International Journal of Trend in Research and Development, Volume 3(2), ISSN: 2394-9333
www.ijtrd.com

Data Deduplication of Indian Names

[1]Priyanka Rathod, [2]Vaishnavi Balasubramanian Konar,[3]Shivani Wankhade, [4]Ruchita Reengesia,
[5]Mr.Sandesh Chavan and [6]Ms.Pramila Shinde,
[1,2,3,4]Department of Information Technology, Shah And Anchor Kutchhi Engineering College, Mumbai, India
[5]Vice President and Head, Product Engineering, Mindcraft Software Pvt. Ltd.
[6]Asst. Professor, Department of Information Technology, Shah And Anchor Kutchhi Engineering College, Mumbai, India

***Abstract*—** Demonstrating an efficient algorithm to deduplicate information which contains Indian names (which are pronounced and spelled differently) and address strings. The algorithm consists of 2 stages -enrollment and deduplication .In both stages name strings and address strings are reduced to generic name and address strings with the help of phonetic based reduction rules. Thus there may be several name strings having same generic name with all the name strings and address forms a bin. At the enrolment stage a database which is an array of bins is efficiently created and each bin is a singly linked list. At the de-duplication stage name strings and address strings are reduced and used to determine the top 'k' best matches.

Inorder to see the performances of the algorithm we can consider any source of data (eg. Bank Applications, Restaurants). It has been observed that phonetic reduction rules could reduce the name strings and address strings by more than 90% .

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage .This technique is used to improve storage utilization and can also be applied to network data transfers. In this process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk.

Keywords: *De-duplicate, Redundancy, Indian, Phonetic*

I. INTRODUCTION

De (deletion) + Duplication = Deletion of duplicate data Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data.[1] It removes duplicate files within and across a file. In most organizations, the storage systems contain duplicate copies of data. For example, the same file may be saved in several different places by different users. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Deduplication takes place on the file level; that is, it eliminates duplicate copies of the same file. This kind of deduplication is sometimes called file-level deduplication or single instance storage (SIS). Deduplication can also take place on the block level, eliminating duplicated blocks of data that occur in non-identical files. Block-level deduplication frees up more space than SIS.

Previously Data deduplication has been performed using Machine learning approaches. [5][6][3] In computing, data de-duplication is a process to eliminate redundant data to improve storage utilization. In the de-duplication process, duplicate data is deleted, leaving only one copy of the data to be stored, along with references to the unique copy of data. In this paper, by demographic data of an individual, we mean a record consisting of two string fields viz. Given Name and Surname of the individual. Demographic de-duplication process determines whether there exists any set of records in the demographic database which are matched with the record of a query data within some tolerable range. For any new demographic data, a negative background search is performed to obtain all its close matches. To accomplish this task, every string (i.e. first name, and last name of the query individual) is considered to be found in the database if any of the following criteria satisfies.

1. Identical strings in the database.
2. String having similar phonetics but difference with alphabets.
3. String which can be constructed from new string by few transformations like insertion, deletion or substitution. [10]

In this paper we are considering Indian Names and their sounds in order to reduce the duplicated data.

II. ALGORITHMS

A. Soundex

Also known as a distance algorithm based on phonetic similarity metrics, Soundex was developed by Odell/Russel in 1918 and is capable of telling if a pronunciation of two strings are alike just by using its consonants, or better, accomplishing the coding by means of its consonants. [2] Algorithm is as follows:

a. Retain the first letter of the word.
b. Capitalize all letters in the word and drop all punctuation marks. Pad the word with rightmost blanks as needed during each procedure step.
c. Change all occurrence of the following letters to '0' (zero):
'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
d. Change letters from the following sets into the digit given:
1 = 'B', 'F', 'P', 'V'
2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'
3 = 'D','T'
4 = 'L'
5 = 'M','N'
6 = 'R'
e. Remove all pairs of digits which occur beside each other from the string that resulted after step (4).
f. Remove all zeros from the string that results from step 5.0 (placed there in step 3)
g. Pad the string that resulted from step (6) with trailing zeros and return only the first four positions, which will be of the form <uppercase letter> <digit> <digit> <digit>.

B. Metaphone

An algorithm for encoding a word so that similar sounding words encode the same is called as metaphone. It's similar to soundex in purpose, but as it knows the basic rules of English pronunciation it's more accurate.[7] The Metaphone algorithm is significantly more complicated than the others because it includes special rules for handling spelling inconsistencies and for looking at combinations of consonants in addition to some vowels. An updated version of the algorithm, called Double Metaphone, goes even further by adding rules for handling some spellings and pronunciations from languages other than English. Metaphone codes are particularly useful where spelling discrepancies may occur in words that sound the same, for example, where information has been captured over the telephone. By considering the pronunciation of the string instead of the exact string value, many minor variances can be overcome. A Metaphone code is therefore a good alternative to the raw data value when performing a duplicate check, making it is easier to identify possible duplicate or 'equivalent' values.

