinical referral process for neurology clinics, identify clinical intent in free-text notes, and others. There are three DBMS in Cogstack. One of them is ES.

Otte W.M. et al. [26] used ES for storing almost 36 million articles from PubMed when developing a reference website [27]. The words of the article abstracts were classified to the part of Patient or Intervention of the Patient-Intervention-Control-Outcome (PICO) Framework. Classification of words is validated by the Cochrane Database of Systematic Reviews (CDSR).

Quan X. et al. [28] developed the AIMedGraph software for storing the knowledge count about genes, genetic alterations and their therapeutic and diagnostic relevance. This software system used ES. There are no details of use of ES in the article.

Palm V. et al. [29] used ES and Kibana for storing, searching and visualizing Digital Imaging and Communications in Medicine (DICOM) data. ES and Kibana are core components of the Kaapana open-source toolkit [30]. Kaapana is a platform for storing and analyzing images.

Purpura A. et al. [31] implemented a system for the discovery of semantic entity associations. The system consist of (1) a sentence annotation system, (2) an ES index, and (3) a querying and filtering component. Associations were extract from 14 million PubMed abstracts.

Sadek J. et al. [32] developed the ScanMedicine software [33]. They uploaded formatted descriptions of clinical cases from more than 10 registers of different countries of the world and medical device approval data of the FDA in ES. ES was used to search. Data are synchronized with MongoDB.

Scheable R. et al. [34] automated translation of the Medical Subject Headings (MeSH) in English into another languages by MeSH browser [35]. ES provides a fuzzy full-text search.

Sileo D., Uma K., and Moens M.F. [36] explored a medical multiple-choice question answering (MCQA). One of pretrained data were Wikipedia medical articles. The articles was indexed with ES.

Upadrista V., Nazir S., and Tianfield H. [37] employed ES as database with patient details, health data, and EHRs. Metadata were stored in two blockchains. Source of data was medical devices.

Yu C. et al. [38] added to their Passionfruit Genomic Database [39] ES for accurately search for genes of interest by keywords.

Yue T. et al. [40] used ES to fuzzy search T-cell receptor similar Complementarity Determining Region 3 (CDR3) sequences with a maximum of 1 mismatch.

Zhang N. and Jankowski M. [41] assigned International Classification of Disease (ICD) diagnosis codes to medical documents in their Medical Document BERT (MDBERT). ES was used to search the SNOMED-CT indications by the raw disease name.

3.3 Logging

Jagan P. et al. [42] designed a rehabilitation complex. The complex includes a rehabilitation device, a control device, and a network data storage with their visualization. The rehabilitation device mounted on wheel chair and gives repetitive rehabilitation therapy in seating and sleeping positions. The control device on Raspberry Pi controls the rehabilitation device and sends its parameters and log to the network data storage. The network data storage is ES. To visualize data from ES, Kibana is used.

Nan J. et al. [43] proposed a practice guideline (PG) to promote the Fast Healthcare Interoperability Resources (FHIR) standard. The PG includes practice design and a architecture. Practice design defines the responsibilities of stakeholders and outlines the complete procedure from data to services. The development architecture for practice design lists the available tools for each practice step and provides direct and actionable recommendations. ES in the architecture stores system logs.

Smyrlis M. et al. [44] presented a solution to estimate the attack surface and resilience of medical applications and systems. They named it Risk Assessment for Medical Applications (RAMA). ES is used for storing of alert messages and data visualization in Kibana.

Ulgu M.M. et al. [45] developed the national Disease Management Platform (DMP) in Turkey. The DMP screens and controls of chronic diseases in adherence to evidence-based clinical guidelines. The platform utilizes ES to store patient personal data and basic attributes, their current screening and monitoring statuses for each disease. ES serves as a system log repository as well.

3.4 Other Use Cases

Al-Agil M. et al. [46] programmed the Smart Watcher software system. Apache NiFi extracts clinical data from various sources and sends them into the ES index. Using Elastic Watcher, doctors are notified when documents with a certain text appear in the index. Examples of such notifications are listed in [46], for example, the appearance of patients with a positive test Covid PCR or Out-of-Hospital Cardiac Arrest.

Chen L. et al. [47] examined mapping strategies of Chinese medical entities to their English counterparts in the Unified Medical Language System (UMLS). ES was one of the strategies. The optimal strategy is the linear combination method with the aid of multiple-source web translation engines.

