Searching Method of Personal Details on the Basis of Fuzzy Comparison

Nataliia Limanova <Nataliya.I.Limanova@gmail.com>,
Maxim Sedov <SedovMN@inbox.ru>
Povolzhskiy State University of Telecommunications and Informatics,
443010, L. Tolstogo, 23, Samara, Russia

Abstract. During the information exchange from one department to another there is a problem of personal identification. This problem concerns people who have partially or completely not coinciding personal details. For the correct comparison of personal data in databases of the source and the receiver it is necessary to perform intellectual search of such data and to bind them to an existing personal identification number. In the article the method and the algorithm of fuzzy search of personal details in databases are offered. The method is based on the modified Levenshtein metrics with use of three operations with symbols: inserts, replacements and removals where all three operations have identical weight. The general flowchart of the algorithm of the fuzzy search with the detailed description of its operation and features is submitted. The developed procedure of identification can be considered as part of the decision-making support system. Procedure doesn't require the operator intervention, gains experience and trains in the process of operation, allowing to exempt specialists completely from low-profile, inefficient, manual operations directly with the sets of personal details which are stored in databases. The built-in system of details priority allows to identify the person in such cases as change of the surname, name, moving and mistakes at manual data input, and in case of partially absent details. Results of technical and economic indicators comparison of the offered method with existing are given. The algorithm is implemented in PL-SQL in the Oracle database 11g and is used since 2007 in commercial operation at the automated information processing in several municipal authorities of the Samara region. In the long term the offered method has potential of successful introduction in systems of global merging of the state or commercial organizations storages for maintaining the uniform database of population of any country of the world. The logical structure of the developed algorithm gives the chance to implement it in any programming language. Features of the offered method allows to apply program procedures on its basis both in small organizations, and in large corporations, everywhere, where is the register of physical persons data.

Keywords: interdepartmental exchange of information; indistinct matching; search of personal details; function of intellectual matching; personal identification number (PIN).

DOI: 10.15514/ISPRAS-2015-27(3)-23

For citation: Limanova Nataliia, Sedov Maxim. Searching Method of Personal Details on the Basis of Fuzzy Comparison. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 3, 2015, pp. 329-342. DOI: 10.15514/ISPRAS-2015-27(3)-23.

1. Introduction

In the course of the interdepartmental information exchange there is an approval problem of the main personal details (full name, birth date, address, passport data, etc.) in databases of various departments. The problem of personal identification has the greatest relevance for physical persons who have partially or completely not coinciding personal details.

For optimum control of big data files, in which the information about physical persons is included, it is necessary to provide centralized storage regulations of such personal details as full name, birth date, address, passport data, etc. Recently various departments – holders of local databases have aimed to combine these arrays for simplification and improvement of work quality. But there is a problem of personal details comparison in different databases. In such cases the elaborated intellectual algorithm of data search in databases or, in the other words, the algorithm of identification of physical persons comes to the aid.

For convenience of data processing to each set of details the so-called personal identification number (PIN) is assigned. In the cases of handling or transferring of physical person data all binding is performed to this PIN. Unfortunately, in Russia, there is no uniform database with personal details of all residents, and therefore in each department the separate register of physical persons is kept, and own PINs are given. The problem arises in the case of residents' information exchange between the organizations. So it is necessary to execute a binding of the entering personal data to the already available information. For an unambiguous binding it is necessary to execute intellectual search of physical person in base receiver which shall consider a set of factors: the mistakes in the case of manual input in the database, the absent or obsolete personal details and etc. It is reasonable to assume that similar search must be implemented in the form of the specialized software [1].

2. Automated search problem

Traditionally this problem is solved by the analysis of identity of the main personal details. There are several details: name, surname, middle name, date of birth, series, passport number and address. Having unambiguously determined coincidence of the existing and new details, it is possible to execute identification of personal details in a database [1][2]. This method of search is carried out manually only in that case when the amount of the transmitted data is small (number of personal details is no more than 30). In case of large volumes of transmitted data the computer comparison of identity of details is used. Such approach allows to determine (50 – 60) % of total number of identifiable personal details. The remained (40 – 50) % is the personal data in which the details in parts or in full don't match. It is more difficult to handle such information manually. Accordingly, the computer search task is divided into three subtasks depending on the type of input data. As a result of comparison the following three types of results can turn out.

