Methods and Algorithms for Searching Arabic Name Entity

Boumedyen Shannaq, Ph.D
Mazoon College "University College"
Computer Science & IT Dept, Sultanate of Oman

ABSTRACT
This research paper is an attempt to revise the basic algorithms which have been developed for name matching, such as English Soundx and n-gram. It's an attempt to developed the basic algorithms to process the Arabic names . The main objective of this work is to develop new algorithm to improve the name matching for Arabic names .The developed algorithm could be used also to process other names from other languages .It's an attempt to develop and enhance work places who in charge to collect customer information such as Civil Service Department ,Bank Systems ,and other organizations. This work will study and analyze the most common used algorithms that used to process personal names. For this purpose this wok builds database of specific Arabic names of 6753 along with 93 training queries and 60 test queries. The basic rustles are based on the comparative analysis of ASoundx and n-gram algorithm respectively . Recall and precision measuring effectiveness have been used to evaluate the proposed algorithm. The obtained results are superior to existing approaches .This work also develop a special software that could be used by entry data worker to help them to process and to select the proper personal Arabic name entry for a particular client .

General Terms
Arabic name matching, Recall, Precision , Algorithms.

Keywords
Name matching algorithms, ArabicSoundx, Soundx , n-gram, Information retrieval Systems.

1. INTRODUCTION
Personal names are easily misspelled. Many Arabic elderly people don't know very well how to write their names correctly.. Workers of Civil Service Department and other similar organizations have to make their best guess at how to register personal information of their customers particularly elders customers and their representative information.. This Research has made an attempt to develop English Soundx to Arabic and propose new language independent approach.

2. CHARACTERISTICS OF THE ARABIC LANGUAGE
One of the earliest and most complete studies of classical Arabic was done by Sibawayh in his great work Known as al-Kitab 'The Book' [1]. The Arabic language is one of the most popular natural languages. More information about the Arabic Language could be found in [2][3][4][5].

3. NAME MATCHING TECHNIQUES
3.1 Asoundx
Asoundx, was developed as describe in [6]. soundex technique developed English Soundx to Arabic, Table 1 illustrate the code set developed for Asoundx.

3.2 N-gram
From [7], introduces two n-gram techniques ," gram-count =|N1 \ N2| , Gram-dist=|N1| + |N2| − 2 |N1 \ N2| " Table 2 show how N-grams yielded by 'boumedian'.

3.3 ArabicSoundx
For building technique for ArabicSoundx, this research Paper used exist technique approach of English Soundx that using English letters. For ArabicSoundx technique this research translate English letters to Arabic letters and choose the first letter to be constant as in English Soundx technique. This research not restricts the length of encoded words to 4, as in the original English Soundex algorithm. This research paper experimented different lengths of encoded words trying to discover better performance. Table 1 shows the code set developed for ArabicSoundex.. For example the code set for ' اسامة ' is 250 ا because first letter is kept constant 0= ا , 2= س , 0= ة , 5= م all zero between codes are deleted, the word must consist of four letters only and code set for ' غولام ' is 450 غ

Table 1. Asoundx code-set

Code	Arabic Letters	English Letters	Category
0	ا,و,ي	a,i,o,e,u	Vowel Letters
1	ف ,ب	b , f	Labial
2	س,ز ,ج,خ,ق,ظ,ص,ك	k,q,z,s,c,j,kh	Guttural &Sibilants
3	ذ,د,ث,,ت,ط,ض	t,d	Dental
4	ل	l	Long Liquid
5	ن,م	m,n	Nasal
6	ر	r	Short Liquid
7	ش	sh	Sharp dental
8	غ	gh	Guttural aspirate
9	ه,ة,ح	h	Aspirate
A	و,ؤ	w	Labial semi-vowel
B	أ,أ,ء,ع,اّ	Based on diacritic	Vowel

Table 2. N-gram yielded by 'boumedian'

Type	n	Grams
Big-grams	2	'bo','ou','um','me','ed','di','ia','an'
Tri-grams	3	'bou','oum','ume','med','edi','dia','ian'

Table 3.Initial ArabicSoundx code-set

Code	Arabic Letters	English Letters	Category
0	ا,و,ي	a,i,o,e,u	Vowel Letters
1	ف ,ب	b , f	Labial
2	س,ز,ج,خ,ظ,ص,ك	k,q,z,s,c,j,kh	Guttural &Sibilants
3	ذ,د,ث,,ت,ط,ض	t,d	Dental
4	ل	l	Long Liquid
5	ن,م	m,n	Nasal
6	ر	r	Short Liquid

