Comprehensive Journal of Science
Volume (9), Issue (33), (Dec, 2024)
ISSN: 3014-6266

مجلة العلوم الشاملة
المجلد (9) العدد 33 ديسمبر 2024
ردمد: 3014-6266

Modified Soundex Algorithm with Pre-processing for Arabic Names

Mohammed F. Al Boashi[1], Saad A. Al Deeb[2]

[1,2] Department of Computer Science, Faculty of Arts and Sciences, Al-Marqab University, Khoms, Libya

[1]mfalboashi@elmergib.edu.ly, [2]saaldeeb@elmergib.edu.ly

تاريخ الاستلام:2024/8/12 -تاريخ المراجعة: 2024/9/13- تاريخ القبول: 2024/11/5- تاريخ للنشر : 2024 /12/17

Abstract

Search algorithms play a crucial role in enabling effective and interactive retrieval and querying experiences. This research introduces a new search function designed specifically for retrieving Arabic names stored within databases. Dubbed the Modified Arabic Soundex Function, it is presented as a method to enhance the search and retrieval process for Arabic names.

Inspired by the English Soundex Function, which originated in 1918, the Modified Soundex Function is designed to address the unique characteristics of Arabic names. By processing the first letter of names, the proposed algorithm evaluates the "degree of closeness" between compared names, allowing it to accommodate variations in spelling and pronunciation inherent to Arabic names.

This research provides a detailed explanation of the steps involved in the proposed algorithm, demonstrating its effectiveness in improving the accuracy and efficiency of the search and retrieval process for Arabic names from databases.

The analysis of the performance of the Arabic name classification system using the Soundex algorithm showed notable results, with an overall precision of 79.1% and a recall of 91.9%. These percentages indicate the system's high ability to classify names accurately, though there is a need to improve precision to reduce errors. Overall, these results enhance the system's potential in phonetic name classification applications, highlighting the necessity for ongoing algorithm improvements.

Keywords: Arabic Soundex, Name Retrieval, Natural Language Processing, Soundex algorithm.

Introduction

Data retrieval plays a vital role in modern information systems, serving as the backbone for numerous critical processes. From large-scale databases to individual applications, the ability to efficiently retrieve data is paramount for ensuring smooth operations and informed decision-making.

Data retrieval enables access to valuable information stored in various repositories. Whether retrieving customer records from a database, accessing financial data for analysis, or retrieving scientific research findings, the ability to retrieve data swiftly and accurately is fundamental for organizations across all sectors [2].

When searching for a specific name in various databases, we may be surprised to find that the name does not exist. This could be due to misspelling the name when entered, the name not being stored in the database, or perhaps the user not being sure of the name.

Conventional search processes do not address this issue. Therefore, we aim to enhance the efficiency of searching for Arabic names. The search will be enhanced to display the correct name if it exists or show closely related names based on a specified degree of proximity. This will be achieved by developing a Soundex algorithm adapted to the Arabic language to determine the degree of closeness between names.

Soundex function

The Soundex algorithm is a phonetic algorithm used to convert words into specific character strings that approximately represent the sound of those words. This algorithm was first developed in 1918 by Robert Russell, primarily designed for searching databases for words or names that may sound similar but are spelled differently.

The Soundex algorithm consists of a set of rules that define how words are converted into character strings. These rules are based on English phonetics and spelling, where each letter in the word is represented by a specific sound key. These rules are applied to the word to generate a standardized character string representing the basic sound of the word [1].

The Soundex algorithm is commonly used in databases to match words that may sound similar but are spelled differently. This improvement enhances researchers' ability to find matching records, especially when they have incomplete information or when there are spelling errors in the input data.

The simplicity and effectiveness of the Soundex algorithm in handling words with similar sounds make it a valuable tool in many applications such as data management systems, online search, and people and location search applications [3][7].