1. The person is found. This conclusion can be created as a result of direct comparison of details, and equality of sets of certain key data. In this case the personal details are attached directly to the corresponding PIN.

2. The person is ambiguously determined. This result is displayed in the presence of mistakes, both in new data, and in the earlier received one. For example, the operator's mistakes in the case of manual input of the main details are possible, data corruption during transmission, incorrect work of package requests in case of information processing, etc. In this case the list of PINs which main details are mostly approached to identifiable data is displayed.

3. The person isn't found. This case shows that this personal details is absent in the database and for a binding of this person to the PIN it is necessary to add him to the available data set with assignment of a new PIN.

When creating an automated complex software, which yields above-mentioned results, the most important was to determine borders between the first and second cases, and also between the second and third. The software working without similar differentiation will put down PINs to all found persons unambiguously, and those whose data are determined ambiguously, are removed in the report for manual handling by the operator. Thus all not found persons will be added to base with assignment of a new PIN. Now let us imagine that in case of any discrepancy of the main details, the data will be provided to the report, or that is even worse, will be added as new. For example, the woman name is Nataliia, she got married, respectively she has replaced her surname, she has moved to other residence and she has changed the passport. Besides, in the database she is registered under the name of Natalya, and in her birth date there is a mistake, an incorrectly specified number. When handling such data the program will decide that it is the new person and will add them with assignment of a new PIN. Of course, any task will set to a new PIN in compliance. As a result it turns out that data on one personal detail is doubled and different PINs of one person operate with different tasks. If the error is not corrected immediately, the number of incorrect data will grow up in the geometric progression. On correction of consequences of operation of such software a large number of competent employees of organization will spend a lot of time and forces [3][4][5].

The wrong identification can also lead to a large number of data in the report of manual working off, to assignment of the PIN to incorrect person and to addition of excessive data. At worst case the consequences of such mistakes can completely paralyze work of organization for indefinite time, at the best case – to take away more than 10% of working hours of specialists for errors correction. The analysis of the existing software showed that there is no single identifier; the universal algorithm of identification is also absent.

3. Mathematical model of searching method on the basis of fuzzy comparison

Some types of the metrics reflecting intuitive concept of similarity of lines are known. The most common are Hamming’s distance, Levenshtein’s metric and distance editing [6][7][8].

Hamming's distance is determined for lines of identical length and is set as number of line items in which symbols don't match. In fact, Hamming's distance is calculated as minimum price of transformation of one line in another when the only one transaction of editing lines – replacement is possible.

In a case when it is required to make comparison of lines of different length, Levenstein's metrics or distance editing are used. These two metrics are very similar on creation and actually are the same metrics, little modified for each case. For example, Levenstein's metrics is determined as minimum price of transformation of one line in another with the use of three transactions: inserts, replacements and removals of a symbol, and all three transactions have identical weight.

The distance editing is modification of Levenstein’s metrics in the case when only two transactions are allowed: insert and removal.

Due to the above, Levenstein’s general metrics which supports all three transactions with line was chosen. For further operation the linguistic variable "similarity of lines" was constructed. It is decided to allocate the following terms: "lines match", "lines almost match", "lines are similar", "lines are similar and dissimilar at the same time", "lines aren't similar".

In the result of the analysis of functions of accessory of linguistic terms there was a need to modify the method of calculation of Levenstein’s metrics. It was required to modify metrics so that the distance between lines depended on length of the compared lines.

Theorem 1:

We will designate by means of size $p(s_1,s_2)$ Levenstein's metrics, and size $\|s_i\|$ – length of line s_i. Then function:

$$r(s_1,s_2)=\frac{p(s_1,s_2)}{\max\{\|s_1\|,\|s_2\|\}}, \tag{1}$$

is the metrics.

