ACHSM Asia-Pacific Journal of Health Management shape

RESEARCH ARTICLE

COMPARISON OF KEYWORD SEARCH TECHNIQUES WITH RESPECT TO ELECTRONIC HEALTH RECORDS

Prachi Gurav, Sanjeev Panandikar

National Institute of Industrial Engineering, Mumbai, India

Correspondence: prachigurav19@gmail.com

ABSTRACT

As the world progresses towards automation, manual search for data from large databases also needs to keep pace. When the database includes health data, even minute aspects need careful scrutiny. Keyword search techniques are helpful in extracting data from large databases. There are two keyword search techniques: Exact and Approximate. When the user wants to search through EHR, a short search time is expected. To this end, this work investigates Metaphone (Exact search) and Similar_Text (approximate search) Techniques. We have applied keyword search to the data, which includes the symptoms and names of medicines. Our results indicate that the search time for Similar_text is better than for Metaphone.

KEYWORDS

Electronic Health Records (EHR), Keyword search, Approximate keyword search, Exact keyword search, Metaphone, Similar_text

INTRODUCTION

As the use of Electronic Health Records (EHR) gains momentum, there are immense opportunities for their use in healthcare research and patient treatment. Every patient-doctor interaction has queries related to the previous ailment and medicines prescribed therefor. A mere 10.9% of the patients could memorize drug names [2], which poses difficulties to the doctor in keeping track of previous medication. This creates an enormous scope for EHR. For instance, if the doctor requires the patient's history of hypertension or needs to access data on previous medication, all the physician needs to do is to pull out the HER, for ready access to the information.

EHR comprises data relating to medical history, demographics, lab reports, etc., all of which make EHR voluminous. A manual search of such huge data would be extremely laborious and time-consuming. Consequently, keyword search is the more efficient and expeditious alternative to manual search.

Keywords are ideas and topics that define what the content is about. They are the words and phrases that users enter into search engines. Such keywords can be classified as exact match keywords and approximate match keywords.

This paper aims to quantify the performance of these two techniques, in the context of EHR. We have used the

Metaphone algorithm for exact keyword search and the Similar_text algorithm for approximate search.

LITERATURE REVIEW

Our search was initiated with "Keyword search in EHR," but did not elicit appropriate and relevant articles. We hence modified our search to "Natural language processing in Electronic Health Records." The timeline selected was 2016-2021. This yielded a total of 1945 articles in the first stage, which were filtered for relevance based on the abstracts, reducing the article count to 378. After the perusal of the full articles, we could select 54 for final review. These 54 articles work on seven different levels, of which the phonetic level works on exact keyword search, and as the morphological level deals with insertion, deletion, suffixes, and prefixes, it falls under the category of approximate keyword search. Of these, one paper used exact keyword search, while 3 papers used approximate keyword search. The following table summarizes the literature review.

TABLE 1: LITERATURE REVIEW OF KEYWORD SEARCH

ARTICLE	CONCEPT	LIMITATIONS	KEYWORD SEARCH TYPE
[4]	Uses a combination of string and phonetic search to analyze unstructured medical data. The results show that the combination produces better results than the traditional string distance metrics for misspelled words. The technique is applied to drug names.	Limited to Portuguese language, drug names	Exact keyword search
[6]	Uses surgical pathology and emergency department notes to identify misspelling by using Levenshtein distance algorithm	Uses small corpora of surgical pathology and emergency department documents.	Uses Levenshtein distance algorithm-approximate keyword search
[3]	Automated HIV risk analysis from EHR	No performance improvement with the use of empirical methods; Unigram model did not account for Unigram; negation consideration; information loss due to use of template notes; external validation of model, lack of interoperability	Used approximate keyword search

The aforesaid studies are confined to a particular disease or department, for instance, Cancer or the Emergency Department. However, when it comes to drug interaction, the doctor requires information regarding the medications currently used by the patient. The physician also needs to study drug interaction, because when the patient is taking multiple medicines, he/she is susceptible to side effects. Drug interaction refers to the impact one medicine has on another. Medicines can also interact with alcohol and even some food items; some of these interactions can be serious, even life-threatening. We have hence taken data from every patient-doctor interaction and applied keyword search to symptoms and medicine names. Applying the Metaphone and Similar_text algorithms, we evaluated the comparative benefits of exact and approximate search.

KEYWORD SEARCH

When we input words to locate information, the words we search with are called 'keywords.' Keywords are the keys to unlocking the information we require. This section focuses on the different algorithms used to match keywords, based on the sound or spelling difference.

