Extending Kosovo Civil Registry String Searching Algorithm

Blerim Rexha, Valon Raça, and Agni Dika

***Abstract*—**Offering e-Government services to citizens is linked primarily to civil registry data. Searching for a citizen's data in civil registry is a common service carried out by string search algorithms using unique keywords such as citizen's name and surname. Similar pronunciation of some Albanian language consonants challenges search on citizen's data, names of which are similarly pronounced, despite different spelling.

This paper presents a novel approach for extending string searching algorithm based on Albanian names in Kosovo Civil Registry. This paper compares Levenshtein distance, American Soundex and extended Soundex algorithm results in a database of 271.000 citizens of Prishtina municipality. The extended algorithm accommodates basic rules of pronunciation in Albanian language and its accuracy and efficiency is better than Levenshtein distance and American Soundex.

***Keywords*—**Civil Registry, Levenshtein, Soundex, String Matching, String Searching.

I. Introduction

Until recently, Kosovo Civil Registry (KCR) was implemented as a distributed standalone system, comprising numerous standalone databases; one per each official entitled to register civil status of citizens of the Republic of Kosovo. The referenced solution was followed by many technical problems, while in use by civil registry officials.

Albanian language (Kosovo's most widely spoken language, and its prime official language), includes many letters, of its 36-letter alphabet, which may have similar pronunciation, making it difficult even for native speakers to distinguish the exact spelling. This problem applies also to person names and surnames, which directly hits the KCR.

Merging the distributed standalone databases and the pronunciation similarity of Albanian names produced a single database of 3.5 million people living in Kosovo, which is not even nearly exact figure of corresponding statistic facts [1]. This increased number of citizens on the KCR is direct consequence of the multiple registration of the same citizen, name of whom may contain 'problematic letters' (letters that are spelled differently, but pronounced similarly).

A comprehensive, error free database of the citizens was a necessity, having in mind that all government-to-citizen services should be based upon this system, including voters' database, database of persons involved in crime, employment bureaus and many other services.

Among one of the applications that links KCR data is the Student Management Information System at University of Prishtina, by including the Citizen Number (CIN), KCR's primary key in their X.509 digital certificates serving for digital signing the student grades in database [2].

II. Related Work

Edit distance commonly referred as Levenshtein distance and the Soundex algorithm present the most widely used and most popular algorithms for string matching.

Levenshtein's distance is a metric algorithm primarily concerned with manipulation of string characters (addition, deletion or substitution) to compute the metric distance between two strings [3]. The distance between strings is calculated based on operations manipulations of the first string characters to get the second string.

Soundex algorithm is a phonetic algorithm. It tries to match strings that are more similar in pronunciation. American Soundex, which is a modified version of the original Soundex, is based in encoding each string in a 4-character code, starting with the first letter of the string and followed by three numbers as they are coded, based on rough pronunciation of English letters [4]. However, Soundex is based on English language phonetics and does not serve well when used for string matching in other languages.

Even though Levenshtein distance and Soundex serve for the same purpose, there's a substantial difference between them on techniques to retrieve similar names. While the former has mathematical approach in achieving the goal of string matching, bypassing the rules and conventions of a specific language grammar, making it universally applicable, the latter is very language bound, applicable only in specific languages through tuning the rules of algorithm towards phonetic rules by concentrating in pronunciation, such as giving importance to the first letter of the name, since it predominantly decides how a name sounds [5].

Dr. techn. Blerim Rexha is associate professor at Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:blerim.rexha@uni-pr.edu).

M.Sc. Valon Raça is teaching assistant at Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:valon.raca@uni-pr.edu).

Dr. Sc. Agni Dika is full professor at Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail:agni.dika@uni-pr.edu) .

III. Civil Registry in Kosovo

Government of Kosovo has implemented an electronic version of Civil Registry (CR) on 2008 [6], phasing out all previous standalone databases (developed in Microsoft Access), which primarily served as certificate producer databases, while not maintaining a comprehensive civil registry.

While Kosovo has around 2.1 million inhabitants [1], the data migration process, from around 500 standalone databases from all civil registration offices, produced a database of 3.5 million records of citizens, which after an anti-redundancy process followed by a subjective verification of records by a national committee shrunk to around 2 million records. However, the data migration and removing duplicates process solved only partially the problem of searching for persons in Civil Registry Database (CRD). The remaining problem is searching for a citizen's data on CRD, while the query returns no matching records.

A. KCR Architecture

System components, respectively the general communication architecture of KCR is presented in Fig. 1.