Study Objectives:

1. Review the original Soundex algorithm and explain how it works

2. Analyze the challenges and potential problems in applying the Soundex algorithm to Arabic names.
3. Apply the Soundex algorithm to a sample of Arabic names and evaluate its efficiency in search and matching operations.
4. Explore the potential uses of the Soundex algorithm in various applications in the Arabic language.

Study Methodology

The study methodology includes a comprehensive review of the literature related to the Soundex algorithm and analysis of previous studies conducted in this field. Additionally, the Soundex algorithm is applied to a sample of Arabic names, and the results are analyzed and evaluated for efficiency in search and matching operations.

Related Work

Researchers H. S. Al-Khalifa and M. Al-Muhtaseb introduced a new phonetic algorithm aimed at improving the retrieval of Arabic names. This study focuses on modifying the Soundex algorithm to accommodate the unique characteristics of Arabic names, including variations in spelling and pronunciation. The modified algorithm enhances the ability to accurately identify phonetically similar names. By making these adjustments, the new algorithm improves the effectiveness of searching in databases containing Arabic names, leading to more accurate and efficient retrieval results [4].

Researchers T. El-Diraby and L. Liao conducted a comparative study of several different phonetic algorithms, including improvements to Soundex, in the context of retrieving Arabic names. The study demonstrates how phonetic algorithms can be enhanced to better match the unique characteristics of the Arabic language. It provides an analysis of the effectiveness of these algorithms in identifying phonetically similar names, highlighting how modifications to Soundex and other algorithms can improve the accuracy of name retrieval in Arabic databases [5].

Pre-Pressing

The proposed system aims to improve results by modifying the Soundex algorithm through adding a preliminary processing step for Arabic names, followed by applying the traditional Soundex algorithm to the names [6].

The preliminary processing steps are:

1. Removing the first letter of the name if it starts with (ا, أ, آ, إ) and the following letter is not (ل), because we noticed they added more confusion.
2. Removing the first two letters if the name starts with (ا, أ, آ, إ) followed by the letter (ل).

Arabic Soundex Algorithm

1. **Retain the First Letter**:

- Keep the first letter of the remaining String after the initial processing.

2. **Replace Consonants with Digits:**
   - Replace certain consonants (excluding the first letter) with digits according to a specific mapping:
     - (ا, أ, إ, آ, ح, ع, غ, ش, و , ي , ه , ة) → 0
     - (ف, ب)→ 1
     - (خ, ج, ز, س, ص, ظ, ق, ك)→ 2
     - (ت, ث, د,ذ,ض,ط)→ 3
     - (ل) → 4
     - (م, ن)→ 5
     - (ر) → 6
3. **Collapse Adjacent Digits:**
   - Collapse adjacent identical digits into a single digit.
4. **Remove Zeros:**
   - Remove all zeros from the resulting string.
5. **Pad or Truncate:**
   - If the resulting string is shorter than 4 characters, pad it with zeros. If it's longer, truncate it to 4 characters.

The final result is a 4-character code representing the sound of the input name.

Experiments

To test the effectiveness of our proposed algorithm, we created a comprehensive dataset consisting of Arabic names. We began by collecting a large list of names, totaling 872 entries from this extensive collection, we filtered out and identified 73 unique individual names.

In addition to this, we expanded our dataset further by accessing and extracting data from various lists of graduates and students from Libyan universities. This effort aimed to ensure a diverse and representative sample of Arabic names. After combining these sources, our final dataset comprised a total of 123 unique Arabic names.

These names were selected to thoroughly evaluate the performance and accuracy of our algorithm. The diverse sources and the large number of names help in assessing how well the algorithm handles real-world data. This process is critical for verifying the algorithm's reliability and effectiveness in practical applications. Overall, the dataset provides a robust foundation for testing and refining our proposed solution.

The set of names was divided into 12 phonetic categories, with each category representing a specific phonetic class. This organization allows for a precise analysis of names based on phonetic similarity, aiding in the evaluation of the proposed system.