3.4 EAsoundex

As shown in Table 3 that ، ش ، غ ،ة ،ه ، و ؤ ، إ ، أ ، آ ، ح ، ء ، ع letters have no counterparts in English .
EAsoundex, and the length of name is still four and all zeros between codes are deleted as in ArabicSoundx**.**

Table 4 . Improved EASoundex code-set

Code	Arabic Letters	English Letters	Category
0	ا,و,ي,أ,إ,و,ؤ,ء,ى,ة	a,i,o,e,u	Vowel Letters
1	ف ,ب	b , f	Labial
2	س,ز,ج,خ,ق,ظ,ص,ك	k,q,z,s,c,j,kh	Guttural &Sibilants
3	ذ,د,ث,,ت,ط,ض	t,d	Dental
4	ل	l	Long Liquid
5	ن,م	m,n	Nasal
6	ر	r	Short Liquid
7	ش	sh	Palatal
8	غ,ع	gh	Velar
9	ه,ة,ح	h	Pharyngeal

Example: the code set according to Table 3 for ' غولام ' is 450 غ , for ' بومدين ' is 530 ب and for ' عبدالغني 'is ع 134.

3.5 WESoundx

In this technique the first letter assume not to be constant is translate to code set according to Table 6, the length of name is four and all zeros between codes 6 are deleted. for example the code set according to
Table 4 for 'غولام' is 8450 , for 'بومدين ' is 1530 and for ' عبدالغني ' is 8134.

3.6 EASoundx(5)

This technique is similar to EASoundx in code set as in Table 3 but the length of the code is extended to be five. For example the code set for 'غولام' is غ 4500
for ' بومدين ' is ب 5350 and for ' عبدالغني ' is ع 1348

3.7 EASoundx(6)

This technique is similar to EASoundx in code set as in Table 6 but the length of the code is extended to be
sex . For example the code set for 'غولام' is غ 45000 , for 'بومدين' is ب 53500 and for 'عبدالغني ' is ع 13485 .

3.8 EASoundx(7)

This technique is similar to EASoundx in code set as in Table 6 but the length of the code is extended to be seven. For example the code set for 'غولام' is
غ 450000 ,for ' بومدين ' is ب 535000 and for 'عبدالغني ' is ع 134850 .

3.9 DNSA (Diagonal Name Search for Arabic)

The main idea of this technique is search diagonals which consists of more consequently numbers of ones
in a matrix.

3.9.1 How DNSA technique work

1- build a matrix :-Size of this matrix depended on the length of the name which we want to find for its similar names in Dataset and the names that are in the Dataset
-Length of rows equal to the length of the query name.
-Length of columns equals to the length of the name in the dataset.
-Element of this matrix consists of zeros and ones
-Value of Element equal to one if any letter of query name is the same letter of any letter name which is in the dataset.
-Value of Element equals to zero if any letter of query name is not the same letter of any letter name which is in the dataset.
2- Calculate sum of the most diagonal which consists
of more consequently numbers of ones in matrix
3- Calculate similarity score according to formula (3.3) for query name and names that exist in the dataset if any name in the dataset have similarity scoreless or equal to zero point two (0.2) is consider as similar name to the query name.

$$Ss = 1 - (sum / L) \qquad (3.3)$$

Ss are the similarity score.
Sum is the summation of ones on the most diagonal which consists of more consequently numbers of ones in matrix. L is the length of the name that exist in the dataset
4-Assign a value for the similarity score.
This Research assume that the value of the similarity score is less or equal to zero point two (0.2) for some reasons that will be explained later.
Example
Query name is ' اسامة'
Exist name is ' أسامه'
Matrix for these names is look like as in Table 5.

Table 5. Matrix for ' اسامة ' and ' أسامة'

	ا	س	ا	م	ة
أ	0	0	1	0	0
س	0	1	0	0	0
ا	1	0	1	0	0
م	0	0	0	1	0
ة	1	0	0	0	1

Similarity Score Calculation
Sum= 4
L=5
Ss=1 – (4/5)
Ss=0.2
Ss < = 0.2 this mean that 'أسامة 'and 'اسامة 'similar to each other according to DNSA technique.

4. EVALUATION RESULTS

Since Database including only Arabic names is not readily available, this research was exclusively correct out using a database borrowed from a Jordanian telecommunication company. The names included in the database were reported without diacritics (Tashkel), with the exceptions of the Arabic letter "Alif" (أ) , for example "بومدين احمد نهار" is split to بومدين احمد نهار , This Research used names from dataset as queries, select names queries randomly and then evaluated the result for relevance Queries which selected according to group of specialists. Also this Research prepared 93 training query according to group of Arabic teacher, these queries was then used for evaluation different technique. Because that all dataset names without diacritics except آ , the diacritics was deleted from set of evaluation queries.

