Journal of Basic Sciences (JBS)
مجلة العلوم الأساسية
Volume 37, Issue 2, July - December 2024
Page no.: 325- 334
2024: (1.44)
Email: jbs@journals.asmarya.edu.ly Website: journals.asmarya.edu.ly/jbs

Adapting Soundex for Arabic Names with Preprocessing for Better Accuracy

Mohammed F. Al Boashi *, Saad A. Al Deeb

Computer Science Dept., Faculty of Sciences, Elmergib University, Khoms - Libya

*** Corresponding author: mfalboashi@elmergib.edu.ly**

Received: Nov. 22, 2024	Accepted: Dec. 06, 2024	Published: Dec. 08, 2024

Abstract:
The Soundex algorithm plays an important role in retrieving names based on phonetic similarity, as it converts names into numerical representations that facilitate comparison based on pronunciation rather than exact spelling. This algorithm helps identify names that sound similar despite variations in their spelling, improving the accuracy of searches in databases. Although initially used for languages like English, it has been adapted to fit the characteristics of other languages, such as Arabic. For Arabic names, the algorithm is enhanced to handle variations in pronunciation and spelling, further improving retrieval effectiveness. In this way, the Soundex algorithm provides an efficient method for classifying and retrieving names based on their phonetic sound rather than relying solely on exact spelling. We developed a system based on preliminary processing of Arabic names to improve precision and recall, taking into account the unique characteristics of the Arabic language. We compiled a list of Arabic names, which were then divided into groups based on their phonetic pronunciation. The proposed system demonstrated effectiveness and accuracy in retrieving and classifying names. Overall, the system showed good performance, with a need for improvements in certain specific cases.

Keywords: Soundex algorithm, Natural Language Processing, Arabic names, Phonetic similarity, Name retrieval.

تكييف Soundex لأسماء عربية مع المعالجة المبدئية لتحسين الدقة

محمد فؤاد البوعيشى*، سعد اعمار الذيب

قسم علوم الحاسوب، كلية الآداب والعلوم قصر الاخيار، جامعة المرقب، الخمس، ليبيا

*** لمراسلة المؤلف: mfalboashi@elmergib.edu.ly**

استلمت: 22 نوفمبر 2024	قبلت: 06 ديسمبر 2024	نشرت: ----- ديسمبر 2024

الملخص

تلعب خوارزمية Soundex دورًا مهمًا في استرجاع الأسماء بناءً على التشابه الصوتي، حيث تقوم بتحويل الأسماء إلى تمثيلات عددية تسهّل المقارنة بناءً على النطق بدلاً من التهجئة الدقيقة. تساعد هذه الخوارزمية في التعرف على الأسماء التي تبدو مشابهة رغم اختلاف تهجئتها، مما يحسن دقة البحث في قواعد البيانات. وعلى الرغم من استخدامها في البداية للغات مثل الإنجليزية، تم تعديلها لتناسب خصائص لغات أخرى مثل اللغة العربية. بالنسبة للأسماء العربية، تم تحسين الخوارزمية للتعامل مع التباين في النطق والتهجئة، مما يعزز فعالية الاسترجاع. بهذه الطريقة، توفر خوارزمية Soundex وسيلة فعّالة لتصنيف واسترجاع الأسماء بناءً على صوتها بدلاً من الاعتماد فقط على التهجئة الدقيقة. قمنا بتطوير نظام يعتمد على المعالجة الأولية لأسماء عربية لتحسين الدقة والاسترجاع، مع الأخذ في الاعتبار الخصائص الفريدة للغة العربية. قمنا بتجميع قائمة بالأسماء العربية، التي تم تقسيمها إلى مجموعات بناءً على نطقها الصوتي. أظهر النظام المقترح فعالية ودقة في استرجاع وتصنيف الأسماء. بشكل عام، أظهر النظام أداء جيدًا، مع الحاجة إلى تحسينات في بعض الحالات المحددة.

الكلمات المفتاحية: خوارزميةSoundex ، معالجة اللغة الطبيعية، الأسماء العربية، التشابه الصوتي، استرجاع الأسماء.

1. Introduction

The Soundex algorithm is one of the most prominent tools used for retrieving names based on phonetic similarity. It works by converting names into numerical representations that make it easier to compare them based on pronunciation rather than exact spelling. Developed in the early 20th century, this algorithm was designed to address the need to classify names that may differ in spelling but share similar sounds. It has proven to be highly effective in various applications such as databases, government records, and medical information systems, where it helps improve search and retrieval accuracy.

The Soundex algorithm reduces the gap between names that may be spelled differently but sound alike, making it easier to retrieve the correct or similar names when inaccurate data is entered. Although originally developed for the English language, there have been numerous efforts to adapt it for other languages, such as Arabic, which has different phonetic and orthographic characteristics [2].

By improving algorithms for name retrieval, systems can enhance the accuracy of search processes and reduce errors arising from mismatched spelling or pronunciation. For this reason, the Soundex algorithm is a vital tool in many applications requiring accurate data retrieval in multilingual environments.

2. Soundex function

The Soundex algorithm is a phonetic algorithm designed to convert words into strings of characters that approximately represent their pronunciation. Developed in 1918 by Robert Russell, its primary purpose was to facilitate searching databases for words or names that may sound similar but are spelled differently.

The Soundex algorithm is based on a set of rules that define how words are transformed into character strings. These rules are grounded in English phonetics and spelling, where each letter in the word is represented by a specific sound key. These rules are applied to generate a standardized character string that reflects the basic sound of the word [1][7].

The Soundex algorithm is widely used in databases to compare words that may be phonetically similar but spelled differently. This improvement helps researchers locate matching records, particularly when the input data is incomplete or contains spelling errors.

