Index-Requisite Data Diagnostics In Information Management Systems

Andrey Chukhray[1] [0000-0002-8075-3664] and Olena Havrylenko[2] [0000-0001-5227-9742]

[1] National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine
achukhray@gmail.com
[2] National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine
o.havrylenko@khai.edu

Abstract. Informational Management Systems (IMS) which are based on legacy systems have a significant problem of dirty data. The data cleansing problem solution in such systems usually starts with the search of similar tuples' clusters. After that for each cluster the reference tuple should be formed for saving in a data warehouse of IMS. Moreover, fail tuples should be returned to the source subsystem with the indication of error location, i. e. concrete invalid requisite. The necessary of such a deep diagnosis determined by the following fact: the reference tuple can be not just one of the existent, but as well the combination of several different tuples requisites. Considering one obtained cluster of similar tuples, a certain multiset can be composed from all of the certain attribute values. The paper represents the method of the multiset's diagnostic in terms of faultless and correctability, based on the majority principle. The method provides the minimum time required for establishing the fact of multiset's incorrectness, moreover it allow defining valid (reference) and failed elements of the multiset.

Keywords: Data Cleansing, Diagnostics, Similar Tuples, Reference Requisite, Multiset.

1 Introduction

Informational Management Systems (IMS) which are based on legacy systems have a significant problem of dirty data. The data cleansing problem solution in such systems usually starts with the search of similar tuples' clusters [1]. After that for each cluster the reference tuple should be formed for saving in a data warehouse of IMS.

Therefore, fast index-requisite diagnostics method's development based on the hashing and cyclic codes [2,3] is quite perspective way of data cleaning problem solution. The main principals of the rational control, which have successful implementation in system engineering [4], should be investigated on the real functioning database of IMS.

2 Problem Statement

Let consider the situation when one cluster tuples consist of three requisites. Then the probability $p(BB)$ of the event BB, which means that one of two data tuples has single or double error in one requisite and second tuple – in another requisite, can be calculated as:

$$p(BB) = 2*(p(A_1)p(D_2)p(D_3)*\, p(D_1)p(A_2)p(D_3) + p(A_1)p(D_2)p(D_3)*$$
$$*p(D_1)p(D_2)p(A_3)*\, p(D_1)p(A_2)p(D_3) + p(D_1)p(D_2)p(A_3)),$$

where $p(A_i) = L_i\pi_c(1-\pi_c)^{L_i-1} + C_{Li}^2(1-\pi_c)^{L_i-2}$ are probabilities of requisites fails with single or double error**,** $p(D_i) = (1-\pi_c)^{L_i}$ are probabilities of requisites fails absence, L_i are the average lengths of the requisites values, $i = \overline{1,3}$, π_c is the possibility of the requisite symbol distortion. For example, when $L_1 = 8$, $L_2 = 6$, $L_3 = 10$ and $\pi_c = 10^{-2}$ $p(BB) \approx 0.025$.

Considering only one of the clusters obtained similar tuples proceeds to the formal statement of the problem.

Let *R1* is the considered cluster, i.e. set including $q \in N$ tuples. For every attribute ρ ($\rho = \overline{1,h}$) of the tuple the corresponding multiset $M_\rho = \{sm_{1\rho}, sm_{2\rho}, ..., sm_{q\rho}\}$ should be formed. Then, based on the principle of majority voting (two among three, three among five etc.) is widely used in diagnosing technical systems [4, 5], it is possible to give definitions of multiset M_ρ correctness.

Definition 1. A multiset M_ρ ($|M_\rho| > 2$) is *faultless*, if all elements are equal, i.e.

$$CORRECT\left(M_\rho\right) = \left\{\forall i \in \{1,..,q\} \forall j \in \{1,..,q\}\ (i \neq j) \Rightarrow (sm_{i\rho} = sm_{j\rho})\right\},$$

where $CORRECT\left(M_\rho\right)$ is a Boolean predicate, $\Rightarrow$ is an implication operator.

