Application of phonetic encoding for analyzing.pdf

Shahidul Islam Khan, Bangladesh University of Engineering and Technology
Application of phonetic encoding for analyzing similarity of Patient’s Data: Bangladesh Perspective

Abir Bin Ayub Khan, Mohammad Sheikh Ghazanfar, Shahidul Islam Khan

Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh

Abstract— Due to illiteracy and lack of standardized healthcare systems, patients in Bangladesh usually, provide misspelled names while making their entry via these systems. Different records with slightly misspelled names are thus generated which makes data mining and others tasks quite challenging and inefficient. In this paper, we have looked into the underlying problem of misspelled names of patients in healthcare systems and proposed a modified version of NameSignificance algorithm. Our proposed algorithm has performed significantly better than the existing solutions like NameSignificance, Modified Soundex, Double Metaphone encoding for Bangla as we founded our algorithm on the phonetic nature of Bengali names written in English translated form. Our algorithm achieved a staggering 77% matching of names whereas relevant algorithm could not pass 70% correct matches. This proposed method could pave the way for better record linkage and data analysis of the medical patient dataset. Our syllable based approach also helped us identify the reason behind the wrong matches appeared through our algorithm which will surely pave the way to purify our algorithm in the future.

Keywords—Record Linkage; phonetic encoding; Name; Bangladesh; Health Data;

I. INTRODUCTION

Generally, unique entity identifiers or keys are not available in all the databases that are linked hence making record linkage a daunting task. So, to do linkage the familiar attributes need to be used. For databases which consist of personal identifying information, these familiar attributes mostly include names, birth date, addresses and other detail.[1-3]

Using names of users in the records of the dataset plays vital role in record linkage. At the same time, errors in names results in faulty outcomes from the dataset. As for health and other public datasets where records contain patient names which is manually inputted it is quite common to have misspelled names in the system.

There are phonetic encoding algorithm which can be found in existing literatures. All these algorithms could be divided into two types with respect to language preferences. Algorithms that are efficient for English/American names explicitly and algorithms that are good for Bengali names. For the first type we have Metaphone, NYSIIS, and Soundex etc. and for the later there have been NameSignificance, Modified Soundex, Double Metaphone encoding for Bangla etc. Among the Bengali versions only NameSignificance have been properly implemented on Medical patient dataset. Others were just theoretical analysis.

In this paper, we have discussed about our proposed algorithm that can check the resemblance among names in Bengali. By experimenting our algorithm on noisy dataset of 65,000 patient records we measured a correct matching of 77% records on the basis of their names which was either misspelled or error-prone.

II. MOTIVATION

In 2009, with the assistance of German Donor GIZ, Bangladesh government initiated the development of National Health Data Warehouse (NHDW). At present, there is exists a gap between the different available digital health record. The inclusion of a data warehouse would bridge this gap make them interoperable. As of now, two open source softwares: OpenMRS and DHIS2 are being utilized to collect medical data from different healthcare organizations under Directorate General of Health Services (DGHS) of Bangladesh Government.[4-6]

Analysis of some real-life scenarios also worked as our motivation to solve this problem.

A. Scenario 1: (Illeterate patient)

According to UNESCO, Bangladesh has literacy rate of only 61.5%. So, a huge majority of people who come for healthcare in rural or developed healthcare centers are presumably illiterate who cannot even pronounce their names correctly let alone writing them.

B. Scenario 2: (Patient cycle)

Various cases emerge in the context of Bangladesh based on the patient cycles discussed in. For treatment or diagnosis, a patient may visit a clinic, hospital, diagnostic center, or private chamber of a doctor. It is very natural for a patient in Bangladesh to have numerous reports scattered in different healthcare organizations.[7]

Let SHL be the sum of all the health records for a single patient in his/her entire life.
SHL can be calculated as follows:

Let us assume,

Life period of a patient = t years
On average visits for health care facility per year = y
Total visits for health care in the whole life period,

\[ T = ty \]

Average Life expectancy in Bangladesh: Female-72 years and Male-70 years.[8]

We can assume the average life expectancy as 71 years form male and female all together.