4.1 Result for ArabicSoundx

The selection of the best ArabicSoundx technique was based on the 'average precision and R-Precision' measurement. Table 6 shows the effectiveness of the different ArabicSoundx techniques. Testing these tybes of ArabicSoundx on training set of 93 queries, for ArabicSoundx and EAsoundex this work recommend a code-set of four letters the first character of every name was chosen, while the next three characters were translated according to Table 3 and Table 4. This research shows that EAsoundex(5) is the best of these techniques as shown in Table 6 based on Average Precision result, but based on R-Precision result the EAsoundex is the best as shown in Tables 7 , 8. However the results for soundex(5), EAsoundex(6) and EAsoundex(7) were very similar.

Table 6. Arbicsoundx version

algorithm	Technique	Average Precision	R-Precision
1	ArabicSoundx	0.2755879	0.1740083095
2	EAsoundex	0.3908905	0.24515494
3	WAsoundex	0.2302164	0.13623578
4	EAsoundex(5)	0.3983568	0.13917755
5	EAsoundex(6)	0.3983568	0.13917755
6	EAsoundex(7)	0.3983568	0.13917755

Table7. Average R-Precision for ArabicSoundx and EAsoundex

algorithem1	algorithem2	AverageRP(T1/T2)(93)
ArabicSoundx	EAsoundex	- 0.071146637

Table8. Average R- Precision for EAsoundx and EAsoundex(5)

algorithem1	algorithem2	AverageRP(T1/T2)(93)
EAsoundex	EAsoundex(5)	0.105977396

As shown in Table 7 and Table 8 the best algorithm of ArabicSoundx algorithms for searching Arabic names is EAsoundex. Figure 1 shows how changing the length of developed ArabicSoundx techniques codes between four and seven according to average precision result from Table 6.

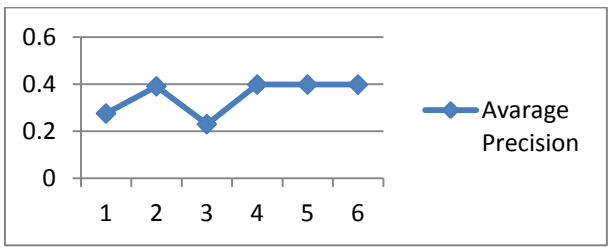

Fig. 1 : Modifying in average precision with code length 4 – 7

Figure 2 shows how modifying of the length of the developed ArabicSoundx techniques codes between four and seven according to R- precision result.

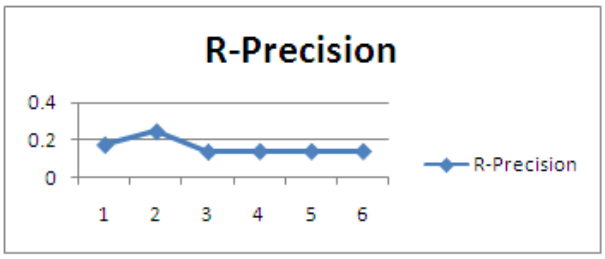

Fig. 2: Change in R- precision with code length 4 – 7

From Table 6. effectiveness of ArabicSoundx techniques is enhanced as codes length is between four and five, there are significant improvements as first character is a constant and the length of the code set is equal to five.

4.2 Asoundx and EAsoundx

Table 9 shows that Asoundx is better than EAsoundx

Table9. Average R- Precision for EAsoundx and EAsoundex(5)

EAsondex	EAsoundex(5)	0.105977397

As shown in Table 7 and Table 8 the best algorithm of ArabicSoundx algorithms for searching Arabic names is EAsoundex. From Table 6. Effectiveness of ArabicSoundx techniques is enhanced as codes length is between four and five, there are significant improvements as first

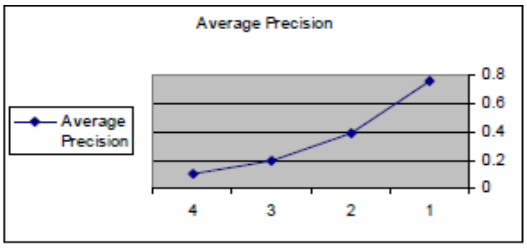

Fig. 3 : Average Precision for versions of DNSA

From Table 9 and in figure 3, when the score is decreased there is an improvement in average precision so , this Research take the best of these different types and calculate R-precision for the best two types Table 10.

Table 10 . effectiveness of DNSA when score is < = 0.2 and <= 0.3

number	technique	Score result	Average Precision	R-Precision
1	DNSA	< = 0.2	0.757976	0.597043010
2	DNSA	< = 0.3	0.392594	0.259129430

Based on Table 12 when the best Score for the DNSA technique is < = 0.2.