Chen T. et al. [15] developed Integrated Medicinal Plantomics (IMP) [48]. There are analysis modules, including tools for gene annotations, sequences, structures, functions, distributions and expressions in IMP. IMP is utilized by medicinal plants or plants with medicinal benefits. ES is incorporated for data retrieval.

Cheng K.Y. et al. [49] stored and analyzed in ES data of DICOM medical images. Data are image headers and generated keywords. The analysis was used to correct a learning model based on clinical data.

Jin Q. et al. [50] proposed the PM-Search (PM – precision medicine) search engine. PM-Search contains titles, authors, and abstracts of PubMed articles. Data are enriched by synonyms via the National Library of Medicine’s web application programming interface in MedlinePlus. ES is the baseline retriever for initial selection of candidate titles and abstracts. Then the evidence reranker selects next level candidates based on their evidence quality from initial candidates.

Linares M. and Santos L. [51] analyzed selfie-related deaths with help of the Heimdllr-Project tool. The tool is a global system of applied epidemiological intelligence for the detection and analysis of events and outbreaks. ES is used as a DBMS and an artificial intelligence tool. This work utilizes full Elastic Stack (ES, Logstash, Kibana).

Raad M. et al. [52] prioritized 100 genes by a relevance score from ES. The genes involve in the pain pathways and cancer pathways.

Scott-Boyer M.P. et al. [53] compared business intelligence (BI) tools that display biological data in text form from ES. The study included Elastic Kibana, Siren Investigate, Microsoft Power BI, Salesforce Tableau, and Apache Superset for four types of biological data. The study did not reveal a significant advantage of one tool over the others.

Srba I. et al. [54] explored false information in medical articles. They used ES to select a subset of the article-claim pairs for further analysis. The criterion of the selection is the relevance score higher than the 2/3 of the maximum score.

Villena F. et al. [55] trained a model to recognize words related to various diseases, and then used ES to assign disease codes to phrases from clinical reports.

3.5 Types of using Elasticsearch as Database Management System

We classified three type of using ES as DBMS:

- primary DBMS is the only or main DBMS that does not synchronize data on a command from another DBMS;
- secondary DBMS is an auxiliary DBMS that synchronizes data upon a command from another DBMS; ES is used for searching in this case;
- parallel DBMS is one of the DBMS that manages some of the data without synchronization with another DBMS.

We aggregate articles by usage types (table 1).

Table 1. Number of articles and shares of using as a database management system

Type	Article count	Share, %
Primary	11	34.4
Secondary	8	25.0
Parallel	13	40.6
Unknown	2	6.3

3.5 Index Document Count

Some articles refer to index document number (table 2).

Table 2. Document number in the Elasticsearch index

Article	Index document count
[21]	60000000
[26]	35900000
[50]	29000000
[55]	18716629
[31]	14000000
[19]	458594
[54]	317000
[24]	30000
Mean	19802778

3.6 Joint Technologies and Libraries

ES is used in conjunction with the following technologies and libraries mentioned in the articles (table 3). We used ES to aggregate the data for this table.

Table 3. Elasticsearch joint technologies and libraries

Technologies and libraries	Count
Redis	4
Apache Nifi, CogStack, Apache Kafka, Neo4j, PostgreSQL, Python	3
Amazon S3, Apache AirFlow, Apache Superset, Flask, Hugging Face, Kubernetes, MongoDB, MySQL, Vue.js	2
Apache Camel, Apache Druid, Apache ECharts, Apache Spark, Apache Tika, Apache ZooKeeper, Amazon DynamoDB, AWS Athena, AWS Lambda, Angular, BlasterJS, D3.js, DCM4chee, Django, Elastic Watcher, Element UI, Ganache, Grafana, igv.js, JBrowse, Java, Keycloak, Log4j, MariaDB, Microsoft Power BI, MinIO, Eclipse Mosquitto, NeoVis, Node.js, OpenResty, plotly.js, PostgREST, PyTorch, React, React-Admin, Ruby on Rails, Salesforce Tableau, Siren Investigate, Spring Boot, vis.js, ZomboDB	1

3.7 Technical Details

Some articles contain descriptions of technical details of using ES (table 4).