So, considering \( t = 71 \) years, and \( y = 15 \) /year

\[ T = 15 \times 71 = 1065 \] times

If one visit produces at least one record, SHL is greater One Thousand for a single patient (1065 to be exact). And if the patient gets more than one records per visit then the count goes even higher.

III. PROBLEM DEFINITION

Bangladesh as a developing country comprises of a huge percentage of population who are illiterate and do not know their full name let alone pronouncing their name correctly even in their native language. Its healthcare system is not fully automated and needs manual entry of patient records by data entry officer in different healthcare centers. All of these facts give rise to a problem where various records are generated for a single patient with slightly misspelled names in each of the record. The problem can be understood from the table below.

Slightly misspelled names of a patient generate different records although all these records are actually of the same patient. To carry out data mining and knowledge discovery tasks on datasets containing this type of records results in faulty outcomes which is not at all useful.

<table>
<thead>
<tr>
<th>Original Patient Name</th>
<th>Misspelled Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abrar Rashid Rafi</td>
<td>Mr. Abrar Rashid Rafi</td>
</tr>
<tr>
<td></td>
<td>Md. Abrar Rashed Rafi</td>
</tr>
<tr>
<td></td>
<td>Abraar Rashid Rafi</td>
</tr>
<tr>
<td></td>
<td>Captain Abrar Rashid Rafi</td>
</tr>
<tr>
<td></td>
<td>Abrar Rashid Rafee</td>
</tr>
<tr>
<td></td>
<td>Abrar Rasid Rafi</td>
</tr>
</tbody>
</table>

IV. RELATED ALGORITHM

Several algorithms have been proposed and implemented in relation to phonetic matching of names. All these algorithms have been developed for different purpose and in different context. We will discuss the major algorithms according to their context below.

A. English language:

- **Metaphone algorithm:** A rough approximation of how a word in English sounds is returned in this algorithm published by Lawrence Philips. The returning result should be the same for words or names that have similar sounds. Wide usage of Metaphone algorithm can be found in Spell Checkers, Search Interfaces, De-duping programs and Genealogy Sites. Double Metaphone and Metaphone 3 are some of the many variations of this algorithm.[9]

- **Match rating Approach:** Whether two names are pronounced similarly can be checked with the help of Match rating approach. It first encodes the given string using match rating codex then compare it.[10]

- **NYSIIS algorithm:** NYSIIS keep tracks of the vowels’ placement in the name that has been given as input. It does that by converting all the vowels to the character ‘A’. This is a fundamental difference between NYSIIS and Soundex. A pure alphabetic code is returned by this algorithm. [11]

B. Bengali language unrelated to healthcare:

- **Modified Soundex algorithm:** Zaman and Khan proposed a Bangla phonetic encoding for better spelling suggestion. In essence it is based on Soundex algorithm and has been modified for Bengali language. [12]

- **Double Metaphone encoding for Bangla:** Another work for misspelled words in Bangla was done by Zaman and khan with the application of Double Metaphone algorithm in Bengali language context.[13]

Both these papers are actually intended for better spell checker word suggestions. There has not been shown any implementation of the algorithms and performance analysis on real world dataset.

C. Bengali names written in English in healthcare:

- **NameSig algorithm:** In the context of Bengali names which are written in English especially in the field of healthcare there has been only one paper so far. In this paper the authors proposed an algorithm called NameSig and applied it to achieve similarity analysis of patients’ data.[14]

V. PROPOSED ALGORITHM: MODIFIED NAMESIGNIFICANCE

Modified NameSignificance algorithm produces a masked string. It finds out the insignificant portions of a name such as salutations and titles and removes them. It considers the remaining portion as significant and applies the syllable division and word mapping on it. Syllable division divides the significant portion of the name into different syllables by treating the letters in the name as English representative of Bengali alphabets. Each syllable is then transformed to a simplified version of the Bengali alphabets
We used two tables for Bengali consonants and vowels in order to map individual letters.

To understand how our algorithm works on a real name of a patient, Figure 1 is used where all the steps of our algorithm have been graphically represented.

In Figure 1, we can see that for a patient named “Mr. Kamrool Islam” which has been given as input in our algorithm generates an output of “kmrlslm” as the Modified NameSignificance code of that name. During this process we could see how the name has been put through modification in different steps of the algorithm.