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

3. Arabic Soundex

The Soundex algorithm, originally developed for the English language, is primarily used to identify words that sound similar despite being spelled differently. However, the Arabic language presents unique phonetic and orthographic characteristics that require modifications to the Soundex algorithm to make it suitable for Arabic. In Arabic, sounds vary based on the letters and phonetic similarities between them, along with the effects that may occur when letters are added or omitted from names or words [4][5].

Applying the Soundex algorithm to Arabic names requires adjustments to the rules to align with the specific phonetics of the language. For example, letters such as "غ" "ع," and "ق" may sound similar but are spelled differently, highlighting the need to adapt the algorithm to account for these phonetic variations. Additionally, similar sounds in Arabic may be written in multiple ways, such as "س" and "ص" or "ج" and "ش," which makes adjusting the algorithm crucial for improving accuracy in retrieval.

By adapting the Soundex algorithm to better fit the Arabic language, search results in databases containing Arabic names or words—especially those that may be spelled differently but sound alike—can be improved. This enhancement allows systems to more accurately identify phonetically similar names and improves data retrieval even in cases of misspellings or inaccuracies in writing.

4. Study Objectives:

1. Developing a Modified Version of the Soundex Algorithm
2. Applying the Soundex Algorithm to a Sample of Arabic Names and Evaluating Its Effectiveness in Search and Matching Operations

5. Study Methodology

The study methodology involves a thorough review of previous research related to the Soundex algorithm, along with an in-depth analysis of existing studies in this field. Additionally, the algorithm is applied to a set of Arabic names, and the results are analyzed and evaluated for their effectiveness in search and matching operations.

6. 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].

7. Proposed System

A new system for the Soundex algorithm has been proposed, which includes preprocessing of Arabic names, where the algorithm was improved to match the phonetic and orthographic characteristics of the Arabic language. The

algorithm was applied after the preprocessing step to classify and retrieve Arabic names based on their phonetic similarity.

7.1 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).
3. Remove the first letter of the name if it starts with (م) and the following letter is (ح or س or ك).
4. If the name ends with "ين" ,"ان", or "ون", remove these suffixes from the name.

7.2 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.

8. Experiments

We compiled a comprehensive dataset containing Arabic names, starting with a list of 1121 names. This list was then filtered to select 204 unique and distinct names. The names were divided into 16 phonetic categories, with each category representing a group of names with similar phonetic characteristics. This classification helps in conducting a precise analysis based on phonetic similarities, thereby enhancing the evaluation of the proposed system.

These names were carefully selected to assess the algorithm's performance with accuracy and effectiveness. The diversity of sources and the large number of names reflect the algorithm's ability to handle real-world data, contributing to testing its reliability and efficiency in practical applications.

9. 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:

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

• **True Positives (TP):** The number of names correctly classified within the same phonetic category.

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

- **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	10	2	0	83.3%	100%
G2	12	4	1	75%	92.3%
G3	14	1	2	93.3%	87.5%
G4	16	2	0	88.9%	100%
G5	11	2	1	84.6%	91.7%
G6	10	0	0	100%	100%
G7	12	2	1	85.7%	92.3%
G8	10	0	1	100%	90.9%
G9	8	1	0	88.9%	100%
G10	11	1	1	91.7%	91.7%
G11	9	2	0	81.8%	100%
G12	7	1	0	87.5%	100%
G13	13	1	1	92.9%	92.9%
G14	12	0	2	100%	85.7%
G15	6	0	0	100%	100%

G16	12	1	1	92.3%	92.3%
Total	173	20	11	89.6%	94%

Precision reflects the accuracy of the system in classifying names into the specified phonetic categories. A precision of 89.6% means that there is a 10.4% 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 94% means that the system successfully found most of the correct names, although it may have missed some.

10. Conclusions

The proposed algorithm showed good performance across different groups, achieving a high recall of 94% overall, indicating its strong ability to identify positive cases. The precision was 89.6%, which is a good indication that most of the positive predictions were correct. However, some groups suffered from false positives (FP) that affected the precision, while false negatives (FN) were minimal. Precision could be improved in some groups with a higher FP rate. Overall, the algorithm provides strong results but needs improvements in certain specific cases.

References

[1] D. E. Knuth, "The Art of Computer Programming," Volume 3: Sorting and Searching, Addison-Wesley, 1998.

[2] J. Yang and M. Bhowmik, "Data Retrieval Techniques in Modern Information Systems," *Journal of Computer and System Sciences*, vol. 81, no. 5, pp. 810-825, 2015.

[3] A. An and S. Gupta, "A Soundex-based Approach for Matching Similar Names," *Journal of Database Management*, vol. 20, no. 3, pp. 1-20, 2009.

[4] H. S. Al-Khalifa and M. Al-Muhtaseb, "A New Phonetic Algorithm for Arabic Name Matching," *International Journal of Computer Applications*, vol. 138, no. 1, pp. 12-18, 2016.

[5] T. El-Diraby and L. Liao, "Comparative Study of Phonetic Algorithms for Arabic Name Retrieval," *International Journal of Computer Applications*, vol. 90, no. 12, pp. 1-7, 2014.

[6] R. M. El-Bakry, S. A. Al-Mansoori, and R. S. Al-Salemi, "A Comparative Study of Arabic Name Normalization Methods," *Journal of King Saud University - Computer and Information Sciences*, 2015.

[7] S. A. Abed and Z. M. Awwad, "A Phonetic Matching Algorithm for Arabic Names," *Journal of Computer and Communications*, vol. 4, no. 9, pp. 54-63, 2016.