Proof (not strict proof):

Because $p(s_1,s_2)$ is a metrics, we have:

$p(s_1,s_2)\geq 0$,

$p(s_1,s_2)=p(s_2,s_1)$,

$p(s_1,s_2)+p(s_2,s_3)\geq p(s_1,s_3)$

for any lines s_1, s_2 and s_3. Considering these ratios and equality (1) we come to a conclusion that $r(s_1,s_2)$ satisfies to the first two axioms determining metrics. It is necessary to prove that for any lines s_1, s_2 and s_3 function $r(s_1,s_2)$ satisfies to a triangle inequality:

$r(s_1,s_2) + r(s_2,s_3) \geq r(s_1,s_3)$.

Write this inequality in the form:

$$\frac{p(s_1,s_2)}{\max\{\|s_1\|,\|s_2\|\}}+\frac{p(s_2,s_3)}{\max\{\|s_2\|,\|s_3\|\}}-\frac{p(s_1,s_3)}{\max\{\|s_1\|,\|s_3\|\}}\geq 0.$$

The following cases are possible:

1. $\|s_1\| \leq \|s_2\| \leq \|s_3\|$
2. $\|s_2\| \leq \|s_3\| \leq \|s_1\|$
3. $\|s_3\| \leq \|s_1\| \leq \|s_2\|$
4. $\|s_2\| \leq \|s_1\| \leq \|s_3\|$
5. $\|s_1\| \leq \|s_3\| \leq \|s_2\|$
6. $\|s_3\| \leq \|s_2\| \leq \|s_1\|$

Consider the first case. We have:

$$\frac{p(s_1,s_2)}{\max\{\|s_1\|,\|s_2\|\}}+\frac{p(s_2,s_3)}{\max\{\|s_2\|,\|s_3\|\}}-\frac{p(s_1,s_3)}{\max\{\|s_1\|,\|s_3\|\}}=\frac{p(s_1,s_2)}{\|s_2\|}+$$

$$+\frac{p(s_2,s_3)}{\|s_3\|}-\frac{p(s_1,s_3)}{\|s_3\|}\geq\frac{1}{\|s_3\|}(p(s_1,s_2)+p(s_2,s_3)-p(s_1,s_3))\geq 0.$$

Thus, for the first case the triangle inequality is carried out. As the second case is similar to the first one, based on similar calculations we draw a conclusion that for the second case the triangle inequality is also carried out.

We will turn to consideration of the third case. So, in the third case we have:

$$r(s_1,s_2)+r(s_2,s_3)-r(s_1,s_3)=\frac{1}{\|s_2\|}(r(s_1,s_2)+r(s_2,s_3))-\frac{1}{\|s_1\|}r(s_1,s_3)). \qquad (2)$$

We'll consider a question when the minimum of the function which is in the right part of this equality is reached. It is clear that if expression of $r(s_1,s_2) + r(s_2,s_3)$ reaches the minimum, and $r(s_1,s_3)$ reaches the maximum, the value of all expression will be minimum. The two specified conditions can be satisfied at the same time if two following statements are carried out at the same time:

- lines s_1 *and* s_3 have no common symbols,
- lines s_1 and s_3 are included as sublines in s_2. Then:

$r(s_1,s_3) = \max\{\|s_1\|,\|s_3\|\}=\|s_1\|$,

$r(s_1,s_2) = \|s_3\| + \|C\|$, $r(s_2,s_3) = \|s_1\| + \|C\|$,

thus, the minimum value of expression (2) will register in a form:

$$\frac{\|s_3\|+\|C\|+\|s_1\|+\|C\|}{\|s_3\|+\|s_1\|+\|C\|}-\frac{\|s_1\|}{\|s_1\|}=\frac{\|C\|}{\|s_3\|+\|s_1\|+\|C\|}\geq 0.$$

Therefore, in the third case for function $r(s_1,s_3)$ a triangle inequality is also carried out. Other cases are similar to the already considered. Thus, function $r(s_1,s_2)$ is the metrics, defined in the set of lines. The theorem is proved.