Definition 2. A multiset M_ρ ($|M_\rho| > 2$) is *correctable*, if more than half, but not all of its elements are equal, i.e.

$$CORRECTED\left(M_\rho\right) = \left\{\exists M_\rho^{'} = \{sm_{1\rho}^{'}, sm_{2\rho}^{'}, ..., sm_{z\rho}^{'}\} \subset M_\rho\,(z > q/2) \wedge (z < q) \wedge\right.$$
$$\left.\wedge\left(\forall i \in \{1,..., z\} \forall j \in \{1,..., z\}\ (i \neq j) \Rightarrow (sm_{i\rho}^{'} = sm_{j\rho}^{'})\right)\right\},$$

where $CORRECTED\left(M_\rho\right)$ is a Boolean predicate, $M_\rho^{'}$ is a subset of M_ρ.

Consequently, an element $sm_{ip} \in M_\rho$ is a reference tuple, if $sm_{i\rho} \in M_\rho^{'}$, as well as an element $sm_{i\rho} \in M_\rho$ is an failed tuple, if $sm_{i\rho} \notin M_\rho^{'}$. It is also obviously, in the case if M_ρ is correctable $M_\rho^{'}$ is unique. In addition, if in M_ρ are equal no more than half of the elements, M_ρ is *not faultless* and *not correctable*.

The objective of this paper is to represent the method of the multiset's diagnostic, which ensures minimal time of M_ρ correctness establishment, based on the given

above definitions, and allow to locate, if it is necessary, reference and failed elements of M_ρ.

Let consider the obviously easiest approach, i.e. pairwise comparison of all the M_ρ elements $sm_{i\rho}$ and $sm_{j\rho}$ for $(i \neq j)$. To do this, we assign the multiset M_ρ another multiset $CNT_\rho = \{cnt_{1\rho}, cnt_{2\rho}, ..., cnt_{q\rho}\}$, such that $cnt_{i\rho} = 1 + \sum_{\substack{j=1 \\ j \neq i}}^{q} eq_{ij}$, $i = \overline{1,q}$ where $eq_{ij} = \begin{cases} 1, & if\ sm_{i\rho} = sm_{j\rho}; \\ 0, & othrwise. \end{cases}$ If $cnt_{1\rho} = q$ then M_ρ is faultless, or M_ρ is correctable if $\exists i \in \{1,...,q\} : cnt_{ip} > q/2$, i.e. there is a reference element. For example, if $M_\rho = \{'Иванов', 'Ивашов', 'Иванов'\}$, then $CNT_\rho = \{2,\ 1,\ 2\}$. $sm_{1\rho}$, $sm_{3\rho}$ are reference elements, $sm_{2\rho}$ is a failed element.

The performance estimation of the pairwise comparisons method [2] by counting the number of symbols comparisons is followed. Let each element $sm_{i\rho} \in M_\rho$ comprises L characters ($L >> 1$). Then the maximum number of character comparisons $C_{pc} = (q-1+q-2+...+1)L = \frac{q^2-q}{2}L$. Consequently, $T_{pc} \approx \frac{q^2-q}{2} L \cdot t_{pc}$, where T_{pc} is the maximum time of the multiset M_p of pairwise comparison elements, t_{pc} is the runtime of the two characters comparison. Since the diagnosing second stage maximum runtime is proportional to q then $T_{cpc} \approx \left(\frac{q^2-q}{2} L + q \right) t_{pc}$ where T_{cpc} is the maximum runtime of the diagnostic procedure based on the method of pairwise requisite's comparison.

A significant improvement of this method is using of a conversion key (hash) [2], which sends requisites to the array indexes (memory address):

$$H : sm_{ip} \to a_{ip} \qquad (2)$$

where H is a transformation (mapping), a_{ip} is an array index corresponding to an element sm_{ip} of the multiset M_p.