Evaluation of Existing Algorithms

We run our proposed algorithm using the test set. We calculate the average precision and recall. For each group, we compute recall and precision as follows:

$$\textbf{PRECISION} = \frac{TP}{TP + FP}$$

$$\textbf{RECALL} = \frac{TP}{TP + FN}$$

- **True Positives (TP):** The number of names correctly classified within the same phonetic category.
- **False Positives (FP):** The number of names incorrectly classified within the same phonetic category.
- **False Negatives (FN):** The number of names that should have been classified within the same phonetic category but were incorrectly classified into a different category.

Table 1: shows the results of the proposed system across different datasets.

Groups	True Positives (TP)	False Positives (FP)	False Negatives (FN)	Precision	Recall
G1	6	3	1	66.6%	85.7%
G2	8	4	0	66.6%	100%
G3	6	1	1	85.7%	85.7%
G4	11	2	0	84.6%	100%
G5	6	2	1	75.0%	85.7%
G6	8	3	1	72.7%	88.8%
G7	9	2	1	81.8%	90.0%
G8	6	2	1	75.0%	85.7%
G9	5	1	0	83.3%	100%
G10	10	2	1	83.3%	90.9%
G11	9	2	0	81.8%	100%
G12	7	0	1	100%	87.5%
Total	91	24	8	79.1%	91.9%

Precision reflects the accuracy of the system in classifying names into the specified phonetic categories. A precision of 79.1% means that there is a 21.9% error rate in the classifications made by the system.

Recall reflects the system's ability to identify all the correct names belonging to each category. A recall of 91.9% means that the system successfully found most of the correct names, although it may have missed some.

Conclusions

Through analyzing the performance of the Arabic name phonetic classification system based on the **Soundex** algorithm, we found that the system delivers notable results in classifying names into specific phonetic categories. The results of **Precision** and **Recall** across different classifications indicate that the system is capable of classifying names accurately and effectively.

The overall **Precision** of the system is 79.1%. This means that 79.1% of the names that the system classified positively were in the correct category. Although this percentage reflects good accuracy, there is room for improvement in reducing classification errors.

The overall **Recall** was 91.9%. This percentage reflects the system's high ability to identify names that should have been classified into the correct categories. This strong performance shows that the system is able to capture most of the correct names, although some names might have been missed.

Based on these results, we can conclude that the system performs excellently in terms of accurately identifying names, though there is a need to improve classification precision to reduce errors. Enhancing precision may require refining our proposed algorithm or integrating it with additional classification techniques to reduce False Positives.

Overall, these results provide a strong foundation to support the use of the system in phonetic name classification applications, emphasizing the importance of continuing to improve algorithms to achieve the best possible performance in the future.

References

1. Knuth, D. E. (1998). The art of computer programming, Volume 3: Sorting and searching. Addison-Wesley.
2. Yang, J., & Bhowmik, M. (2015). Data retrieval techniques in modern information systems. Journal of Computer and System Sciences, 81(5), 810-825.
3. An, A., & Gupta, S. (2009). A Soundex-based approach for matching similar names. Journal of Database Management, 20(3), 1-20.

4. Al-Khalifa, H. S., & Al-Muhtaseb, M. (2016). A new phonetic algorithm for Arabic name matching. International Journal of Computer Applications, 138(1), 12-18.
5. El-Diraby, T., & Liao, L. (2014). Comparative study of phonetic algorithms for Arabic name retrieval. International Journal of Computer Applications, 90(12), 1-7.
6. El-Bakry, R. M., Al-Mansoori, S. A., & Al-Salemi, R. S. (2015). A comparative study of Arabic name normalization methods. Journal of King Saud University - Computer and Information Sciences.
7. Abed, S. A., & Awwad, Z. M. (2016). A phonetic matching algorithm for Arabic names. Journal of Computer and Communications, 4(9), 54-63.