EXACT KEYWORD SEARCH TECHNIQUES:

[1] studied the SoundEx technique. While translating a string into canonical form, a code of maximum 4-letters is used. The algorithm depends on the first character. This technique has a few limitations like noise intolerance, differing transcription systems, names with particles, silent consonants, name syntax inconsistencies, weak precision, etc. It is suitable for applications with high false-positive and high false negatives.

[5] Henry name matching is based on the Rusell SoundEx method, with the important difference that the earlier method used a 3-letter code. This technique is suitable for the French language.[5] investigated the Metaphone algorithm, in which the system ignores the vowels after the first letter or retains vowels as they are if the string starts with a vowel. It ignores double letters. It substitutes 'o' for 'th' and 'X' for 'sh'. [5] explored the K-approximation method, which attempts to ascertain the difference between the entered string and a string that is part of the text. [5] mentioned Guth name matching, which performs a letter-by-letter comparison. As per the findings of [5], the Metaphone algorithm has the highest accuracy and average execution time. Though the SoundEx too has the same accuracy as Metaphone, due to the aforesaid limitations of the SoundEx algorithm, we considered the Metaphone algorithm for implementation and comparison.

APPROXIMATE KEYWORD SEARCH TECHNIQUES:

A similar text algorithm system checks for variations in the string by insertion, deletion, and substitution. The number of matching characters is calculated by finding the longest first common substring and repeating the procedure for the prefixes and the suffixes, recursively. The lengths of all the common sub-strings found are added. The Levenshtein algorithm checks the similarity of two strings by calculating single-letter edits (insertion, deletion, substitution).

IMPLEMENTATION AND RESULTS

We have implemented the algorithms using the WAMP server with PHP and MySQL. We have created databases in MySQL incorporating information relating to doctors, patients, and relatives. The patient database includes fields such as the name of the patient, his/her unique id, date of record insertion, symptoms, doctor's specialization, medicines, and 'Medication_till_date'. To complete this database, we sourced data from 'webMD' and drug.com websites. This data includes names of medicines and symptoms to which the medicines are applied. A Random function was used to create records in Excel. We imported this database to MySQL in the WAMP server.

After implementing the algorithms, we applied them to the database of 1062 records. The results are depicted in Table 2 below.

TABLE 2: PERFORMANCE OF KEYWORD SEARCH TECHNIQUES

		SEARCH TIME IN SECONDS		
Medicine	**Occurrence**	**Metaphone**	**Similar_Text**	**Comments**
Atovaquone	7	0.01473	0.00691	
Atovaquone and Proguanil	12	0.02778	0.01146	No result for Metaphone

Clindamycin	7	0.03208	0.01146	
Doxycycline	7	0.03208	0.01189	
Doxycycline tablets and capsules	49	0.05114	0.01531	No result for Metaphone
primaquine	7	0.02184	0.00978	
Adoxa CK	7	0.02314	0.01035	No result for Metaphone
Adoxa Pak	7	0.02314	0.01272	No result for Metaphone
Adoxa TT	7	0.02129	0.01044	No result for Metaphone
Alodox	6	0.02038	0.00971	
Amoxicillin	7	0.03210	0.01041	
Amoxicillin and Clavulanate Potassium	26	0.03303	0.01202	No result for Metaphone
AmoxicillinA	43	0.02196	0.01002	
Amoxil	7	0.02020	0.00944	
Artemether and Lumefantrine	7	0.02997	0.01176	No result for Metaphone
Avidoxy	7	0.01605	0.01605	
Azelastine HCL drops	19	0.02664	0.01001	No result for Metaphone
Azithromycin	43	0.02543	0.01023	
Carbinoxamine syrup	23	0.02779	0.01007	No result for Metaphone
Crocin	25	0.01199	0.00916	
Cyproheptadine HCL	682	0.04867	0.00657	
Desloratadine	690	0.01533	0.00876	
Doryx	7	0.00919	0.00507	
Doxycycline delayed released tablets	954	0.01646	0.00873	No result for Metaphone
Emadine	19	0.01022	0.00445	
Hydroxycloroquine	19	0.01772	0.00986	
Hydroxyzine HCL	7	0.01913	0.0913	No result for Metaphone
Levocetrizine Dihydrochloride	1062	0.02193	0.00730	No result for Metaphone
Livostine	7	0.01597	0.00807	
Mefloquine	7	0.01464	0.00862	
Metronidazole	1062	0.01502	0.00884	
Morgidox	13	0.02264	0.01023	
Moxatag	7	0.01077	0.00662	
Oracea	7	0.01077	0.00419	
Paracetamol	7	0.01636	0.00616	
Qunidine	1062	0.00487	0.00251	