The main challenge deal with the key conversion is that the set of possible values is much greater than the set of possible memory locations. Therefore it is necessary to choose a mapping H, which allow:

- to detect common errors of source data, entering by human-operator in input fields, i.e., reference argument sm_{ip} and argument sm_{rp}, failed with any of the specified error, always guarantee different results $a_{ip} \neq a_{rp}$;
- to establish definitely the difference of sm_{ip} and sm_{rp} in the case of different results $a_{ip} \neq a_{rp}$;

– to produce the same addresses for random source elements with the difference of arbitrarily small probability;
– to conclude that the probability $sm_{ip} \neq sm_{rp}$ is arbitrarily small for the equal addresses ($a_{ip} = a_{rp}$).

When mapping H is chosen, the problem of the M_p correctness establishment and the reference and failed M_p elements search can be effectively solved by using diagnostic models [4] linking errors indirect signs with the direct ones.

3 The Choice Of The Requisites-To-Indices Reflection

The stated problem should be considered as a task of error-correcting coding theory: construction of the predetermined code with detecting ability for transmitting discrete information through a noisy channel [3].

Indeed, $\left(sm_{ip}, a_{ip}\right)$ can be regarded as permissible sequence of redundancy code, where sm_{ip} are data bits, $a_{ip} = H(sm_{ip})$ are checking bits. However, in contrast to transmission over the communication channel, whereby both data bits and checking bits could be corrupted because of possible noise, in this case only information bits could be changed, i.e. sm_{ip} is received as $sm_{rp} \neq sm_{ip}$.

The most reliable are the cyclic codes having high detecting ability and widely used in practice because of less complicated coding/decoding devices schemes in comparison with other coding techniques [3]. Constructing the cyclic code for a given number u of data bits the shortest length of code combinations w is determined to provide a predetermined multiplicity of error detection. This problem is reduced to the determination of needed generating polynomial $G(x)$ of degrees $w-u$.

For cyclic codes data bits transformation to test bits has the form as following:

$$H\left(M_{ip}(x)\right) = \left(M_{ip}(x) + x^{w-u}\right) \bmod G(x), \qquad (3)$$

where $M_{ip}(x)$ is polynomial of a dummy variable x corresponding to the data bits, sm_{ip} , mod is an operator getting remainder of the polynomials division.

Thus, it is need to choose a mapping H such as:

1. If $H(M_{rp}(x)) \neq H(M_{ip}(x))$, then $M_{rp}(x) \neq M_{ip}(x)$.
2. For frequently occurring error classes $E_з$ and $M_{rp}(x) = M_{ip}(x) + E_з$ a combination $\left(sm_{ip}, a_{ip}\right)$ is excepted, i.e. $H(M_{rp}(x)) \neq H(M_{ip}(x))$.
3. For random noise $M_{rp}(x) = M_{ip}(x) + E_c$ the probability of $\left(m_{rp}, a_{ip}\right)$ permission is arbitrarily small, i.e. $p(H(M_{rp}(x)) = H(M_{ip}(x))) \to 0$ where E_c is some random noise.

4. The equity $sm_{ip} = sm_{rp}$ is ensured from $H(M_{rp}(x)) = H(M_{ip}(x))$ with a probability close to one.

It is necessary to show further that the first requirement is satisfied if $deg(G(x)) > 0$ where $deq(G(x))$ is a generator polynomial $G(x)$ degree. If $H(M_{rp}(x)) \neq H(M_{ip}(x))$, then $(M_{rp}(x)x^{w-u}) \bmod G(x) \neq (M_{ip}(x)x^{w-u}) \bmod G(x)$. Using the distributive property of the operator mod, it can be obtained $(M_{rp}(x)x^{w-u} + M_{ip}(x)x^{w-u}) \bmod G(x) \neq 0$. Let's supposed, that $M_{rp}(x)x^{w-u} + M_{ip}(x)x^{w-u} = 0$, then $0 \bmod G(x) \neq 0$ is a contradiction, since the condition $\deg(G(x)) > 0$ and the definition of $0 \bmod G(x) = 0$. Consequently, $M_{rp}(x)x^{w-u} + M_{ip}(x)x^{w-u} \neq 0$, $M_{rp}(x)x^{w-u} \neq M_{ip}(x)x^{w-u}$ and $M_{rp}(x) \neq M_{ip}(x)$.