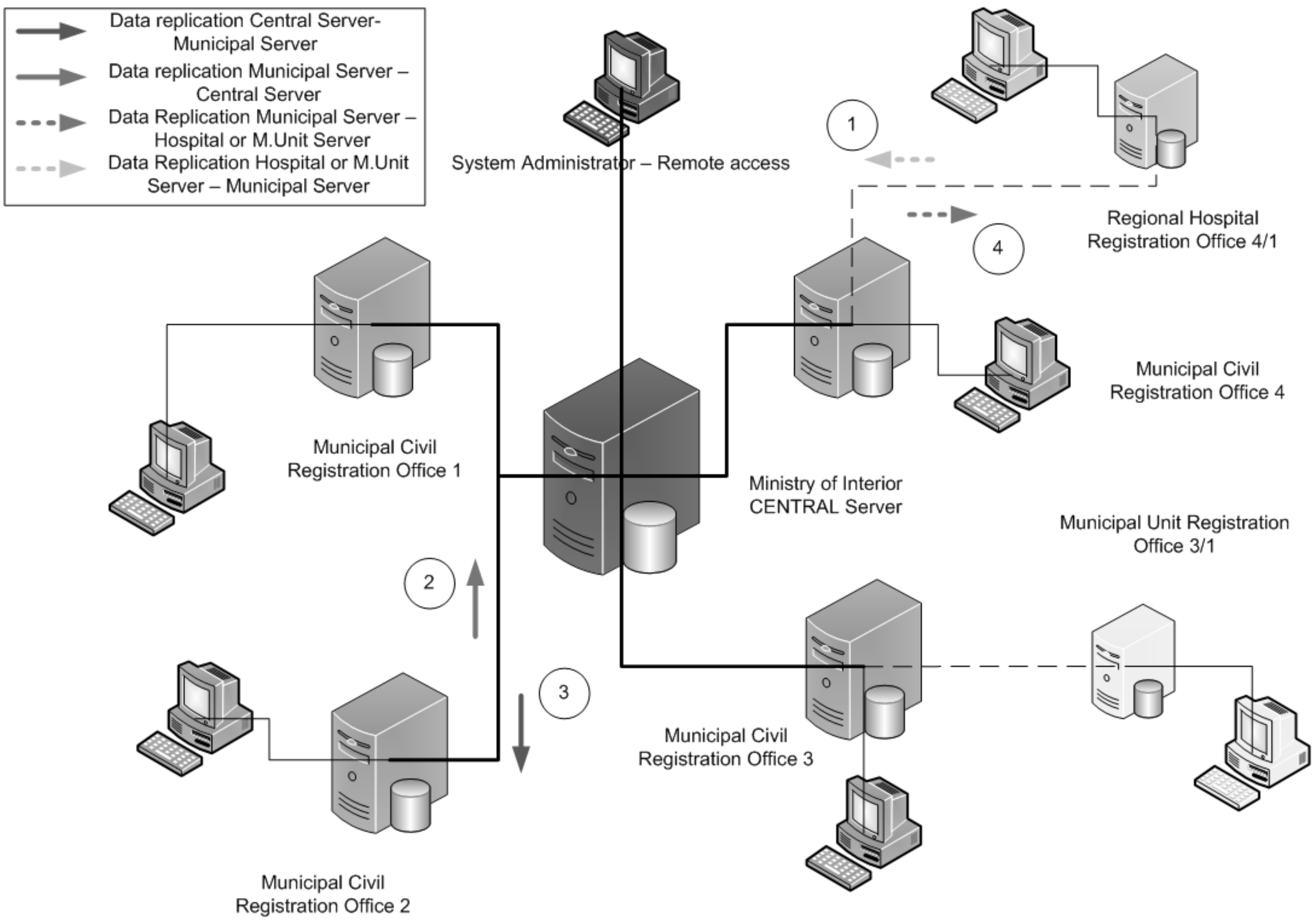

Data replication Central Server- Municipal Server
Data replication Municipal Server – Central Server
Data Replication Municipal Server – Hospital or M.Unit Server
Data Replication Hospital or M.Unit Server – Municipal Server
System Administrator – Remote access
1
4
Regional Hospital Registration Office 4/1
Municipal Civil Registration Office 4
Municipal Civil Registration Office 1
Ministry of Interior CENTRAL Server
Municipal Unit Registration Office 3/1
2
3
Municipal Civil Registration Office 3
Municipal Civil Registration Office 2

Fig. 1. Kosovo Civil Registry Information System – general architecture

Most of the data is accessed by registration officials through municipal servers, since most of citizens are registered and request certificates in the municipalities where they were born. But in cases when a citizen of one municipality requests a certificate or applies for the registration of a civil fact in another municipality such as birth, marriage or death, data are accessed through central server. Updates of the data that are accessed through central server are going to be replicated to all databases containing that record, in order to maintain data consistency and to prevent manipulation.

In the central database server are stored personal data of all citizens of Kosovo. All municipalities store data about the citizens of their administrative territory in their respective municipal databases, which in turn is a subset of data of all citizens. Large municipalities such as capital Prishtina are decentralized, having many municipal units (districts) dealing with the administrative issues and citizen-related services. Thus, the district databases hold the most updated data, and are the source data for the municipal databases. District databases are replicated daily to the municipal database.

Each municipality database server is replicated daily to central server. This process occurs overnight [7].

The database schema is primarily concentrated in facts about citizens, thus storing most essential data in Citizen Entity (Table). This table contains citizen's fundamental data such as name, surname, date of birth and other civil status facts. Also being referred to as First Data Tier, it is the most accessed entity in the database serving for thousands of searches, hundreds of insert and update statements daily (combined for all offices of CR in Kosovo). The second tier of the database comprises of entities which store essential data regarding civil facts such as time of birth, institution where the birth was delivered, place of marriage, place of death and other attributes, all of them stored and organized in separate entities for each fact, as determined in Law on Civil Registers [8]. The third tier of the database is concerned with historical data such as updates of data, or history of certificates issued per person, particularly given that some certificates cannot be issued more than once on six months period.