The processor allows you to specify the maximum length of the Metaphone code (up to a maximum of 12 characters) so that it can be focused solely on the first few syllables or words of complex data rather than the entire column, and so that you can control the sensitivity of the phonetic similarity between values.[8]

C. Double Metaphone

An updated version of the metaphone algorithm, called Double Metaphone, goes even further by adding rules for handling some spellings and pronunciations from languages other than English. The name comes from the fact that it produces two encodings for each name. So it doesn't have the robustness of Daitch-Mokotoff which can have many encodings, but it is certainly more robust than the earlier systems that have only one encoding per name. Another new feature of double metaphone is that it includes foreign pronunciations, but it lumps all the foreign rules together and doesn't distinguish which rule corresponds to which language. And Phillips dropped one of the improvements of his earlier Metaphone – namely the encoding is again restricted to the initial part of the name. [8]

D. Daitch Mokotoff

D-M Soundex was developed by genealogist Gary Mokotoff and later improved by genealogist Randy Daitch because of the problems they encountered while trying to apply the Russell Soundex to Jews with Germanic or Slavic surnames (such as Moskowitz vs. Moskovitz or Levine vs.Lewin). Results of D-M Soundex are returned in a numeric format between 100000 and 999999. This algorithm is much more complex than Soundex. [4]

The Daitch-Mokotoff Soundex Coding Chart

Letter	Alternate Spelling	Start of a name	Before a vowel	Any other situation
NC = not coded				
AI	AJ, AY	0	1	NC
AU		0	7	NC
Ą	(Polish a-ogonek)	NC	NC	6 or NC
A		0	NC	NC
B		7	7	7
CHS		5	54	54
CH	Try KH (5) and TCH (4)			
CK	Try K (5) and TSK (45)			
CZ	CS, CSZ, CZS	4	4	4
C	Try K (5) and TZ (4)			
DRZ	DRS	4	4	4
DS	DSH, DSZ	4	4	4
DZ	DZH, DZS	4	4	4
D	DT	3	3	3
EI	EJ, EY	0	1	NC
EU		1	1	NC
Ę	(Polish e-ogonek)	NC	NC	6 or NC
E		0	NC	NC
FB		7	7	7
F		7	7	7
G		5	5	5
H		5	5	NC
IA	IE, IO, IU	1	NC	NC
I		0	NC	NC
J	Try Y (1) and DZH (4)			
KS		5	54	54
KH		5	5	5
K		5	5	5
L		8	8	8
MN			66	66
M		6	6	6
NM			66	66
N		6	6	6
OI	OJ, OY	0	1	NC
O		0	NC	NC
P	PF, PH	7	7	7
Q		5	5	5
RZ, RS	Try RTZ (94) and ZH (4)			
R		9	9	9
SCHTSCH	SCHTSH, SCHTCH	2	4	4

Letter	Alternate Spelling	Start of a name	Before a vowel	Any other situation
NC = not coded				
SCH		4	4	4
SHTCH	SHCH, SHTSH	2	4	4
SHT	SCHT, SCHD	2	43	43
SH		4	4	4
STCH	STSCH, SC	2	4	4
STRZ	STRS, STSH	2	4	4
ST		2	43	43
SZCZ	SZCS	2	4	4
SZT	SHD, SZD, SD	2	43	43
SZ		4	4	4
S		4	4	4
TCH	TTCH, TTSCH	4	4	4
TH		3	3	3
TRZ	TRS	4	4	4
TSCH	TSH	4	4	4
TS	TTS, TTSZ, TC	4	4	4
TZ	TTZ, TZS, TSZ	4	4	4
Ţ	(Romanian t-cedilla)	3 or 4	3 or 4	3 or 4
T		3	3	3
UI	UJ, UY	0	1	NC
U	UE	0	NC	NC
V		7	7	7
W		7	7	7
X		5	54	54
Y		1	NC	NC
ZDZ	ZDZH, ZHDZH	2	4	4
ZD	ZHD	2	43	43
ZH	ZS, ZSCH, ZSH	4	4	4
Z		4	4	4
Letter	Alternate Spelling	Start of a name	Before a vowel	Any other situation

E. Dynamic Programming

Sometimes, divide-and-conquer leads to overlapping subproblems and thus to redundant computations. It is not uncommon that the redundancies accumulate and cause an exponential amount of wasted time. We can avoid the waste using a simple idea:
solve each sub-problem only once.To be able to do that, we have to add a certain amount of book-keeping to remember sub-problems we have already solved. The technical name for this design paradigm is dynamic programming.