Before finding out conditions that satisfy the second requirement, it is need to introduce auxiliary statements.

Statement 1. If $A(x) \bmod G(x) \neq 0$ and $B(x) \bmod G(x) \neq 0$, then $(A(x)B(x)) \bmod G(x) \neq 0$.

Proof. Let's supposed, that $A(x) \bmod G(x) \neq 0$ and $B(x) \bmod G(x) \neq 0$, but $(A(x)B(x)) \bmod G(x) = 0$, then $A(x) \neq W(x)G(x)$, where $W(x)$ is a polynomial. Multiplying both sides of this inequality by $B(x)$, it could be obtained $A(x)B(x) \neq W(x)G(x)B(x)$. Next, let take the reminder by dividing both sides by $G(x)$ as $(A(x)B(x)) \bmod G(x) \neq (W(x)G(x)B(x)) \bmod G(x)$, then $0 \neq 0$ is a contradiction. Consequently, $(A(x)B(x)) \bmod G(x) \neq 0$, Q.E.D.

Statement 2. If $G(x) = x^c + ... + 1$ then $G(x)W(x) = x^d + ... + x^f + ... + \alpha$, where $c, d, f \in N$, $d > f$, $d = c + \deg(W(x))$, $\alpha \in \{0,1\}$, $W(x)$ is a polynomial.

Proof. Let represent $G(x)$ as sequence of ones. Then multiplication by modulo 2 of $G(x)$ and $W(x)$ may be considered as a modulo 2 addition with a shift: $G(x) \times W(x) = \overline{W(x)}_1 \oplus \overline{W(x)}_2 \oplus ... \oplus \overline{W(x)}_n$, where $\oplus$ is operation of addition by modulo 2, $\overline{W(x)}_i$ are right shifts by i of $W(x)$.

It should be mentioned that the lowest significant bit of the first term and the highest significant bit of the last term are not compensated, therefore, $G(x)W(x)$ represented in the form $x^d + ... + x^f + ... + \alpha$, Q.E.M.

Statement 3. If $G(x) = x^{w-u} + ... + 1$, then for any single error $E(x) = x^i$. $i \in \{w-u, ..., w-1\}$, such that $M_{rp}(x) = M_{ip}(x) + E(x)$, $H(M_{rp}(x)) \neq H(M_{ip}(x))$ is performed.

Proof. Considered conditions $H(M_{rp}(x)) = ((M_{ip}(x) + x^i)x^{w-u}) \bmod G(x)$ and $H(M_{ip}(x)) = ((M_{ip}(x))x^{w-u}) \bmod G(x)$, let's supposed that condition $H(M_{rp}(x)) = H(M_{ip}(x))$ is true. Then equality

$((M_{ip}(x)+x^i)x^{w-u}) \bmod G(x) = (M_{ip}(x)x^{w-u}) \bmod G(x)$ is true as well, from which it is followed that $(x^{i+w-u}) \bmod G(x) = 0$ and, therefore, $x^{i+w-u} = G(x)W(x)$. On the other hand, according to statement 2 $G(x)W(x) = x^d + ... + x^f + ... + \alpha$ and $x^{i+w-u} \neq x^d + ... + x^f + ... + \alpha$. Consequently, $H(M_{rp}(x)) \neq H(M_{ip}(x))$, Q.E.M.

Statement 4. If $G(x) = x^{w-u} + ... + 1$, then for packet type error $E(x) = x^i + ... + x^{i-p+1}$, $p \leq w-u$, $i \in \{w-u+p-1, ..., w-1\}$, for which $M_{rp}(x) = M_{ip}(x) + E(x)$ is true, $H(M_{rp}(x)) \neq H(M_{ip}(x))$ is performed.