Officials at municipal and other (lower) level Civil Registration Offices use a Windows platform application, which serves as a front-end for collecting data regarding the civil registration, or for searching on citizens' database for issuing various certificates [7].

In order to provide better government services, by increasing the quality of services and collaboration of government with its citizen, which has a direct impact on quality of governance [9], a web portal for KCR is developed to help citizens apply for certificates, and is considered to be amongst most accomplished e-services that government offers online [10]. KCR has a similar architecture as presented in [11]. Using KCR web portal, citizens may apply online for issuance of certificates. In return the certificate is printed and certified by the CR Officials in 24-hours period from the application time, and is ready for delivery the next day.

The web portal is an independent and standalone system and up until now can only list the submissions made online, but not verify the citizens' data automatically. The remaining part of the job (data verification, retrieval and certificate printing) has to be completed manually by officials of CR.

While there is no Certification Authority (to issue digital certificates) created by Kosovo Government, issuance of electronic civil certificates, digitally signed by officials is not a practice.

B. *Searching for citizen records on CRD*

Data stored in CRD are frequently searched by municipal civil registry officials, as well as other national and municipal officers working for other government and public agencies, as depicted in Fig. 2.

CR officials and other actors of the system cannot search for a large group of people, or do not have access to lists containing bulk data of citizens, due to security reasons, preventing massive extraction of data, thus violating privacy law [7].

Assuring privacy and security of citizen's data, searching by citizen personal number may also not be allowed, as well as with birth dates, especially for e-government services and public information systems, which are secondary actors of the system. Localizing a specific citizen record by matching only name and surname may produce an empty dataset, having in mind that many Albanian names and surnames may be misspelled during search, due to similarity in pronunciation of some letters. Thus, a search for citizen data should allow implementing of certain advanced string matching techniques to help officials localize easier citizen records, by providing a limited list of records, which are similar to each other in terms of citizen name and surname attributes.

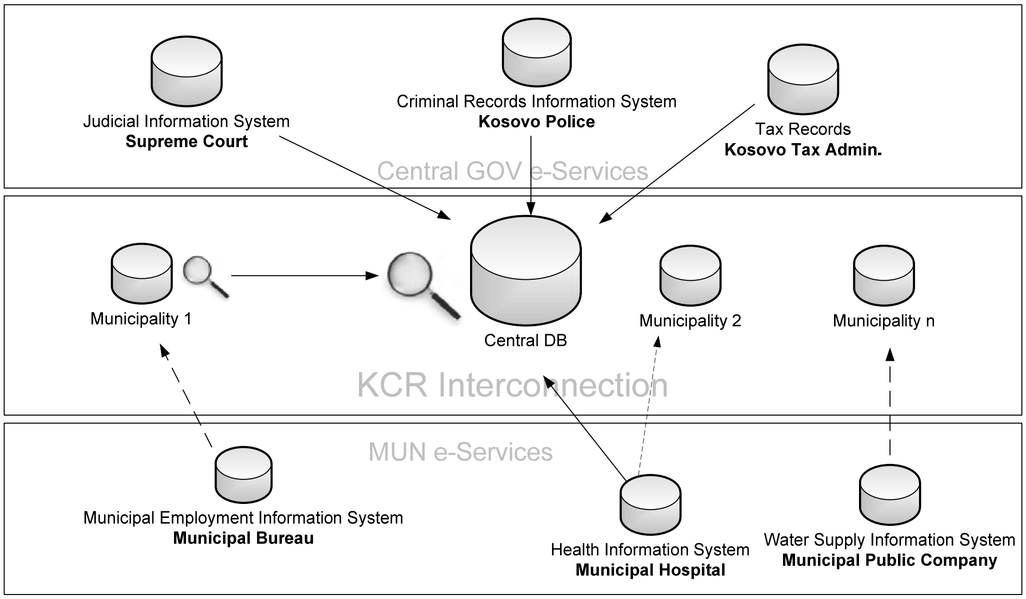


Fig. 2. Searching CRD from native subsystems and other government e-services

C. Using Levenshtein distance in KCR for string matching

Levenshtein distance is the actual algorithm that is used by KCR system for string matching while searching for citizen data. Levenshtein distance is used in the context of KCR Information System for searching on citizen names registered on KCR, personal data of whom cannot be found by using a simple text-search query, due to misspelling of the names, either during the registration or search process.

Levenshtein distance returns a set of good results for variants of a name in Albanian language, but it provides also names which are not expected to be listed as similar to the given name, being verified by a process of manual verification in citizen's database.

Table I presents query results for variants of name *Xhemajl [d͡ʒemajl],* a common name in Kosovo transliterated from Arabic name *Jamal*.

Table I is divided on three parts. Part a) of the Table I represents results for one (1) Levenshtein distance, respectively all the names on the database that can be transformed from the name *Xhemajl* with addition of only one letter, deletion of only one letter from the name or substitution of any letter within the name with another letter. Part b) of the Table I represents results for the same name for variants that can be obtained with at most two operations, and part c) represents results obtained with at most three transformations.
Search for matches with four Levenshtein distances have not been presented, due to a large list with no relevant results, having in mind that most of the names in Albanian are shorter than eight letters, and names that must have at least 4 letters not in common does not represent a set of variants that should be considered for similarity.