Table 4. Technical details for using Elasticsearch

Article	Details
[24]	The standard tokenizer, the n-gram token filter (from 1 to 4)
[52]	ES version 7.11
[45]	6 servers in a cluster
[31]	The keyword data type
[21]	Boolean query on title and description; the title boost is two times more than the boost of the description
[50]	Boolean query with should and must, keyword matching, the dis_max query with synonyms and boost, the tie_breaker 0.8, the title field boost 3.0

3.8 Information Standards, Databases, and Tools in Medicine

Data formats in the healthcare system are standardized by Health Level 7 (HL7) [56]. The data contain different concepts. The concepts and their relations are collected in the Unified Medical Language System (UMLS) [57]. UMLS uses dictionaries, including the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) dictionary [58]. SNOMED-CT is a collection of multilingual clinical healthcare terminology. Another part of the UMLS is the Medical Subject Headings (MeSH) [59]. MeSH is produced by the National Library of Medicine (NLM) since 1960. It's used for cataloging documents (text, media) and as an index to search these documents. Medical Concept Annotation Tool (MedCAT) extracts data from EHRs and link it to terms of SNOMED-CT or UMLS [60–61].

These standards, databases, and tools are used in some of the surveyed articles (table 5).

Table 5. Number of articles mentioning using MedCAT, MeSH, SNOMED-CT, UMLS

Information standards, databases, and tools	Count
UMLS	5
MedCAT, MeSH	3
SNOMED-CT	2

4. Discussions

ES is used in medical applications mainly for searching and logging. This doesn't differentiate medicine from other uses of ES.

Most often, ES is used as a parallel database (41%) without synchronization with other DBMSs.

On average, researchers store nearly 2 million documents in a ES index.

ES is used with other DBMSs (Redis, Neo4j, PostgreSQL), data transfer systems (Apache Nifi, Apache Kafka), and medical databases and tools such as UMLS, MedCAT, MeSH, SNOMED-CT.

Conclusion

1. We surveyed 34 articles about the use of ES in medical applications, published in 2022-2024. Articles were found via the Google Scholar.
2. We analyzed the articles for function of ES, type of use of ES as a DBMS, number of documents in the index, joint technologies and libraries, technical details, use of medical standards, databases, and tools.

ES is widely used in medical applications mainly for text searching and logging.

References

1. Rogers E.M. Informatization, Globalization, and Privatization in the New Millennium. In Asian Journal of Communication, 2000, 10(2):71-92. DOI: 10.1080/01292980009364785.
2. Wickramasinghe N., Geisler E. (eds.) Encyclopedia of Healthcare Information Systems. IGI Global, 2008.
3. Badgett B. et al. Medical Search Engines. In Wickramasinghe N., Geisler E. (eds.) Encyclopedia of Healthcare Information Systems. IGI Global, 2008:873-881.
4. Lombardi C. et al. Search Engine as a Diagnostic Tool in Difficult Immunological and Allergologic Cases: Is Google Useful? In Internal Medicine Journal, 2009, 39(7):459-464.
5. Thangaraj M., Devi M.K. Survey on Electronic Medical Record Search Engine in Healthcare. In Journal of Positive School Psychology, 2022, 6(5):3360-3380.
6. Wang L. et al. Using Internet Search Engines to Obtain Medical Information: A Comparative Study. In Journal of Medical Internet Research, 2012, 14(3):e74. DOI: 10.2196/jmir.1943.
7. Prutzkow A.V. Informatsionno-poiskovaja sistema Elasticsearch: ucheb. posobie. 2-e izd. M.: Kurs, 2024. 392 s.
8. Prutzkow A.V. Sposob vyjavlenija knig dlja nachal'nogo osvoenija kompleksa programm Elastic Stack i ego rezul'taty // International Journal of Open Information Technologies. 2023. T. 11. № 11. S. 53-57. DOI: 10.13140/RG.2.2.17164.73601.
9. Prutzkow A.V. Informatsionno-poiskovoe myshlenie: kak uskorit' poisk v seti Internet i ne vygoret' // Informatsionnoe obschestvo. 2024. № 4. S. 50-60. DOI: 10.52605/16059921_2024_04_55.
10. Prutzkow A.V. Sposob poiska adresov s nepolnym tekstom zaprosa v sisteme Elasticsearch // Informatsionnyj obmen v mezhdistsiplinarnykh issledovanijakh II: sb. tr. Vseros. nauch.-prakt. konf. s mezhdunar. uchastiem. Rjazan': Akad. FSIN Rossii, 2023. S. 246-250.
11. Prutzkow A.V. Primer uproschenija zaprosa agregatsii posle normalizatsii struktury dokumenta v sisteme Elasticsearch // Sovremennye tekhnologii v nauke i obrazovanii – STNO-2024: sb. tr. 7-go mezhdunar. nauch.-tekhn. foruma: v 10 t. Rjazan': RGRTU, 2024. T. 4. S. 6-12.
12. Sivarajkumar S. et al. Clinical Information Retrieval: A Literature Review. In Journal of Healthcare Informatics Research, 2024. DOI: 10.1007/s41666-024-00159-4.
13. Mazein I. et al. The Use of Graph Databases in Systems Biology: A Systematic Review. 2024.
14. Cao C. et al. PGS-Depot: A Comprehensive Resource for Polygenic Scores Constructed by Summary Statistics Based Methods. In Nucleic Acids Research, 2024, 52(D1):D963-D971.
15. Chen T. et al. IMP: Bridging the Gap for Medicinal Plant Genomics. In Nucleic Acids Research, 2024, 52(D1):D1347-D1354.
16. Knox C. et al. Drugbank 6.0: The Drugbank Knowledgebase for 2024. In Nucleic Acids Research, 2024, 52(D1):D1265-D1275.
17. Alskaf E. et al. Machine Learning Outcome Prediction Using Stress Perfusion Cardiac Magnetic Resonance Reports and Natural Language Processing of Electronic Health Records. In Informatics in Medicine Unlocked, 2024, 44:101418.
18. PGS Depot. URL: http://www.pgsdepot.net. Accessed 2024-04-09.
19. Dilmaç F., Alpkoçak A. Automatic Keyword Assignment System for Medical Research Articles Using Nearest-Neighbor Searches. In Turkish Journal of Electrical Engineering and Computer Sciences, 2022, 30(5):10. DOI: 10.55730/1300-0632.3907.