Proof. Let's supposed that $H(M_{rp}(x)) = H(M_{ip}(x))$, then $((M_{ip}(x)+x^i + ... + x^{i-p+1})x^{w-u}) \bmod G(x) = (M_{ip}(x)x^{w-u}) \bmod G(x)$, hence $(x^{w-u+i-p+1}(x^{p-1} + ... + 1)) \bmod G(x) = 0$. Each of the factors: $x^{w-u+i-p+1}$ is not evenly divisible by $G(x)$; $x^{p-1} + ... + 1$ also not divisible by $G(x)$, because of the $p-1 < w-u$ and therefore, $(x^{w-u+i-p+1}(x^{p-1} + ... + 1)) \bmod G(x) \neq 0$. It is a contradiction, and hence, $H(M_{rp}(x)) \neq H(M_{ip}(x))$.

Statement 4. If for requisite it is used 8 bits to represent one character, sm_{rp} differ from sm_{ip} by any single transcription and $G(x) = x^{w-u} + ... + 1$, where $w-u \geq 8$, then $a_{rp} \neq a_{ip}$.

Proof. Any single transcription can be represented as $E(x) = x^i + ... + x^{i-p+1}$, where $p \leq 8$. Consequently, in accordance with the statement 3, $H(M_{rp}(x)) \neq H(M_{ip}(x))$ and therefore $a_{rp} \neq a_{ip}$.

Statement 5. If for requisite it is used 8 bits to represent one character, sm_{rp} differ from sm_{ip} by any transposition or double transcription of adjacent characters and $G(x) = x^{w-u} + ... + 1$, where $w-u \geq 16$, then $a_{rp} \neq a_{ip}$.

Proof. Any transposition or double transcription of adjacent symbols can be represented as $E(x) = x^i + ... + x^{i-p+1}$, where $p \leq 16$. Consequently, in accordance with the statement 3, $H(M_{rp}(x)) \neq H(M_{ip}(x))$ and therefore $a_{rp} \neq a_{ip}$.

Considering the third requirement for independent input of two values sm_{ip} and sm_{rp}, let's supposed that all valid requisites sm_{ip} are equal and H uniformly send them to the full range of possible addresses a_{ip}. In this case, each a_{ip} corresponds to $2^{(u-(w-u))}$ sm_{ip}. Then there $2^u 2^u$ options independent of input values sm_{ip} and sm_{rp}, among which:

a) 2^u identical values input options, i.e., ($sm_{ip} = sm_{rp}, a_{ip} = a_{rp}$);

b) $2^u(2^u - 2^{u-(w-u)})$ options, in which errors are detected, i.e. $\left(sm_{ip} \neq sm_{rp}, a_{ip} \neq a_{rp}\right)$;

c) $2^u(2^{u-(w-u)}-1)$ options, in which errors are not detected, i.e. $\left(sm_{ip} \neq sm_{rp}, a_{ip} = a_{rp}\right)$.

Further there are described computations of the probability of different outcomes independent input values sm_{ip} and sm_{rp}. The probability of entering identical values is $p(sm_{ip} = sm_{rp}, a_{ip} = a_{rp}) = \frac{2^u}{2^u 2^u} = \frac{1}{2^u}$. The probability of the case, in which error is detected, is $p(sm_{ip} \neq sm_{rp}, a_{ip} = a_{rp}) = \frac{1}{2^8} - \frac{1}{2^{48}} \approx 0,004\,\text{s}$ $p(sm_{ip} \neq sm_{rp}, a_{ip} \neq a_{rp}) = \frac{2^u(2^u - 2^{u-(w-u)})}{2^u 2^u} = 1 - \frac{1}{2^{w-u}}$. The probability of the case, in which error is not detected, is $p(sm_{ip} \neq sm_{rp}, a_{ip} = a_{rp}) = \frac{2^u(2^{u-(w-u)}-1)}{2^u 2^u} = \frac{1}{2^{w-u}} - \frac{1}{2^u}$. For example, when $w = 56,\ u = 48$ and $w = 64,\ u = 48$ $p(sm_{ip} \neq sm_{rp}, a_{ip} = a_{rp}) = \frac{1}{2^{16}} - \frac{1}{2^{48}} \approx 1,5 \cdot 10^{-5}$.