Variants presented in bold in Table I represent the *very similar (VS)* variants of the name being searched (best results). Other variants are categorized as *similar (S)*, being variants that are listed because of their similarity to the given name (good results), but not the perfect matches. The category of names that is depicted as *different (D)* includes the names which may be morphologically similar to the given name, but which are bad matches of it (bad results). Queried name is marked with O (original name).

Table I. Variants of name *Xhemajl*, for various distances of Levenshtein algorithm

a) - D(1), Xhemajl		**b) - D(2), Xhemajl**							
Xhemaijl	VS	Çemajl	S	Xhelal	D	**Xhemail**	VS	**Xhemajlie**	V
Xhemail	VS	**Gjemajl**	V	Xhema	S	**Xhemaj**	VS	**Xhemal**	V
Xhemaj	VS	Kemajl	S	**Xhemahil**	VS	***Xhemajl***	O	**Xhemali**	V
Xhemajl	O	Qemajl	S	**Xhemaijl**	VS	**Xhemajli**	VS	Xhemil	S
Xhemajli	VS								
Xhemal	VS								
c) - D(3), Xhemajl									
Çemajl	S	Kemal	S	Xhalal	D	**Xhemajlie**	VS	Xhemko	D
Çemal	S	Kemall	S	Xhela	D	**Xhemajlije**	VS	Xhemush	D
Dimajl	D	Qemail	S	Xhelal	D	**Xhemal**	VS	Xhenata	D
Gjemail	VS	Qemajl	S	Xhelil	D	**Xhemali**	VS	Xhenej	D
Gjemajl	VS	Qemal	S	Xhem	D	Xhemë	S	Xhesa	D
Gjemajll	VS	Qemall	S	Xhema	S	Xhemi	S	Xheva	D
Gjemal	VS	Shehal	D	**Xhemahil**	VS	Xhemil	S	Xhevad	D
Hema	D	Shema	D	**Xhemaijl**	VS	Xhemila	S	Xhevaile	S
Imajl	D	Shemsje	D	**Xhemail**	VS	Xhemile	S	Xhevair	D
Ismajl	D	Shenaj	D	**Xhemaj**	VS	Xhemilja	S	Xhevat	D
Kemail	S	Sheval	D	***Xhemajl***	O	Xhemille	S	Xhezair	D
Kemajl	S	Smajl	D	**Xhemajli**	VS	Xhemka	D	Zhelal	D

The database used for this paper contains 271.000 citizens, of which 15.839 are name variants. Manual verification in database for the variants of the name *Xhemajl* results on 13 very similar variants, not including the given name. Search for the given name, *Xhemajl*, with one Levenshtein distance returned 5 very similar variants out of 13. Running Levenshtein algorithm with two distances produced 9 very similar variants out of 13, while for three Levenshtein distances produced 13 out of 13 possible relevant variants of the name in the database.

Levenshtein distance returns a set of good results for variants of a name in Albanian language, but it provides also names, which are not expected to be listed as similar to the given name by manual processing, which is a drawback, because it may produce a large amount of non-relevant variants. In the referenced case for the three Levenshtein distances it produced 46 variants which are totally different names.

In the Fig. 3 are depicted results represented by Levenshtein distance algorithm for each of relevant distances presented in this section.

As seen in Fig. 3 almost all very similar variants are contained in subset produced by searching with three Levenshtein distances in citizen's database. Result sets produced by each Levenshtein distance are shown as subsets of each other, for example the D(1) is a subset of D(2), thus D(2) contains D(1) result set.

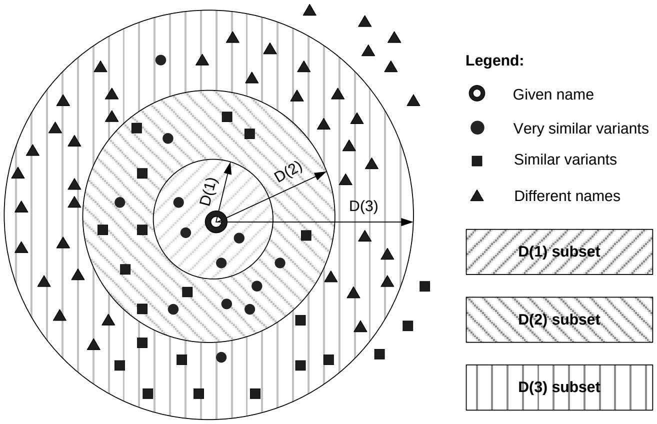
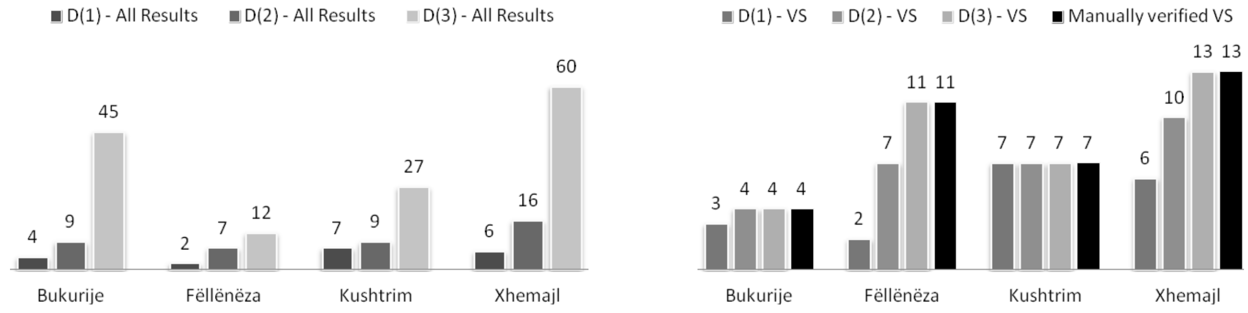