20. Debauche O. et al. RAMi: A New Real-Time Internet of Medical Things Architecture for Elderly Patient Monitoring. In Information, 2022, 13(9):423. DOI: 10.3390/info13090423.
21. Guo Q., Cao S., Yi Z. A Medical Question Answering System Using Large Language Models and Knowledge Graphs. In International Journal of Intelligent Systems, 2022, 37(11):8548-8564. DOI: 10.1002/int.22955.
22. DrugBank Online | Database for Drug and Drug Target Info. URL: https://go.drugbank.com. Accessed 2024-04-09.
23. Lee H.Y. et al. Internet of Medical Things-Based Real-Time Digital Health Service for Precision Medicine: Empirical Studies Using MEDBIZ Platform. In Digital Health, 2023, 9:20552076221149659.
24. Lee W. et al. Automated Clinical Impression Generation for Medical Signal Data Searches. In Applied Sciences, 2023, 13(15):8931.
25. Noor K. et al. Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: Cogstack at University College London Hospitals. In JMIR Medical Informatics, 2022, 10(8):e38122. DOI: 10.2196/38122.
26. Otte W.M. et al. Fast Clinical Trial Identification Using Fuzzy-Search Elastic Searches: Retrospective Validation with High-Quality Cochrane Benchmark. In medRxiv, 2023:2023.09.06.23295135.
27. Browse Page. URL: https://evidencehunt.com/browse/. Accessed 2024-04-09.
28. Quan X. et al. AIMedGraph: A Comprehensive Multi-Relational Knowledge Graph for Precision Medicine. In Database, 2023, 2023:baad006. DOI: 10.1093/database/baad006.
29. Palm V. et al. AI-Supported Comprehensive Detection and Quantification of Biomarkers of Subclinical Widespread Diseases at Chest CT for Preventive Medicine. In Healthcare, 2022, 10(11):2166. DOI: 10.3390/healthcare10112166.
30. Scherer J. at al. Joint Imaging Platform for Federated Clinical Data Analytics. In JCO Clinical Cancer Informatics, 4:1027–1038.DOI: 10.1200/CCI.20.00045.
31. Purpura A., Bonin F., Bettencourt-Silva J. Accelerating the Discovery of Semantic Associations from Medical Literature: Mining Relations between Diseases and Symptoms. In the Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022:77-89.
32. Sadek J. et al. ScanMedicine: An Online Search System for Medical Innovation. In Contemporary Clinical Trials, 2023, 125:107042. DOI: 10.1016/j.cct.2022.107042.
33. ScanMedicine (NIHRIO) Searching Global Medical Innovation. URL: http://www.scanmedicine.com. Accessed 2024-04-09.
34. Scheible R. et al. A Multilingual Browser Platform for Medical Subject Headings. In Informatics and Technology in Clinical Care and Public Health, 2022, 289:384.
35. MeSH Browser. URL: https://mesh-browser.de/. Accessed 2024-04-09.
36. Sileo D., Uma K., Moens M.F. Generating Multiple-Choice Questions for Medical Question Answering with Distractors and Cue-Masking. In arXiv Preprint arXiv:2303.07069, 2023.
37. Upadrista V., Nazir S., Tianfield H. Consortium Blockchain for Reliable Remote Health Monitoring, 2024. DOI: 10.21203/rs.3.rs-2297411/v1.
38. Yu C. et al. Passionfruit Genomic Database (PGD): A Comprehensive Resource for Passionfruit Genomics. In BMC Genomics, 2024, 25(1):157.
39. Passionfruit Genomic Database. URL: http://passionfruit.com.cn. Accessed 2024-04-09.
40. Yue T. et al. TCRosetta: An Integrated Analysis and Annotation Platform for T-Cell Receptor Sequences. In Genomics, Proteomics & Bioinformatics, 2024:qzae013.
41. Zhang N., Jankowski M. Hierarchical BERT for Medical Document Understanding. In arXiv preprint arXiv:2204.09600, 2022.
42. Jagan P. et al. XoRehab: IoT Enabled Wheelchair Based Lower Limb Rehabilitation System. In 45th International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2023:1-5.
43. Nan J. et al. Designing Interoperable Health Care Services based on Fast Healthcare Interoperability Resources: Literature Review. In JMIR Medical Informatics, 2023, 11(1):e44842.
44. Smyrlis M. et al. RAMA: A Risk Assessment Solution for Healthcare Organizations. In International Journal of Information Security, 2024:1-18. DOI: 10.1007/s10207-024-00820-4.
45. Ulgu M.M. et al. A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report. In JMIR Medical Informatics, 2024, 12(1):e49986. DOI: 10.2196/49986.