Note: function $r(s_1,s_2)$ belongs to the interval [0,1] for any lines s_1 and s_2.

In the offered algorithm this metrics is applied for operation with line personal details which includes full name, address, document, etc. Therefore the linguistic variable constructed with use of this metrics allows to process requests of search for the person similar to other person in details. Having accepted such request from the user, we actually receive two values: the value of a required detail and the radius of search.

4. Algorithm of the searching method on the basis of fuzzy comparison

The Fig. 1 shows the integrated flowchart of developed algorithm of searching method on the basis of fuzzy comparison. The offered algorithm is presented in the form of process of Data Mining and includes the following stages [9]:

1. analysis of subject domain;
2. problem definition;
3. preparation of data;
4. creation of models;
5. check and assessment of models;
6. model choice;
7. application of model;
9. correction and updating of model.

Consider these steps in details.

1. The subject domain represents data sets with the main personal details in the different organizations and departments.

2. The task of search consists in conditions of single personal identification number absence to search of the details set in one database according to personal details in the other database.

3. Preparation of data represents the organization of the integrated selection including about 300-500 sets, remotely similar to the required. The code fragment organizing programmatically such selection is given below:

```
CURSOR persons
 IS SELECT p.person_id, p.lastname, p.firstname, p.patronymic, p.birthdate
    FROM work.person p
WHERE (((SOUNDEX(TO_TRANSLIT(p.lastname)) =
SOUNDEX (TO_TRANSLIT(fo_Lastname)))
        AND (SOUNDEX (TO_TRANSLIT(p.firstname)) =
SOUNDEX(TO_TRANSLIT(fo_Firstname))))
```

```
        OR ((SOUNDEX(TO_TRANSLIT(p.lastname)) =
SOUNDEX(TO_TRANSLIT(fo_Lastname)))
      AND (SOUNDEX(TO_TRANSLIT(p.patronymic)) =
SOUNDEX(TO_TRANSLIT(fo_Patronymic))))
        OR ((SOUNDEX(TO_TRANSLIT(p.firstname)) =
SOUNDEX(TO_TRANSLIT(fo_Firstname)))
      AND (SOUNDEX(TO_TRANSLIT(p.patronymic)) =
SOUNDEX(TO_TRANSLIT(fo_Patronymic)))));
```