4.3 Comparison (Average Precision)

Table 13 shows research result after testing set of 93 queries, 93 user query and 5501 dataset of Arabic names. Table 11 and figure 4 show that DNSA technique with score value (0.2) is the best of these techniques for the user. figure 8 shows the variant Average Precision for these algorithms.

4.4 Comparison (R - Precision)

Table 12 and figure 4 shows that the R-Precision measured which selected the best algorithm that will retrieve the best result for the query .

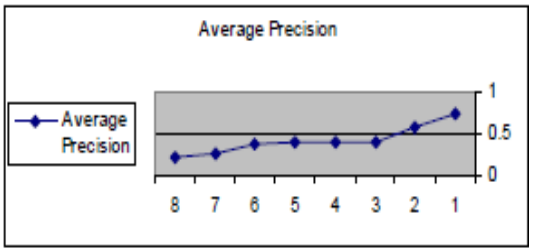

Fig. 4 : Average Precision of 93 queries for variant techniques

Table 11. Average Precision for 93 queries

number	Algorithm	Average Precision
1	DNSA(0.2)	0.757976437995639
2	Asoundx	0.581102181645489
3	EAsoundex(5)	0.398356809119298
4	EAsoundex(6)	0.398356809119298
5	EAsoundex(7)	0.398356809119298
6	EAsoundex	0.390890578899611
7	ArabicSoundx	0.275587997348391
8	WAsoundex	0.230216419466555

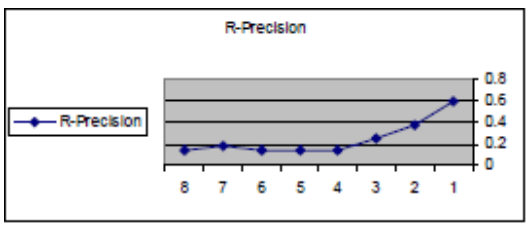

Fig. 5 : R- Precision of 93 queries for variant techniques

Table 12. R- Precision of 93 queries for variant techniques

number	Algorithm	R-Precision
1	DNSA(0.2)	0.597043010752688
2	Asounx	0.374473032940775
3	EAsoundex	0.245154946033961
4	EAsoundex(5)	0.139177550547196
5	EAsoundex(6)	0.139177550547196
6	EAsoundex(7)	0.139177550547196
7	ArabicSoundx	0.174008309513587
8	WAsoundex	0.136235781470184

5. CONCLUSION

This research paper has made an attempt to develop two new techniques for searching the Arabic names, and compare them to other existing techniques, the Arabicsoundx techniques which extended to EAsoundx , Easoundx(5), Easondx(6) , Easondx(7) and Wasoundx based on English Soundx . Another technique DNSA with score value equal to (0.2) . For evaluation purposes, this research constructed a collection of 6753 Arabic names along with 93 user queries and 93 testing set of queries. Research result shows that DNSA algorithm is superior to all other algorithms , on the other hand DNSA technique is have a smaller recall value , this mean that DNSA technique is suitable for the user , but its unsuitable for the system. For Arabicsondx which is based on English Soundex will work only with Arabic names not like DNSA which it's a languages independent. This research also shows that all types of Arabicsoundx is superior to n-gram and trigram, but not superior to Asoudx. On the contrary the collection of data set and training set of query have an important factor that contributes to bringing out this result. Most Arabic names selected for query, were of length between 4 and 5. This explains why Esoundx(5),Esondx(6)and Esoundx(7) have always the same result. Also this research concludes that there is few number of errors with Arabic names, because these names are selected without diacritics(Tashkel) .

6. REFERENCES

[1] Whitaker B. Arabic words and the Roman alphabet. http://www.al-bab.com/arab/language/roman1.htm

[2] Al-Shawakfa, E and Evans, M.2001. The Dialoguer: An Interactive Bilingual Interface to a Network Operating System. International Journal of Expert Systems.

[3] Salman H.AL-ANI, 1970. Arabic Phonology an Acoustical and Physiological Investigation by Indiana university.

[4] Nicholas Awde, Putros Samano, 1998. THE ARABIC ALPHABET How to Read and Write IT.

[5] AbiFares, H. 1998. Comparison of Latin and Arabic scripts. http://www.sakkal.com/articles/Arabic_Type_ Articl e/Arabic_Type3.html

[6] Syed Uzair Aqeel, Steve Beitzel, Eric Jensen,Ophir Frieder, David Grossman.On the Development of Name Search Techniques for Arabic.

[7] Zobel, J., & Dart, P. 1995. Finding Approximate Matches in Large Lexicons. Software - Practice and Experience.