Fig. 3. Graphical representation of the distribution of results – Levenshtein distance

Fig.4 depicts trends of result sets produced by Levenshtein distances on randomly selected Albanian names. As its depicted in part a) of the Fig. 4, size of the resulting set produced after querying with three Levenshtein distances is growing almost exponentially, while the size of the subset of very similar variants as depicted in part b) of the Fig. 4 is remaining stable or growing slightly. Very similar variants returned by querying with three Levenshtein distance are almost always all possible variants, and correspond to manually verified records. While it seems like the solution of the problem in finding relevant misspelled names, unwanted variants appearing in this result set, reduce significantly the chances for an official to manually determine which variant is the one requested each time a search procedure is commenced.

According to experimental findings the resulting subset of very similar variants in a result set retrieved by querying with three Levenshtein distances ranges from 10% - 30% of overall results being qualified as best matches. These figures change when applied for very uncommon names such as the case with name *Fëllënëza*, by giving almost 90% of the expected results. Experimental results also show that when querying for common names which are shorter than 6 letters, applying Levenshtein algorithm with three distances returns only single digit (3% - 8%) percentage of valid variants.



a)

b)

Fig. 4. Trends of Levenshtein result sets for randomly selected Albanian names: a) result sets with all variants returned; b) result sets with only very similar variants returned compared to manually verified variants

Result sets retrieved by querying with two Levenshtein distances always produce a significantly smaller set of variants, and usually contain 50%-70% of the very similar variants based on experimental results. Yet, the result sets retrieved on application of Levenshtein algorithm with two distances for common names shorter than 6 letters generate subsets that contain from 10-25% of the expected results. Figures produced for longer names return perfect sets, but these names represent only a margin amount of Albanian names.

Result sets retrieved by querying with one Levenshtein distance, almost always doesn't include all the possible entries similar to the given name. These queries tend to produce smaller set which is convenient for officials while trying to determine the most proper variant, but those sets do not include generally more than 50% of the expected results, and tend sometimes to include variants that morphologically are very similar variants to the given name, but practically different names such as *Arban* and *Artan*, both of them very common names.

Until now, KCR applies queries with two Leveshtein distances, leaving a valuable subset of relevant results out of final list, in order to ensure usability of the search tool.

Sometimes, KCR system applies edit distance search on both names and surnames simultaneously, with two edit distances on both strings, which produces a more refined set of results for a given citizen record, but the problem of the performance of algorithm is raised, due to increase in complexity of the searching algorithm, consequently raising the processing time from seconds to minutes.

D. Using American Soundex in KCR for string matching

American Soundex is widely in use for string matching based on pronunciation of English words and it is built-in function of most popular database management systems. American Soundex is not suitable solution for string matching in Albanian, even it does present better results than Levenshtein distance in names which are misspelled primarily in vowels, due to lack of encoding in Soundex algorithm for vowels. American Soundex is considered in the context of this paper, also for its efficiency and very good performance of execution, even in very large databases.

American Soundex is very inappropriate when using it to search for the variants of names which contain problematic letters such as similarly pronounced consonants. But it also fails on some parts of Albanian names, due to similar pronunciation of some syllables containing both vowels and consonants such as *aj* and *ai* in *Xhemajl (Xhemail).* While Soundex codes *j,* it doesn't code *i* (vowels are not coded), thus excluding a most common variant of the name.

Table II presents the results for given names *Xhemajl*, *Fëllënëza* and *Lulëzim*. *Xhemajl* is one of the names, constituting consonants of which are mostly misspelled, while *Fëllënëza's* vowels are mostly misspelled. *Fëllënëza* and *Lulëzim* are pure Albanian names, *Xhemajl* is not.

Table II. Result sets for variants produced by American Soundex for Albanian names

a) Xhemajl		***b) Fëllënëza-***				***c) Lulëzim***			
Xhansel	D	Falanik	D	**Fëllenza**	VS	Liljana	D	**Lulzim**	VS
Xhemaijl	VS	**Fellanza**	VS	**Fëllënza**	VS	Llokman	D	**Lulzime**	VS
Xhemajl	O	**Fëllanza**	VS	**Fëllënzë**	VS	Llukman	D	**Lulzon**	VS
Xhemajli	VS	**Fëllanzë**	VS	**Fllanza**	VS	Lulegzim	S	Lulzum	S
Xhemajlie	VS	**Fëllënëza**	O	**Fllanzë**	VS	**Lulëzim**	O		
Xhemajlije	VS	**Fëllënxa**	VS	**Fllënza**	VS	**Luljzim**	VS		
		Fellenza	VS	Flonja	VS	Lulxona	S		

For the given name *Xhemajl*, as presented in Table II, part a), where in most cases consonants that constitute this name are misspelled, American Soundex produced a set of variants, 4 of which are relevant out of 13 variants represented in result set. It also produced a name which is a bad match of the given name.