Let consider now independent input information elements sm_{ip}, sm_{rp} and sm_{sp}, each with equal probability takes any of the valid values and is transformed uniformly on the entire range of possible addresses. As previously, each a_{ip} corresponds to $2^{(u-(w-u))}$ sm_{ip}. Totally, there are $2^u 2^u 2^u$ values for sm_{ip}, sm_{rp} and sm_{sp} inputs, among which:

– 2^u identical values input options, i.e., ($sm_{ip} = sm_{rp} = sm_{sp}, a_{ip} = a_{rp} = a_{sp}$);

– the cases, in which errors are not detected, are following:

a) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} \neq a_{rp},\ a_{ip} \neq a_{sp},\ a_{rp} = a_{sp};$ b) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} \neq a_{rp},\ a_{ip} = a_{sp},\ a_{rp} \neq a_{sp};$.

c) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} \neq a_{sp},\ a_{rp} \neq a_{sp};$, d) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} = a_{sp},\ a_{rp} = a_{sp};$

e) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} = sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} = a_{sp},\ a_{rp} = a_{sp};$ f) $sm_{ip} \neq sm_{rp}, sm_{ip} = sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} = a_{sp},\ a_{rp} = a_{sp};$

g) $sm_{ip} = sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} = a_{sp},\ a_{rp} = a_{sp}.$

i.e. cases a, b, c include $2^u(2^u - 2^{u-(w-u)})(2^{u-(w-u)} - 1)$ options; g – $2^u(2^{u-(w-u)} - 1)(2^{u-(w-u)} - 2)$ options; e, f – $2^u(2^{u-(w-u)} - 1)$ options.

– the cases, in which errors are detected, are following:

a) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} \neq a_{rp},\ a_{ip} \neq a_{sp},\ a_{rp} \neq a_{sp};$ b) $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} = sm_{sp},$ $a_{ip} \neq a_{rp},\ a_{ip} \neq a_{sp},\ a_{rp} = a_{sp};$

c) $sm_{ip} \neq sm_{rp}, sm_{ip} = sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} \neq a_{rp},\ a_{ip} = a_{sp},\ a_{rp} \neq a_{sp};$ d) $sm_{ip} = sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp},$ $a_{ip} = a_{rp},\ a_{ip} \neq a_{sp},\ a_{rp} \neq a_{sp},$

i.e. case a includes $2^u(2^u - 2^{u-(w-u)})(2^u - 2^{u-(w-u)} - 2^{u-(w-u)})$ options; cases b, c, d – $2^u(2^u - 2^{u-(w-u)})$ options.

Further there are described computations of probabilities of the different diagnoses with independent input values requisites sm_{ip}. sm_{rp} and sm_{sp}. The probability of entering identical values of requisites is equal to $\frac{2^u}{2^{3u}} = \frac{1}{2^{2u}}$. The probability of the case, when the error are skipped, is equal to $\frac{3 \cdot 2^u (2^u - 2^{u-(w-u)})(2^{u-(w-u)} - 1)}{2^{3u}} +$ $+ \frac{2^u (2^{u-(w-u)} - 1)(2^{u-(w-u)} - 2) + 3 \cdot 2^u (2^{u-(w-u)} - 1)}{2^{3u}} = \left(\frac{1}{2^{w-u}} - \frac{1}{2^u}\right)\left(3 - \frac{1}{2^{w-u-1}} + \frac{1}{2^u}\right)$. The probability of error detection by comparing indices obtained for the three requisites is equal to $\frac{2^u (2^u - 2^{u-(w-u)})(2^u - 2^{u-(w-u)} - 2^{u-(w-u)})}{2^{3u}} + \frac{3 \cdot 2^u (2^u - 2^{u-(w-u)})}{2^{3u}} =$ $= \left(1 - \frac{1}{2^{w-u}}\right)\left(1 - \frac{1}{2^{w-u-1}} + \frac{3}{2^u}\right)$. For example, when $w = 64,\ \ u = 48$, the probability of the case, when an error is skipped, $p\left[a \vee b \vee c \vee d\right] =$ $= \left(\frac{1}{2^{16}} - \frac{1}{2^{48}}\right)\left(3 - \frac{1}{2^{15}} + \frac{1}{2^{48}}\right) \approx 4,5 \cdot 10^{-5}$.