Our proposed Modified NameSignificance algorithm takes the name of the patient as input. It then goes on covering all the steps to generate a modified NameSignificance code of the given name as output.

**Input:**
Name of a patient

**Output:**
Modified NameSignificance code

**Steps:**
1. Remove Salutation and Title
2. Remove any repetitive vowel and consonant in the name
3. Split the name into syllables based on English representative letters of Bengali vowels and consonants
4. Split the syllables into Bengali vowels and consonants
5. Apply simplified mapping of the individual letters

VI. DATASET

To validate our proposed algorithm, we applied it with two types of datasets. One is a real-world dataset comprising of 65,000 patient records. The second dataset we used is a synthetic dataset consisting of 450 patient records which has been collected by an independent online survey. Each of the record in both the dataset uses five identifying features to identify a patient in a record linkage system. The features are Name of Patient, Age, Gender, Contact Number and Address. We have mainly focused on Name of the Patient feature to correctly identify a patient from different records.

VII. RESULTS AND DISCUSSION

We have implemented the algorithms in Python. The machine that have been used for the experiment was a Windows machine running 64-bit Windows 10 operating system on an Intel® Core™ i7-5500U CPU @ 2.40 GHz. The machine has an 8.00 GB RAM capacity and 1 TB hard-disk space.

Breaking down the names of the patients into syllables is a unique and ground-breaking characteristic of Modified NameSignificance algorithm. We have tested our algorithm on the real-world dataset which contains a lot of errors and found out that our algorithm got 77.7% right matches for the misspelled names in different records.

Then we moved on to find out the reason behind our wrong matches. Fortunately, we have successfully figured out that among the 22.3% wrong matches we could predict 11.7% records which could not be matched with the help of our syllable division feature of the algorithm. We have matched the syllable counts of the actual name and wrongly matched duplicate names. This gave us the desired insight that we were looking for.
Then we applied the NameSignificance algorithm on the same real-world dataset. This algorithm gave a 62.2% correct match and 37.8% wrong match for the noisy dataset. In the paper also, the authors commented that NameSignificance had issues dealing with noise in the dataset.

In Table II, we have shown the comparison of Name Value, Match Rating Codex, NYSIIS and Metaphone algorithm with our algorithm. It is clear that our algorithm performs better than all these algorithms for the natural dataset of patients where we could match the output with the original name of the patients. In regards to time efficiency our algorithm is behind of NYSIIS only which is not a major issue when it comes to the quality of the output.

By applying Modified NameSignificance on the small dataset gave us more justifying results. It contained around 500 patient records. There were 87 actual patients and we were able to map all the records to their respective original names. The repetition of each name in the dataset could be seen from Figure 3.

TABLE II. SEVERAL RECORDS WITH MISSPELLED NAMES OF SAME PATIENT

<table>
<thead>
<tr>
<th>Name of Algorithm</th>
<th>Right matches (%)</th>
<th>Running time(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified NameSignificance</td>
<td>77.7</td>
<td>0.85</td>
</tr>
<tr>
<td>Name Value</td>
<td>62.2</td>
<td>0.91</td>
</tr>
<tr>
<td>Match Rating Codex</td>
<td>62.80</td>
<td>1.03</td>
</tr>
<tr>
<td>NYSIIS</td>
<td>51.69</td>
<td>0.77</td>
</tr>
<tr>
<td>Metaphone</td>
<td>49.76</td>
<td>1.01</td>
</tr>
</tbody>
</table>

VIII. CONCLUSION

Different records with slightly misspelled names are generated because of misspelled names which makes data mining and others tasks quite daunting. In this paper, we have looked into the underlying problem of misspelled names of patients in healthcare systems and proposed a modified version of NameSignificance algorithm. Our proposed algorithm has performed significantly better than the existing solutions as we founded our algorithm on the phonetic nature of Bengali names written in English transliterated form. Our algorithm achieved a staggering 77% matching of names whereas relevant algorithm could not pass 70% correct matches. Our syllable based approach also helped us identify the reason behind the wrong matches appeared through our algorithm which will surely pave the way to purify our algorithm in the future.

REFERENCES