In case of the name *Fëllënëza*, as presented in Table II, part b), where in most cases vowels that constitute the referenced name are misspelled, American Soundex produced a set with 11 variants out of 11 relevant variants that are represented in the database. It also produced 2 bad matches, but it's important to stress that it produced more relevant results than the result set generated by applying Levenshtein algorithm with three distances for the same name.

In the third case, as presented in Table II, part c), American Soundex produced a mixed set for the given name *Lulëzim*, with roughly same number of relevant and non-relevant results. In most cases experiments returned bad results for string matching in Albanian language using American Soundex, due to Albanian names being heavily constituted with consonants and different phonetic rules from English language.

American Soundex fails to return a good result set even in English language when queried for short names, due to lack of at least 2-3 consonants or at best 4 consonants which define (sometimes uniquely) the resulting code of the American Soundex. 0

Although, Levenshtein algorithm returns better results (in terms of very similar variants) in most cases comparing to American Soundex, the latter performs better in returning results faster due to reduced complexity compared to Levenshtein, and generates sets concentrated more in very similar results, by largely reducing the number of the irrelevant variants or bad matches.

Subsets containing very similar variants present almost 60% - 80% of the returned results, based on experimental results on common Albanan names. But, those variants presents only a small subset of the result set of relevant names present in database and considered similar to the given name.

Fig. 5 depicts the results represented by American Soundex for string matching in Albanian names, generalized based on experimental results. As presented in Fig. 5, roughly half of the presented results are erroneous or bad matches, while a good portion of very similar variants of the given name is not represented at all in the result set (small circles outside of the big shadowed circle).

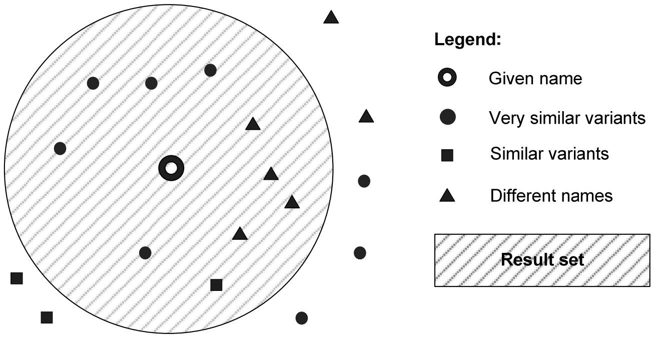


Fig. 5. Graphical representation of the distribution of results – American Soundex

IV. Albanian Soundex

Albanian language is an Indo-European language spoken in Albania, Kosovo and parts of Macedonia, Greece, Montenegro and Serbia.

The Levenshtein algorithm fails to deliver good results on Albanian names such as *Fëllënëza*, which contains vowel *ë* that is not stressed most of the time. The American Soundex fails to deliver good results on Albanian names such as *Xhemajl*, which contains consonants that are mostly misspelled due to their similar pronunciation. Experimental observations show that we have either too much variants represented in the result set, making harder to distinguish between good and bad matches (Levenshtein algorithm), or we have a small result set which does not contain even the very good matches of a specific name (American Soundex)

In this paper is presented a novel approach that returns better results than Levenshtein distance and American Soundex when used for string matching on Albanian names by adjusting the American Soundex coding rules for supporting Albanian language phonetics. This modified Soundex algorithm is named ALSoundex.

A. ALSoundex coding rules

Albanian alphabet contains 36 letters, of which 7 are vowels and 29 consonants. Out of 29 consonants 9 of them are two-character letters.

Unlike English phonetics, where same letters may be pronounced differently, Albanian phonology has a determined sound for each letter, despite their position in the word or surrounding characters.

The problem with string search on names and surnames is generated from consonants that are pronounced similarly like *gj [ɟ]* and *xh [d͡ʒ]* or *q [3]* and *ç [t͡ʃ]*, and the vowel *ë [ə]* (Indo-European schwa) which is not stressed sometimes, therefore not spelled [12]. These problems are present much more with speakers of Gheg dialect, widely spoken dialect of Albanian language in Kosovo.

In Table III are presented Albanian letters as categorized in ALSoundex versus English letters in American Soundex.

Table III. American Soundex vs. ALSoundex coding

Code	American Soundex	ALSoundex
1	B, F, P, V	B, F, P, V
2	C, G, J, K, Q, S, X, Z	C, S, X, Z
3	D, T	G, K
4	L	D, T
5	M, N	L
6	R	M, N
7	-	R
8	-	Gj [ɟ], Xh [d͡ʒ], Sh [ʃ], Zh [ʒ]
9	-	Q [c], Ç [t͡ʃ]

Inherited rules from the American Soundex coding [13]:

- Vowels are not coded despite they were coded on original Soundex [14],
- Consecutive letters are coded only once (*Rr* is coded as *R*, and *Ll* is coded as *L*),
- If there are insufficient characters to be coded, the code is filled with zeros.

Discarded rule from the American Soundex coding:

- Letters that have the same code, and are placed side-by-side are ignored [4].

