EFFICIENT NAME SEARCH ALGORITHM OF PAKISTANI NAMES

Submitted By
Nadia Yousaf
ID: 13021050029
Supervisor
Dr. Tahir Ejaz
School of Science and Technology
University of Management and Technology
Lahore, Pakistan
August, 2015 ii

EFFICIENT NAME SEARCH ALGORITHM OF PAKISTANI NAMES

By

Nadia Yousaf (ID: 13021050029)

A thesis submitted to the faculty of the Department of Computer Science for the degree of MS in Computer Science

Thesis Committee:

Dr. Tahir Ejaz, Chairman _____

Dr. Adnan Abid _____

Mr. Mirza Mubashar Baig _____

School of Science and Technology

University of Management and Technology

Lahore, Pakistan

August, 2015 iii

Abstract

In many applications the crucial and vital role is played by name matching. In every profession the information is retrieved and data is stored in repositories in English. This data can be the names of persons working there or any other type of data. Many algorithms have been developed to match the names because names in every application create unavoidable variations and errors. Spelling, pattern or phonetic modifications are name variations that are considered to develop many algorithms. English language is covered in mostly existing techniques. Pakistan's official language is English and mother language is Urdu this is the reason that all government documents, data storage, business and professional activities use English. Due to it storage of Urdu names in English is mandatory. Matching names such as Pakistani names against names stored in computer databases or files (written in Roman Urdu), can create the large variety of possible spelling variations. For example, the Muslim Pakistani name *"Mohamed"* can be represented as *"Mohammed," "Muhhamad," "Muhamud,"* etc. More sophisticated techniques are required to accommodate the large possible variations in spellings. In this research different methods for string matching are discussed. In this research *an efficient approach PPNM (Pakistani Phonetic Name Matching)* is proposed which phonetically match name strings by using set of preprocessing rules proposed in this thesis for Urdu language.

No specific technique has been designed and implemented for Pakistani names up to now. Another contribution of this thesis is the creation of a new dataset for Pakistani names which covers the variations of spellings against these names. This approach is implemented and then justified by performing number of experiments using the created dataset. After comparing this approach with Edit distance technique for name searching, it can be called an efficient approach for Pakistani names. iv

Dedication

To my respected parents whose utmost love, care and struggle against all odds brought me to this height of knowledge with the blessings and help of the
ALLAH ALMIGHTY v

Acknowledgement

The author expresses her gratitude, appreciation and sincere thanks for Dr. Tahir Ejaz for his supervision of this research work. His advice, guidance and assistance in the preparation of this thesis are thankfully acknowledged.

The author also wishes to thank the advisory committee members Mr. Mubashar Baig and Dr. Adnan Abid for their suggestions and advice at every stage. Guidance and assistance for write up provided by Dr. Adnan Abid is thankfully acknowledged.

The author also acknowledges her family for their constant encouragement and support in the making of this work.

.

able of Contents

Introduction 5 1
1.1 Motivation 5
1.2 Contributions to Field..... 6
1.3 Problem Statement 6
1.4 Thesis Overview 8
Related Work 9 2
2.1 Phonetic Encoding Technique 9
2.1.1 Russell Soundex Code Algorithm 9
2.1.2 Henry Name Matching and the Daitch Mokotoff Coding Method 9
2.1.3 Metaphone Coding Technique 10
2.2 Pattern Matching Techniques 10
2.2.1 K-String Algorithm 11
2.2.2 Q-Grams 11
2.2.3 Edit Distance Technique 11
2.2.4 Gloria Guth's Algorithm 11
2.3 Combination of Phonetic Encoding and Pattern Matching Techniques 12
2.3.1 Editex Technique 12
2.3.2 Syllable Alignment Pattern Searching 12
Methodology 14 3
3.1 Proposed Strategy 14
3.2 Preprocessing 15
3.3 Matching..... 16
3.3.1 Phonetic Distance..... 16
3.3.2 Matching 18
Urdu Names’ Repository 22 4
4.1 Dataset 22
4.1.1 Dataset Statistics 32
Evaluation and Discussion..... 34 52

5.1 Experimental Design 34
5.1.1 Operating Parameters 34
5.1.2 Evaluating Measures 36
5.2 Experimental Results..... 37
5.2.1 Threshold Values for PPNM Approach 37
5.2.2 Threshold Values for edit distance technique 38
5.2.3 Comparison of Thresholds for PPNM and Edit Distance 39
5.2.4 Results of Experiments for PPNM and Edit Distance 40
5.2.5 Comparison of Number of Variants for PPNM and Edit Distance 42
Conclusion and Future Directions 43 6
References 45 7
APPENDIX 47 8 3

List of Figures

Figure 3-1 Pakistani Phonetic Name Matching Approach 14
Figure 3-2 Algorithm for PPNM Approach 18
Figure 3-3 Algorithm for Rules 19
Figure 3-4 Algorithm for Matching 20
Figure 3-5 Algorithm for Phonetic Distance Computation..... 21
Figure 4-1 Dataset Statistics 33
Figure 5-1 Average Edit Distances for all Names 35
Figure 5-2 Results of Threshold Values for PPNM 39
Figure 5-3 Results of Threshold values for Edit Distance 39
Figure 5-4 Results of Number of Variants for PPNM 42
Figure 5-5 Results of Number of Variants for Edit Distance 42 4

List of Tables

Table 2-1 Metaphone Coding Technique 10
Table 3-1 Assignment of Substitution Cost (a)..... 17
Table 3-2 Assigned Substitution Cost (b) 18
Table 4-1 Name Variations for Pakistani Names in Roman Urdu 23
Table 5-1 Operating Parameters 36
Table 5-2 Defining TP, TN, FP, FN 36
Table 5-3 Results of Threshold Values for PPNM 38
Table 5-4 Results of Threshold Values for Edit Distance 38
Table 5-5 Results of Number of Variants for PPNM 41
Table 5-6 Results of Number of Variants for Edit Distance .....