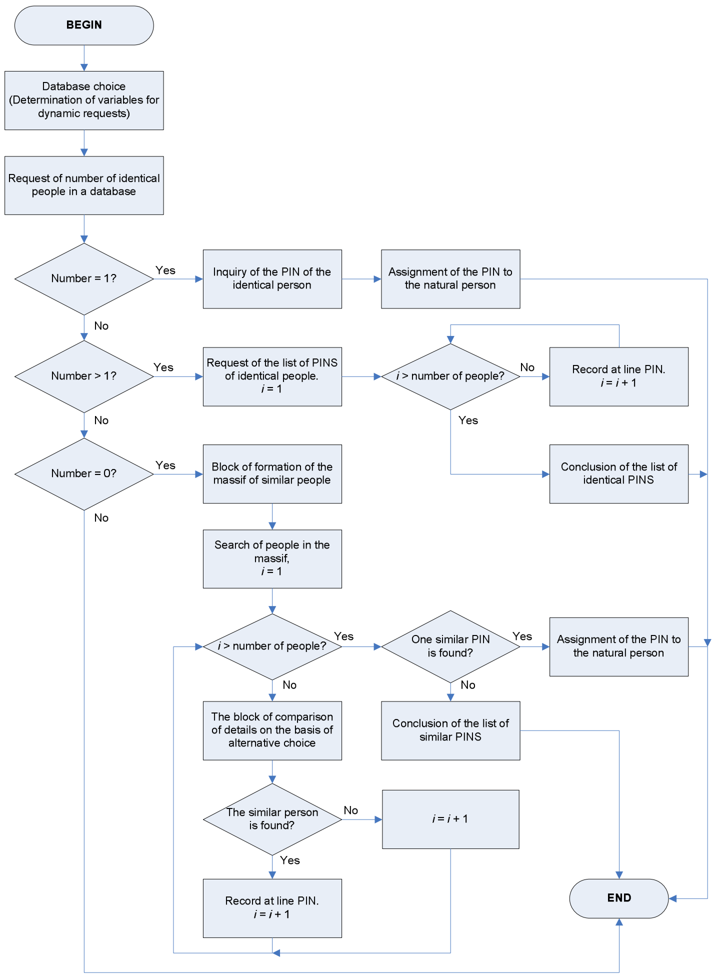

Fig. 1. The integrated flowchart of developed algorithm of search method on the basis of fuzzy comparison.

4. Creation of models consists in detection of regularities in the analysis of data, obtained as the result of step 3, shown in this data set and perhaps suitable for future sets.

5. Check and assessment of models represent testing of regularities for quantity of data sets satisfying with them. The more sets are suitable for specific models the more valuable are revealed regularly.

6. The choice of model consists in detection of the most significant regularities for further using in case of identification procedure future starts.

7. The model application represents regularity using received and approved in case of last start of identification procedure in the current data sets.

8. Correction and updating of models consist in the analysis of result of regularity appendix to a new data set, and, if necessary, correction of model for circle expansion of suitable sets by fuzzy search of personal details compliance.

Programmatically it looks approximately like this (with use of dynamic SQL):

```
-- Perform fast identification
 OPEN cur_Ref_fast_ident
  FOR  'SELECT t.'||v_Col_pin||'
      FROM '||v_Table||' t
    WHERE UPPER(TRIM(t.'||v_Col_lastname||')) =
UPPER(TRIM('''||fo_Lastname||'''))
AND UPPER(TRIM(t.'||v_Col_firstname||')) =
UPPER(TRIM('''||fo_Firstname||'''))
      AND NVL(UPPER(TRIM(t.'||v_Col_patronymic||')), ''_'') =
NVL(UPPER(TRIM('''||fo_Patronymic||''')), ''_'')
      AND t.'||v_Col_birthdate||' =
'''||TO_CHAR(fo_Birthdate, 'dd.mm.yyyy')||'''';
  FETCH cur_Ref_fast_ident  BULK COLLECT
    INTO c_fast_ident;
  CLOSE cur_Ref_fast_ident;
 -- Depending on the number of pins of identical persons
 IF (NVL(c_fast_ident.count, 0) = 1) THEN
   fout_Pin := c_fast_ident(1);
 ELSIF (NVL(c_fast_ident.count, 0) > 1) THEN
   FOR i IN c_fast_ident.first..c_fast_ident.last LOOP
    fout_Pin_list:=fout_Pin_list||TO_CHAR(c_fast_ident(i))||' ';
   END LOOP;
 -- If fast identification didn't yield results
 ELSIF (NVL(c_fast_ident.count, 0) = 0) THEN
    -- write down data from the cursor in collection
    OPEN cur_Ref_full_ident FOR v_Cur_ident;
```

```
FETCH cur_Ref_full_ident BULK COLLECT
  INTO c_full_ident;
CLOSE cur_Ref_full_ident;
 IF (NVL(c_full_ident.count, 0) > 0) THEN
   FOR i IN c_full_ident.first..c_full_ident.last LOOP
-- Perform complete identification
-- The block of comparison of details on the basis of
   alternative choice (see  Fig. 1)
CASE
…
WHEN (UPPER(TRIM(c_full_ident(i).ima)) = UPPER(TRIM(fo_Firstname))
   AND UPPER(TRIM(c_full_ident(i).oth)) = UPPER(TRIM(fo_Patronymic))
   AND ((analyzer_two_number(TO_NUMBER
(TO_CHAR(c_full_ident(i).dtr, 'ddmmyyyy')),
TO_NUMBER(TO_CHAR(fo_Birthdate, 'ddmmyyyy'))) = 1
   AND analyzer_two_number(c_full_ident(i).nom, fo_Passport_number) = 1) OR
((analyzer_two_number(TO_NUMBER
(TO_CHAR(c_full_ident(i).dtr, 'ddmmyyyy')),
TO_NUMBER(TO_CHAR(fo_Birthdate, 'ddmmyyyy'))) = 1
  OR analyzer_two_number(c_full_ident(i).nom,
   fo_Passport_number) = 1)
   AND c_full_ident(i).dom = fo_House
   AND c_full_ident(i).kva = fo_Flat)))
   THEN fout_Pin_list := fout_Pin_list||TO_CHAR(c_full_ident(i).pin)||' ';
   …
WHEN (UPPER(TRIM(c_full_ident(i).fam)) = UPPER(TRIM(fo_Lastname))
   AND UPPER(TRIM(c_full_ident(i).ima)) = UPPER(TRIM(fo_Firstname))
   AND analyzer_two_string(c_full_ident(i).oth, fo_Patronymic) = 1)
   THEN v_Pin_list_sim := v_Pin_list_sim||TO_CHAR(c_full_ident(i).pin)||' ';
   …
 ELSE NULL;
END CASE;
```