Modified rules from the American Soundex coding:

- All coded words consist of five characters,
- Initial character is the first letter of the name searched, except when names start with *g (gj)*, *x (xh)*, *q* or *ç,*
- Coding after the first character, consists of 4 numbers,
- *H* and *J* are the only one-character consonants that are not coded, due to their use with other character to form two-character letters,
- *Th*, *Dh*, *Nj* are not coded.

Among the new rules added to the original Soundex is coding two-character letters which are very present in Albanian names. Fig. 6 shows a part of the code handling the coding of two-character letters, for letters ending with character *H*.

```
case (‘C’, ‘S’, ‘X’, ‘Z’):
  if(chrName[i] = (‘S’, ‘X’, ‘Z’)
          AND chrName[i+1] = ‘H’)
  {
    code := 8;
  }
  else
  {
    code := 2;
  }
break;
```

Fig. 6. Coding two-character letters

Letters such as *Gj* and *Xh* or *Sh* and *Zh* share a very common sound, and they were coded accordingly to the 8th category of ALSoundex code.

Letters such as *Q* and *Ç*, were coded accordingly to the 9th category of ALSoundex code.

Th, Dh and *Nj* were not coded because they have a very similar sound to the first character of the letter, or their combination with *h* respectively *j,* is expected to deliver a similar sound as if pronounced one after another (*Th* → *T+h or T, Dh* → *D+h or D* and *Nj* → *N+j).*

Among one-character letters *G* and *K* were categorized in a different category in regard to American Soundex code, due to their different pronunciation in Albanian language in comparison to other letters of the second category of American Soundex.

B. Experimental results

Searching for similarities for the name *Xhemajl* (coded as 86500) with ALSoundex, returns a result set of 24 variants, 13 of which are very similar variants of the name, respectively all of the variants of this name existing in the database are included in the result set. 11 other variants are similar to the given name, while no variants that are bad matches are included in the result set. Generated result set, while containing all very similar variants is significantly smaller set than the one produced when applying three Levenshtein distances. Otherwise, American Soundex has grouped the very similar variants of the given name in four disjunctive sets, with only four variants represented on the same set as the original name, as presented in Table IV, part a).

Table IV: Result sets for variants produced by ALsoundex for Albanian names

a) ALSoundex('Xhemajl')

Variant	Rated Importance	ALSoundex code	Am. Soundex code
Gjemajl	VS	86500	G252
Gjemail	VS	86500	G254
Gjemal	VS	86500	G254
Gjemali	VS	86500	G254
Gjemil	S	86500	G254
Gjemila	S	86500	G254
Gjemile	S	86500	G254
Gjemilja	S	86500	G254
Xhemaijl	VS	86500	X524
Xhemajl	O	86500	X524
Xhemajli	VS	86500	X524
Xhemajlie	VS	86500	X524
Xhemajlije	VS	86500	X524
Xhemahil	VS	86500	X540
Xhemail	VS	86500	X540
Xhemal	VS	86500	X540
Xhemali	VS	86500	X540
Xhemil	S	86500	X540
Xhemila	S	86500	X540
Xhemile	S	86500	X540
Xhemille	S	86500	X540
Xhmile	S	86500	X540
Xhamilja	S	86500	X542
Xhemilja	S	86500	X542

b) ALSoundex('Fëllënëza')

Variant	Rated Importance	ALSoundex code	Am. Soundex code
Fellanza	VS	F5620	F452
Fëllanza	VS	F5620	F452
Fëllanzë	VS	F5620	F452
Fëllënëza	O	F5620	F452
Fëllënxa	VS	F5620	F452
Fellenza	VS	F5620	F452
Fëllenza	VS	F5620	F452
Fëllënza	VS	F5620	F452
Fëllënzë	VS	F5620	F452
Fllanza	VS	F5620	F452
Fllanzë	VS	F5620	F452
Fllënza	VSc	F5620	F452

c) ALSoundex('Lulëzim')

Variant	Rated Importance	ALSoundex code	Am. Soundex code
Lulëzim	O	L5260	L425
Luljzim	VS	L5260	L425
Lulxona	S	L5260	L425
Lulzim	VS	L5260	L425
Lulzime	VS	L5260	L425
Lulzon	VS	L5260	L425
Lulzum	S	L5260	L425

ALSoundex returns perfect result set, as presented in Table IV, part b), for the name *Fëllënëza*, giving 12 very similar variants out of 12 verified variants of the name in the database. The ALSoundex result set for this name is a subset of the American Soundex set, discarding bad matches returned by the American Soundex. Bad matches produced by American Soundex were not shown, due to their irrelevance in regards to ALSoundex.

Regarding the given name *Lulëzim*, as presented in Table IV, part c), ALSoundex returned 4 very similar variants out of 4 verified variants of the name *Lulëzim* in the database. In this case ALSoundex returned also 2 similar variants to the given name, while no bad matches were returned by ALsoundex.

Based on the experimental results, averagely 78% of the elements of the result set returned by ALSoundex are very similar variants of the given name, while 22% of the set elements are mainly similar variants, and in rare cases different names or bad matches. Bad matches are most probable when searching for the names, which do not contain more than 2 consonants. These results show that ALSoundex subsets containing similar variants and different names are amongst smaller subsets in comparison to Levenshtein and American Soundex, making ALSoundex more efficient, while generating a list of suggestions with elements that are expected to be listed.