46. Al-Agil M. et al. Enhancing Clinical Data Retrieval with Smart Watchers: A NiFi-Based ETL Pipeline for Elasticsearch Queries. 2023.
47. Chen L. et al. Mapping Chinese Medical Entities to the Unified Medical Language System. In Health Data Science, 2023, 3:0011. DOI: 10.34133/hds.0011.
48. IMP. URL: https://www.bic.ac.cn/IMP. Accessed 2024-04-15.
49. Cheng K.Y. et al. An Image Retrieval Pipeline in a Medical Data Integration Center. In Studies in Health Technology and Informatics, 2024, 310:1388-1389. DOI:10.3233/SHTI231208.
50. Jin Q. et al. State-of-the-Art Evidence Retriever for Precision Medicine: Algorithm Development and Validation. In JMIR Medical Informatics, 2022, 10(12):e40743. DOI: 10.2196/40743.
51. Linares M., Santos L. Selfie-Related Deaths using Web Epidemiological Intelligence Tool (2008–2021): A Cross-Sectional Study. In Journal of Travel Medicine, 2022, 1:4.
52. Raad M. et al. Personalized Medicine in Cancer Pain Management. In Journal of Personalized Medicine, 2023, 13(8):1201.
53. Scott-Boyer M.P. et al. Use of Elasticsearch-Based Business Intelligence Tools for Integration and Visualization of Biological Data. In Briefings in Bioinformatics, 2023, 24(6):bbad348. DOI: 10.1093/bib/bbad348.
54. Srba I. et al. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. In the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022:2949-2959.
55. Villena F. et al. Automatic Coding at Scale: Design and Deployment of a Nationwide System for Normalizing Referrals in the Chilean Public Healthcare System. In arXiv, 2023:2307.05560.
56. Beeler W. HL7 Version 3 – an Object-Oriented Methodology for Collaborative Standards Development. In International Journal of Medical Information, 1998, 48:151–161.
57. Schuyler P. et al. The UMLS Metathesaurus: Representing Different Views of Biomedical Concepts. In Bulletin of the Medical Library Association, 1993, 81:217–22
58. Donnelly K. et al. SNOMED-CT: The Advanced Terminology and Coding System for Ehealth. In Studies in Health Technology and Informatics, 2006, 121:279.
59. Rogers F.B. Medical Subject Headings. In Bulletin of the Medical Library Association, 1963, 51:114-116.
60. Fodeh S.J. et al. MedCAT: A Framework for High Level Conceptualization of Medical Notes. In IEEE 13th International Conference on Data Mining Workshops, 2013:274-280. DOI: 10.1109/ICDMW.2013.89.
61. Kraljevic Z et al. Multi-Domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. In Artificial Intelligence in Medicine, 2021, 117:102083. DOI: 10.1016/j.artmed.2021.102083.

ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА ELASTICSEARCH В МЕДИЦИНЕ: ТЕХНИЧЕСКИЙ ОБЗОР

Пруцков Александр Викторович

Доктор технических наук, доцент
Рязанский государственный радиотехнический университет им. В. Ф. Уткина, кафедра вычислительной и прикладной математики, профессор
Липецкий государственный педагогический университет имени П. П. Семенова-Тян-Шанского, кафедра информатики, информационных технологий и защиты информации, профессор
Рязань, Российская Федерация
mail@prutzkow.com

Аннотация

Информатизация воздействует на все сферы жизни общества, в том числе и медицину. Одним из направлений информатизации медицины является использование информационных систем. Elasticsearch (ES) – это информационно-поисковая система. Она используется для хранения и поиска данных в информационных системах. Проанализированы 34 статьи о использовании ES в медицине, опубликованные в 2022–2024 годах. Статьи на английском языке и в свободном доступе были отобраны по типу, описанию использования ES. ES использует, главным образом, для поиска и протоколирования работы. В 41% статей ES используется как СУБД, хранящая только часть данных, в 34% – как первичная СУБД, в 25% – как вторичная СУБД. В среднем в ES хранится почти 2 миллиона документов. Чаще всего с ES используются Redis, Apache Nifi, CogStack, Apache Kafka, Neo4j, PostgreSQL, Python. Собраны некоторые технические детали использования ES (индексаторы, типы данных, виды запросов). ES используется вместе с медицинскими базами данными и инструментами, такими как UMLS, MedCAT, MeSH, SNOMED-CT.

Ключевые слова

Elasticsearch, Elastic Stack, медицина, информатизация, информационно-поисковые системы, информационный поиск, UMLS, MedCAT, MeSH, SNOMED-CT, обзор

Литература

1. Rogers E.M. Informatization, Globalization, and Privatization in the New Millennium. In Asian Journal of Communication, 2000, 10(2):71-92. DOI: 10.1080/01292980009364785.
2. Wickramasinghe N., Geisler E. (eds.) Encyclopedia of Healthcare Information Systems. IGI Global, 2008.
3. Badgett B. et al. Medical Search Engines. In Wickramasinghe N., Geisler E. (eds.) Encyclopedia of Healthcare Information Systems. IGI Global, 2008:873-881.
4. Lombardi C. et al. Search Engine as a Diagnostic Tool in Difficult Immunological and Allergologic Cases: Is Google Useful? In Internal Medicine Journal, 2009, 39(7):459-464.
5. Thangaraj M., Devi M.K. Survey on Electronic Medical Record Search Engine in Healthcare. In Journal of Positive School Psychology, 2022, 6(5):3360-3380.
6. Wang L. et al. Using Internet Search Engines to Obtain Medical Information: A Comparative Study. In Journal of Medical Internet Research, 2012, 14(3):e74. DOI: 10.2196/jmir.1943.
7. Пруцков А.В. Информационно-поисковая система Elasticsearch: учеб. пособие. 2-е изд. М.: Курс, 2024. 392 с.
8. Пруцков А.В. Способ выявления книг для начального освоения комплекса программ Elastic Stack и его результаты // International Journal of Open Information Technologies. 2023. Т. 11. № 11. С. 53-57. DOI: 10.13140/RG.2.2.17164.73601.
9. Пруцков А.В. Информационно-поисковое мышление: как ускорить поиск в сети Интернет и не выгореть // Информационное общество. 2024. № 4. С. 50-60. DOI: 10.52605/16059921_2024_04_55.
10. Пруцков А.В. Способ поиска адресов с неполным текстом запроса в системе Elasticsearch // Информационный обмен в междисциплинарных исследованиях II: сб. тр. Всерос. науч.-практ. конф. с междунар. участием. Рязань: Акад. ФСИН России, 2023. С. 246-250.
11. Пруцков А.В. Пример упрощения запроса агрегации после нормализации структуры документа в системе Elasticsearch // Современные технологии в науке и образовании –

СТНО-2024: сб. тр. 7-го междунар. науч.-техн. форума: в 10 т. Рязань: РГРТУ, 2024. Т. 4. С. 6-12.