Edit distance:

We illustrate dynamic programming using the edit distance problem. Assume a finite set of characters or letters,Σ, which we refer to as the alphabet, and we consider strings or words formed by concatenating finitely many characters from the alphabet. The edit distance between two words is the minimum number of letter insertions, letter deletions, and letter substitutions required to transform one word to the other. For example, the edit distance between FOOD and MONEY is at most four:

FOOD→MOOD→MOND→MONED→MONEY

A better way to display the editing process is the gap representation that places the words one above the other, with a gap in the first word for every insertion and a gap in the second word for every deletion:
F O O - D
M O N E Y

The structure of dynamic programming is again similar to divide-and-conquer, except that the sub-problems to be solved overlap. As a consequence, we get different recursive paths to the same sub-problems. To develop a dynamic programming algorithm that avoids redundant solutions, we generally proceed in two steps:

- We formulate the problem recursively. In other words, we write down the answer to the whole problem as a combination of the answers to smaller sub problems.
- We build solutions from bottom up. Starting with thebase cases, we work our way up to the final solution and (usually) store intermediate solutions in a table.

For dynamic programming to be effective, we need a structure that leads to at most some polynomial number of different subproblems. Most commonly, we deal with sequences, which have linearly many prefixes and suffixes and quadratically many contiguous substrings.[9]

III. THE PROPOSED ALGORITHM FOR DATA DEDUPLICATION

Central Database and Input File:

- Input will be taken from a database which will be provided by the organization
- The data from the database will then be converted into .csv file

Load in Memory:

- The data thus obtained will be loaded into the memory.
- This is done to enhance the speed of the process.

Create Sets Segregation of data Based on the topography of the names:

- North Indian

- South Indian

For each of these topographies there will be different algorithms as each place needs different algorithm for reduction .

Reduce and Identify close match:

- Various algorithms will be applied to reduce the given data.
- An algorithm to reducing the data on the form of parent string having all the child strings under it and identifying the closest match which are then clubbed together using a linked list.

Display Duplicates:

- All the duplicates of the given name will be displayed on request.

Write to Database:

- Any changes such as insertion of a new name to the parent string, a new parent string will be recorded back to the database.

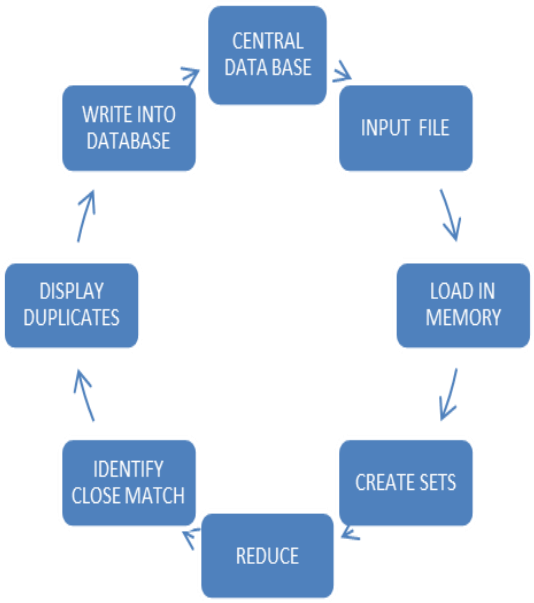

Figure 1: Lifecycle of the proposed algorithm

CONCLUSION

This paper has proposed an algorithm for de-duplication of Indian names.It becomes a daunting task since Indian names are pronounced and spelled differently in different regions. Simple phonetic reduction and exact matching fails to identify potential duplicates.Hence in this paper each name string is reduced to a generic name string which is put in a stack and linked to it's child names by a linked list. Finally, name string of a query data is searched in the linked list to compute the top k best matches.

References

[1] The Insider's Guide to Data Deduplication: Volume 1 Paperback – Import, 29 Oct 2010 by Larry

[2] Multilingual Names Database Searching Enhancement Ivo R. Draganov, Antoaneta A. Popova, and Lubomir L. Ivanov+

[3] Sarawagi, S., Bhamidipaty, A.: Interactive deduplication using active learning. In: Proceedings of the eighth International Conference on Knowledge discovery and data mining(ACM SIGKDD). pp. 269–278. ACM (2002)

[4] G. Mokotoff, Soundexing and Genealogy http://www.avotaynu.com/soundex.html

[5] Winkler, W.: Matching and record linkage. Wiley Online Library (1993) .

[6] Winkler, W.: The state of record linkage and current research problems. In: Statistical Research Division, US Census Bureau. Citeseer (1999)

[7] http://dictionary.reference.com/browse/metaphone

[8] http://stevemorse.org/phonetics/bmpm2.htm

[9] https://www.cs.duke.edu/courses/fall08/cps230/Lectures/L-04.pdf

[10] An Efficient Algorithm for De-duplication of Demographic Data Vandana Dixit Kaushik, Amit Bendale, Aditya Nigam, Phalguni Gupta