In another view, subsets containing perfect matches (very similar variants to the given name), returned by the ALSoundex, contain in average 94% of the manually verified very similar variants in the database. As presented in Fig. 7 there are almost no relevant results that are outside the result set returned by ALSoundex.

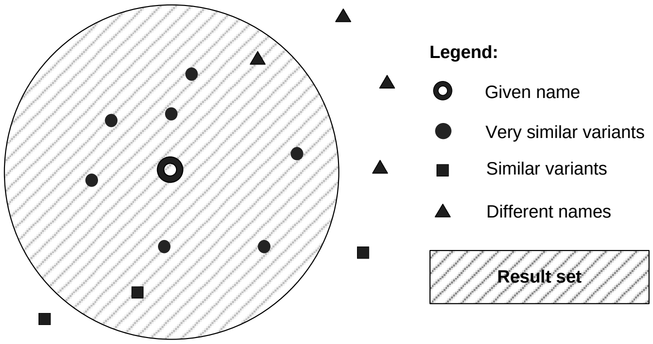


Fig. 7. Graphical representation of the distribution of results – Albanian Soundex

V. Conclusions

The proposed solution for string matching on Albanian names, ALSoundex algorithm was tested against Prishtina municipality database. The returned result sets are more complete, i.e. they contain all variants of citizen names for a given name. A comparison between returned result sets for a given name Xhemajl for all three algorithms: (i) Levenshtein D(2), (ii) American Soundex and (iii) ALSoundex is presented in Fig. 8. The small black circles, representing best matches, are all member of ALSoundex set.

The ALSoundex algorithm is also more efficient than Levenshtein. In Prishtina municipality database that contains 271.000 citizen records, returned result set for the given name Xhemajl is retrieved within one second, whereby the Levenshtein returns result within 47 seconds.

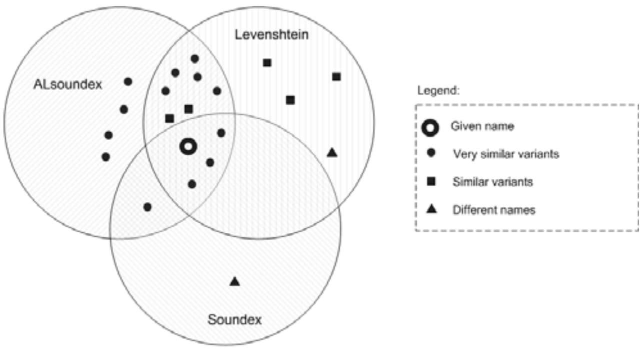


Fig. 8. Distribution of the results for the name *Xhemajl* – intersection of result sets

The quality of result set returned by ALSoundex is more relevant for a given name than Levenshtein's result set and by far more qualitative than American Soundex result set.

References

[1] Statistical Office of Kosovo. 2011. Key Indicators of Population. Available from: http://esk.rks-gov.net/eng/

[2] B. Rexha, H. Lajqi and M. Limani. Implementing Data Security in Student Lifecycle Management System at the University of Prishtina. WSEAS Transactions on Information Science and Applications, ISSN: 1790-0832, vol. 7, pp. 965-974, July 2010.

[3] P. E. Black, 1999. Levenshtein distance. *Dictionary of Algorithms and Data Structures*. U.S. National Institute of Standards and Technology. Available from: http://www.nist.gov/dads/HTML/Levenshtein.html

[4] U.S. National Archives. The Soundex Indexing System. Available from: http://www.archives.gov/research/census/soundex.html

[5] R. Balakrishnan. Country Wise Classification of Human Names. Proceedings of the 5th WSEAS Int. Conf. on Artificiall Intelligence, Knowledge Engineering and Data Bases, Madrid, February 2006, pp. 411-416.

[6] Government of Kosovo. 2008. Strategy on e-Governance 2009-2015. Available from: http://map.ks-gov.net/en/page.aspx?id=160

[7] B. Rexha, V. Raça and M. Arifaj, 2008. Kosovo Civil Registry System Specification

[8] Assembly of Republic of –Kosovo. 2004. Law on Civil Registers. Available from: http://www.assembly-kosova.org/common/docs/ligjet/2004_46_en.pdf

[9] A. S. Drigas, L. G. Koukianakis, Y. V. Papagerasimou. An E-government Web Portal. WSEAS Transactions on Environment and Development, vol. 1, pp. 150-154, 2005.

[10] United Nations Development Program. 2010. eGovernance and ICT Usage Report for South East Europe - 2nd Edition. Available from: http://www.undp.ba/index.aspx?PID=36&RID=107

[11] B. Rexha, A. Ahmeti, L.Ahmedi and V. Komoni. Developing electricity forecast web tool for Kosovo market. WSEAS Transactions on Information Science and Applications, ISSN: 1790-0832, vol. 8, pp. 55-64, February 2011.

[12] A. Kostallari, M. Domi, E. Qabej and E. Lafe, 1974. Drejtshkrimi i Gjuhes Shqipe (Albanian Language Grammar).

[13] G. Mokotoff., 1997. Soundexing and Genealogy. Available from: http://www.avotaynu.com/soundex.htm

[14] R. C. Russell, 1918. U. S. Patent 1261167 Available from: http://v3.espacenet.com/publicationDetails/biblio?CC=US&NR=1261167&KC=&FT=E