12. Sivarajkumar S. et al. Clinical Information Retrieval: A Literature Review. In Journal of Healthcare Informatics Research, 2024. DOI: 10.1007/s41666-024-00159-4.
13. Mazein I. et al. The Use of Graph Databases in Systems Biology: A Systematic Review. 2024.
14. Cao C. et al. PGS-Depot: A Comprehensive Resource for Polygenic Scores Constructed by Summary Statistics Based Methods. In Nucleic Acids Research, 2024, 52(D1):D963-D971.
15. Chen T. et al. IMP: Bridging the Gap for Medicinal Plant Genomics. In Nucleic Acids Research, 2024, 52(D1):D1347-D1354.
16. Knox C. et al. Drugbank 6.0: The Drugbank Knowledgebase for 2024. In Nucleic Acids Research, 2024, 52(D1):D1265-D1275.
17. Alskaf E. et al. Machine Learning Outcome Prediction Using Stress Perfusion Cardiac Magnetic Resonance Reports and Natural Language Processing of Electronic Health Records. In Informatics in Medicine Unlocked, 2024, 44:101418.
18. PGS Depot. URL: http://www.pgsdepot.net. Accessed 2024-04-09.
19. Dilmaç F., Alpkoçak A. Automatic Keyword Assignment System for Medical Research Articles Using Nearest-Neighbor Searches. In Turkish Journal of Electrical Engineering and Computer Sciences, 2022, 30(5):10. DOI: 10.55730/1300-0632.3907.
20. Debauche O. et al. RAMi: A New Real-Time Internet of Medical Things Architecture for Elderly Patient Monitoring. In Information, 2022, 13(9):423. DOI: 10.3390/info13090423.
21. Guo Q., Cao S., Yi Z. A Medical Question Answering System Using Large Language Models and Knowledge Graphs. In International Journal of Intelligent Systems, 2022, 37(11):8548-8564. DOI: 10.1002/int.22955.
22. DrugBank Online | Database for Drug and Drug Target Info. URL: https://go.drugbank.com. Accessed 2024-04-09.
23. Lee H.Y. et al. Internet of Medical Things-Based Real-Time Digital Health Service for Precision Medicine: Empirical Studies Using MEDBIZ Platform. In Digital Health, 2023, 9:20552076221149659.
24. Lee W. et al. Automated Clinical Impression Generation for Medical Signal Data Searches. In Applied Sciences, 2023, 13(15):8931.
25. Noor K. et al. Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: Cogstack at University College London Hospitals. In JMIR Medical Informatics, 2022, 10(8):e38122. DOI: 10.2196/38122.
26. Otte W.M. et al. Fast Clinical Trial Identification Using Fuzzy-Search Elastic Searches: Retrospective Validation with High-Quality Cochrane Benchmark. In medRxiv, 2023:2023.09.06.23295135.
27. Browse Page. URL: https://evidencehunt.com/browse/. Accessed 2024-04-09.
28. Quan X. et al. AIMedGraph: A Comprehensive Multi-Relational Knowledge Graph for Precision Medicine. In Database, 2023, 2023:baad006. DOI: 10.1093/database/baad006.
29. Palm V. et al. AI-Supported Comprehensive Detection and Quantification of Biomarkers of Subclinical Widespread Diseases at Chest CT for Preventive Medicine. In Healthcare, 2022, 10(11):2166. DOI: 10.3390/healthcare10112166.
30. Scherer J. at al. Joint Imaging Platform for Federated Clinical Data Analytics. In JCO Clinical Cancer Informatics, 4:1027–1038.DOI: 10.1200/CCI.20.00045.
31. Purpura A., Bonin F., Bettencourt-Silva J. Accelerating the Discovery of Semantic Associations from Medical Literature: Mining Relations between Diseases and Symptoms. In the Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022:77-89.
32. Sadek J. et al. ScanMedicine: An Online Search System for Medical Innovation. In Contemporary Clinical Trials, 2023, 125:107042. DOI: 10.1016/j.cct.2022.107042.
33. ScanMedicine (NIHRIO) Searching Global Medical Innovation. URL: http://www.scanmedicine.com. Accessed 2024-04-09.
34. Scheible R. et al. A Multilingual Browser Platform for Medical Subject Headings. In Informatics and Technology in Clinical Care and Public Health, 2022, 289:384.
35. MeSH Browser. URL: https://mesh-browser.de/. Accessed 2024-04-09.
36. Sileo D., Uma K., Moens M.F. Generating Multiple-Choice Questions for Medical Question Answering with Distractors and Cue-Masking. In arXiv Preprint arXiv:2303.07069, 2023.