Rantack	13	0.01960	0.00940	
Sinarest	13	0.07571	0.01583	
Trimox	7	0.02822	0.00954	

TABLE 3: COMPARISON OF ALGORITHMS:

POINTS	METAPHONE ALGORITHM	SIMILAR_TEXT
Concept	Search based on sound	Search based on character sequences
Result	Displays words with a similar sound	Displays words with a similar character sequence
Best	When spelling matches with sound	When character sequences match
Execution time for 1000 records	0.039 seconds	0.013 seconds
Execution time for string length 3(Min Length)	0.014 seconds	0.006 seconds
Execution time for string length 14 (Max length)	0.035 seconds	0.010
Execution time for String length 8 (average length)	0.016 seconds	0.006 seconds
Spaces in keywords	Not accepted	Accepted
Execution time for string length 14 with 1 character change	0.024 seconds	0.010 seconds
Execution time for string length 14 with 2 character changes	0.024 seconds	0.010 seconds
Execution time for string length 14 with 1 character change	0.029 seconds	0.010 seconds

DISCUSSION

The main contribution of this work is the use of keyword search to expedite access to and perusal of EHR. This system attempts to search through the entire EHR. We used both the exact and approximate keyword searches. Our results demonstrated that the approximate keyword search with Similar_Text is faster than the exact keyword search, using Metaphone. We can hence conclude that the search time for the Metaphone algorithm depends on string length. However, in the case of Similar_Text, the search time remains constant for the minimum and average string lengths (3 and 8 characters respectively). It changes when the user desires to search a string with the maximum number of characters. In the case of a misspelled string with 1 and 2-character change using Metaphone, the search time remains constant, i.e., 0.024 seconds. When the maximum length string with a 3-character change is searched, it took 0.029 seconds. Search time for 1, 2, and 3-character changes remained constant for Similar_Text, at 0.010 seconds. For Similar_Text, the in-between string spaces are accepted, which is not possible with Metaphone.

We have used only 1000 records. It is possible to apply and verify results with larger datasets. When we entered the keyword Adoxa tt, the system checked for titi and not for the sound 't.' Further, it did not consider the in-between spaces, which resulted in more false negatives
with Metaphone.

Our keyword search technique has several potential clinical applications. For example, it can be used to assist physicians at the point of care to quickly review the patient's history. Additionally, this system facilitates studying drug interaction. For example, when the patient is under medication for hypertension and also suffers from an allergy, the physician can explore whether the allergy could be due to the beta-blocker in the hypertension medicine. The system helps review patient history when the patient forgets to present the file of previous prescriptions and cannot remember the names of the medicines.

LIMITATIONS AND FUTURE WORK:

By using the keyword search, we are dealing with the phonetic and morphological levels of NLP; to study drug interaction in greater depth, the application of a pragmatic level of NLP to EHR is essential. This is a topic for future study. For example, from the previous symptoms and medications, the system could predict the possible side effects the patient could experience. Future work should study possible performance improvement of Metaphone when there is the inclusion of spaces.

CONCLUSION

For studying the performance of keyword search techniques in her, we used two techniques: exact and approximate search, with Metaphone and Similar_Text. This approach demonstrates the potential support to physicians to have a quick overview of patient history and to prepare a new treatment approach. Similar_Text is faster than Metaphone, which is helpful during an emergency when timely retrieval of information is critical. As Metaphone has not worked well on texts with spaces, there is scope for improving its performance.

References:

1. A.J.Lait, B. R. (n.d.). An assessment of name matching algorithms.
2. Ahmet Akici, S. K. Patient knowledge about drugs prescribed at primary healthcare facilities. *Pharmacoepidemiology and drug safety, 13*, 871-876. (2004).
3. Daniel J. Felleer, J. Using clinical notes and natural language processing for automated HIV risk assessment. *J. Acquire Immune Defic Syndr., 72*(2), 160-166. (2018).
4. Hegler Tissot, R. Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in portuguese. *Journal of biomedical semantics, 10(suppl 1)*(17). (2019).
5. Hettiarachchi, G. SPARCL: An improved approach for matching sinhalese words and names in record clustering and linkage. (2012).
6. Workman, T.E., Shao, Y., Divita, G. *et al.* An efficient prototype method to identify and correct misspellings in clinical text. *BMC Res Notes* **12,** 42 (2019).