In developed implementation of algorithm in PL-SQL DBMS Oracle 11g [10] language, key functions are allocated for logically selected procedures ANALYZER TWO STRING and ANALYZER TWO NUMBER, created on the basis of the modified method calculation of Levenstein's metrics which allow carrying out intellectual comparison of two similar lines or numbers, taking into account possible inaccuracies or errors of input. These procedures can be applied not only for identification of details, but also everywhere where full text search with fuzzy set input data is required.

5. Technical and economic indicators of proposed algorithm

For the comparative analysis of developed algorithm consider technology of identification on the basis of direct comparison. Using this technology the emphasis goes on speed of records handling, but not on quality of decision making by system. As a result, after completion of procedure on the basis of direct comparison, there are many data (about 20-30% of total quantity of the lines) not connected with initial which need to be fulfilled manually that is extremely difficult in the case of large volumes of the processed data.

When comparing working indicators of two algorithms it is revealed:

Algorithm of direct comparison:

Data processing speed: ~ 100 000 lines per hour;

Identification accuracy (probability of exact searching method): ~ 80%

Algorithm of identification on the basis of fuzzy comparison:

Data processing speed: ~ 80 000 lines per hour;

Identification accuracy (probability of exact searching method): ~ 99,9%

It is possible to draw a conclusion that, operator's work in manual operation of results is minimized in developed algorithm i.e. though the speed of handling is slightly less, but the algorithm allows to significantly unload operators at the expense of intellectual system of decision making that can't offer algorithm of direct comparison. When comparing economic characteristics of the developed software on the basis of described algorithm with procedure of direct comparison for annual amount of identification of 1 200 000 physical persons the following data were obtained: labor costs on information processing by the method of fuzzy comparison in comparison with method of direct comparison are reduced by 6,7 times, absolute decrease in labor costs constituted 1 446 hours, annual costs when using the fuzzy comparison method decreased by 3 times in comparison with the similar period of application of the direct comparison method, annual economic effect exceeded 580 000 rub. For descriptive reasons some cost indicators which are created when using the software developed and applied are displayed on the chart provided on Fig. 2. Sizes of costs are postponed on ordinate axis in rubles.

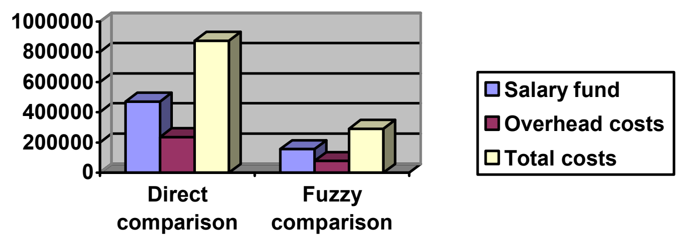

Fig. 2. The chart for the comparative analysis of cost indicators when using methods of direct and fuzzy comparison.