37. Upadrista V., Nazir S., Tianfield H. Consortium Blockchain for Reliable Remote Health Monitoring, 2024. DOI: 10.21203/rs.3.rs-2297411/v1.
38. Yu C. et al. Passionfruit Genomic Database (PGD): A Comprehensive Resource for Passionfruit Genomics. In BMC Genomics, 2024, 25(1):157.
39. Passionfruit Genomic Database. URL: http://passionfruit.com.cn. Accessed 2024-04-09.
40. Yue T. et al. TCRosetta: An Integrated Analysis and Annotation Platform for T-Cell Receptor Sequences. In Genomics, Proteomics & Bioinformatics, 2024:qzae013.
41. Zhang N., Jankowski M. Hierarchical BERT for Medical Document Understanding. In arXiv preprint arXiv:2204.09600, 2022.
42. Jagan P. et al. XoRehab: IoT Enabled Wheelchair Based Lower Limb Rehabilitation System. In 45th International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2023:1-5.
43. Nan J. et al. Designing Interoperable Health Care Services based on Fast Healthcare Interoperability Resources: Literature Review. In JMIR Medical Informatics, 2023, 11(1):e44842.
44. Smyrlis M. et al. RAMA: A Risk Assessment Solution for Healthcare Organizations. In International Journal of Information Security, 2024:1-18. DOI: 10.1007/s10207-024-00820-4.
45. Ulgu M.M. et al. A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report. In JMIR Medical Informatics, 2024, 12(1):e49986. DOI: 10.2196/49986.
46. Al-Agil M. et al. Enhancing Clinical Data Retrieval with Smart Watchers: A NiFi-Based ETL Pipeline for Elasticsearch Queries. 2023.
47. Chen L. et al. Mapping Chinese Medical Entities to the Unified Medical Language System. In Health Data Science, 2023, 3:0011. DOI: 10.34133/hds.0011.
48. IMP. URL: https://www.bic.ac.cn/IMP. Accessed 2024-04-15.
49. Cheng K.Y. et al. An Image Retrieval Pipeline in a Medical Data Integration Center. In Studies in Health Technology and Informatics, 2024, 310:1388-1389. DOI:10.3233/SHTI231208.
50. Jin Q. et al. State-of-the-Art Evidence Retriever for Precision Medicine: Algorithm Development and Validation. In JMIR Medical Informatics, 2022, 10(12):e40743. DOI: 10.2196/40743.
51. Linares M., Santos L. Selfie-Related Deaths using Web Epidemiological Intelligence Tool (2008–2021): A Cross-Sectional Study. In Journal of Travel Medicine, 2022, 1:4.
52. Raad M. et al. Personalized Medicine in Cancer Pain Management. In Journal of Personalized Medicine, 2023, 13(8):1201.
53. Scott-Boyer M.P. et al. Use of Elasticsearch-Based Business Intelligence Tools for Integration and Visualization of Biological Data. In Briefings in Bioinformatics, 2023, 24(6):bbad348. DOI: 10.1093/bib/bbad348.
54. Srba I. et al. Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. In the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022:2949-2959.
55. Villena F. et al. Automatic Coding at Scale: Design and Deployment of a Nationwide System for Normalizing Referrals in the Chilean Public Healthcare System. In arXiv, 2023:2307.05560.
56. Beeler W. HL7 Version 3 – an Object-Oriented Methodology for Collaborative Standards Development. In International Journal of Medical Information, 1998, 48:151–161.
57. Schuyler P. et al. The UMLS Metathesaurus: Representing Different Views of Biomedical Concepts. In Bulletin of the Medical Library Association, 1993, 81:217–22
58. Donnelly K. et al. SNOMED-CT: The Advanced Terminology and Coding System for Ehealth. In Studies in Health Technology and Informatics, 2006, 121:279.
59. Rogers F.B. Medical Subject Headings. In Bulletin of the Medical Library Association, 1963, 51:114-116.
60. Fodeh S.J. et al. MedCAT: A Framework for High Level Conceptualization of Medical Notes. In IEEE 13th International Conference on Data Mining Workshops, 2013:274-280. DOI: 10.1109/ICDMW.2013.89.
61. Kraljevic Z et al. Multi-Domain Clinical Natural Language Processing with MedCAT: The Medical Concept Annotation Toolkit. In Artificial Intelligence in Medicine, 2021, 117:102083. DOI: 10.1016/j.artmed.2021.102083.