6. Conclusions

The considered method and algorithm are based on fuzzy comparison and on the metrics of Levenshtein. The algorithm, developed in the form of Data Mining process, allows defining people quickly according to earlier carried out search. The built-in system of personal details priority gives the opportunity to identify person in such cases as changing of surname, name, moving, mistakes from manual data input and if personal details are partially absent also.

Self-training systems allow releasing human resources for accomplishment of creative tasks. In this area the Data Mining technology provides a full range of theoretical and practical means for choice, development or use of intellectual computer systems.

The procedure of identification from this article can be considered as part of the system of decision support. The procedure does not require the operator intervention, gains experience and learns in the process of operation, allowing to completely exempt specialists from low-profile, inefficient, manual operation directly with the sets of personal details which are stored in databases.

The developed method and algorithm show good results when fields with different information inside (name, address, postcode, phone etc) are compared. Indeed, any symbolical value, whether it be full name, number of the passport or address, it is possible to present in the form of string. In the course of two strings comparison with the help of the offered algorithm, the distinctions of these lines are revealed, such as the admissions of separate symbols or incorrect single symbols which can arise at typographical errors in a manual data set. I.e., from the point of view of symbol-to-symbol comparison, there is no difference between comparison of two passport numbers or two surnames.

In long terms, this algorithm has the possibility of successful implementation in systems of global merger of storages of the state or commercial organizations, for maintaining a single database of the population of any country of the world. The logical structure of developed algorithm allows realizing it in any popular programming language. Features of algorithm allows applying program procedures on its basis both in small organizations, and in large corporations, everywhere, where the register of physical persons data is conducted and staticized. Possible examples of use: portal of state services, medical electronic systems, personnel and accounting systems of accounting of employees, bank systems of data storage on clients, etc.

The algorithm was carried out by PL-SQL of Oracle 11g database management system. The developed software realized the offered method of the computer search of personal data on the basis of fuzzy comparison was implemented and successfully operates since 2007 in the municipal institution «City information center» in Togliatti town of Samara region.

References

[1]. Selection of materials on the international experience of legislative regulation of use of systems of the personality’s identification (http://www.kongord.ru/Index/Prison/SViP.htm).

[2]. The report on accomplishment of research, developmental work "Development of mechanisms of unambiguous identification of data on the physical persons and real estates which are stored in various information systems of public authorities and local government (http://www.nisse.ru/business/article/article_464.html).

[3]. Regulations on personal identification number of the citizens of the Russian Federation living or staying in the territory of St. Petersburg (http://iac.spb.ru/shablon.asp?subpage=171&id=40&dir=0).

[4]. The "Moscow Social Card" project (http://www.soccard.ru).

[5]. The collection of theses of city scientific and practical conference of students, graduate students, teachers of higher education institutions and specialists of local government offices of Tolyatti "Informatization in the social sphere" (http://it-exclusive.ru/idperson/docs/stat.doc).

[6]. Hamming R. V. The theory of coding and the theory of information, trans. Edited by BS Tsybakov, Radio and Communications, 1983.

[7]. Levenstein V. I. Binary codes with correction of losses, inserts and replacements of symbols, reports of Academy of Sciences of the USSR vol.163, 1965.

[8]. Boytsov L.M. Analysis of lines, http://itman.narod.ru/articles/infoscope/string_search.1-3.html.

[9]. Chubukova I.A.,"Data Mining", training course, publishing house of Internet university of information technologies (http://www.intuit.ru/), 2006.

[10]. Scott Urman, "ORACLE 9i - Programming in PL / SQL", tutorial, Oracle Press – publishing house "Lory", 2004.

Метод поиска реквизитов физических лиц в базах данных на основе нечёткого сравнения

Наталия Лиманова <Nataliya.I.Limanova@gmail.com>,
Максим Седов <SedovMN@inbox.ru>
Поволжский государственный университет телекоммуникаций и информатики,
443010, Россия, г. Самара, ул. Л. Толстого, д. 23.

Аннотация. При передаче данных от одного чреждения к другому возникает проблема персональной идентификации физических лиц, у которых частично или полностью не совпадают реквизиты. Для правильного сопоставления персональных данных в базах данных источника и приемника необходимо выполнить интеллектуальный поиск таких данных и привязку к уже имеющимся персональным идентификационным номерам. В статье предлагаются метод и алгоритм нечеткого поиска реквизитов физических лиц в базах данных. Метод основан на модифицированной метрике Левенштейна с использованием трех операций: вставки, замены и удаления символов, где все три операции имеют одинаковый вес. Представлена общая схема алгоритма поиска на основе нечеткого сравнения с подробным описанием его работы и особенностей. Разработанную процедуру идентификации можно рассматривать как часть системы поддержки принятия решений. Процедура не требует вмешательства оператора, накапливает опыт и самообучается в процессе работы, позволяя, тем самым, полностью освободить специалистов от низкопрофильной, неэффективной ручной работы напрямую с наборами реквизитов физических лиц, хранящимися в базах данных. Встроенная система приоритета реквизитов позволяет идентифицировать человека в таких случаях, как смена фамилии, имени, переезд, ошибки при ручном вводе данных, а также при частично отсутствующих реквизитах. Приведены результаты сравнения технических и экономических показателей предложенного метода с существующими. Алгоритм реализован на языке PL-SQL в СУБД Oracle 11g и используется с 2007 года в промышленной эксплуатации при автоматизированной обработке информации в нескольких муниципальных учреждениях Самарской области. В перспективе предложенный метод обладает возможностью успешного внедрения в системы глобального объединения хранилищ государственных или коммерческих организаций для ведения единой базы данных населения любой страны мира. Логическая структура разработанного алгоритма дает возможность реализовать его на любом языке программирования. Масштабируемость алгоритма позволяет применять программные процедуры на его основе, как в малых организациях, так и в крупных корпорациях, везде, где ведётся и актуализируется реестр персональных данных физических лиц.

Keywords: interdepartmental exchange of information; indistinct matching; search of personal details; function of intellectual matching; personal identification number (PIN).

DOI: 10.15514/ISPRAS-2015-27(3)-23

Для цитирования: Лиманова Наталия, Седов Максим. Метод поиска реквизитов физических лиц в базах данных на основе нечёткого сравнения. Труды ИСП РАН, том 27, вып. 3, 2015 г., стр. 329-342 (на английском языке). DOI: 10.15514/ISPRAS-2015-27(3)-23.

Список литературы

[1]. Подборка материалов о международном опыте законодательного регулирования использования систем идентификации личности (http://www.kongord.ru/Index/Prison/SViP.htm).

[2]. Отчёт о выполнении научно-исследовательской, опытно-конструкторской работы «Разработка механизмов однозначной идентификации данных о физических лицах и объектах недвижимости, хранящихся в различных информационных системах органов государственной власти и местного самоуправления (http://www.nisse.ru/business/article/article_464.html).

[3]. Положение о персональном идентификационном номере граждан Российской Федерации, проживающих или пребывающих на территории Санкт-Петербурга (http://iac.spb.ru/shablon.asp?subpage=171&id=40&dir=0).

[4]. Проект "Социальная карта москвича" (http://www.soccard.ru).

[5]. Сборник тезисов городской научно-практической конференции студентов, аспирантов, преподавателей вузов и специалистов муниципальных учреждений г.Тольятти «Информатизация в социальной сфере» (http://it-exclusive.ru/idperson/docs/stat.doc).

[6]. Хемминг Р.В. Теория кодирования и теория информации, пер. с англ. Под ред. Б.С. Цыбакова, Радио и связь, 1983.

[7]. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов, доклады АН СССР т.163, 1965.

[8]. Бойцов Л.М. Анализ строк, http://itman.narod.ru/articles/infoscope/string_search.1-3.html.

[9]. Чубукова И.А., “Data Mining”, учебный курс, издательство Интернет-университета информационных технологий (http://www.intuit.ru/), 2006.

[10]. Скотт Урман, “ORACLE 9*i* – Программирование на языке PL/SQL”, учебное пособие, Oracle Press – издательство “Лори”, 2004.