Considering the fourth requirement in the case of independent input of two values sm_{ip} and sm_{rp}, it should be calculated probability of case, if $a_{ip} = a_{rp}$, then $sm_{ip} = sm_{rp}$. The Bayes' formula [7] allows to calculate posteriori conditional probability of the presence of unconditional priori one.

Let the event ER_1 is equal $sm_{ip} = sm_{rp}$, event ER_2 – $sm_{ip} \neq sm_{rp}$, event EI – $a_{ip} = a_{rp}$. Then $p(ER_1 \mid EI) = \frac{p(ER_1)\, p(EI \mid ER_1)}{\sum_{i=1}^{2} p(ER_i)\, p(EI \mid ER_i)}$. Let's supposed, that sm_{ip} and sm_{rp}, which consist of L characters, are independently entered by two human-operators based on the same original document. Then $p(ER_1)$ can be calculated as the probability of error-free entry of two requisites, i.e. $p(ER_1) = (1 - \pi_c)^{2L}$, where π_c is the possibility of mistakes in the human information (errors per symbol), hence $p(ER_2) = 1 - (1 - \pi_c)^{2L}$. For example, for $L = 6$ and $\pi_c = 10^{-2}$ $p(ER1) \approx 0,88$. It is obvious that $p(EI \mid ER_1) = p(a_{ip} = a_{rp} \mid m_{ip} = m_{rp}) = 1$, as, for equal requisites $(M_{ip}(x) = M_{rp}(x))$ it is impossible to obtain different indexes

$(H(M_{ip}(x)) \neq H(M_{rp}(x)))$,according to (3). For calculations of $p(EI \mid ER_2) = p(a_{ip} = a_{rp} \mid sm_{ip} \neq sm_{rp})$ it is possible to use the fact of equal probabilities of all admissible sm_{ip}, sm_{rp} and their uniform mapping to the corresponding ranges a_{ip}, a_{rp}.

According to the formula of conditional probability [7], $p(a_{ip} = a_{rp} \mid sm_{ip} \neq sm_{rp}) = \dfrac{p(a_{ip} = a_{rp}, sm_{ip} \neq sm_{rp})}{p(sm_{ip} \neq sm_{rp})} = \dfrac{2^{u-(w-u)} - 1}{2^u - 1}$. Then $p(ER_1 \mid EI) = \dfrac{(1-\pi_c)^{2L} \cdot 1}{(1-\pi_c)^{2L} \cdot 1 + (1-(1-\pi_c)^{2L})\left(\dfrac{2^{u-(w-u)} - 1}{2^u - 1}\right)}$. For example, for $u = 48$, $w = 64$, $L = 6$ $p(ER_1 \mid EI) \approx 0,999998$.

Considering the fourth requirement in the case of independent input of three values sm_{ip}, sm_{rp} and sm_{sp}, it is necessary to calculate the probability of $sm_{ip} = sm_{rp} = sm_{sp}$, when $a_{ip} = a_{rp} = a_{sp}$. Let the event $E3R_1$ is equal values $sm_{ip} = sm_{rp} = sm_{sp}$, event $E3R_2$ – $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} = sm_{sp}$, event $E3R_3$ – $sm_{ip} \neq sm_{rp}, sm_{ip} = sm_{sp}, sm_{rp} \neq sm_{sp}$, event $E3R_4$ – $sm_{ip} = sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp}$, event $E3R_5$ – $sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp}$, event $E3I$ – $a_{ip} = a_{rp} = a_{sp}$. Then $p(E3R_1 \mid E3I) = \dfrac{p(E3R_1)\, p(E3I \mid E3R_1)}{\sum_{i=1}^{5} p(E3R_i)\, p(E3I \mid E3R_i)}$. Let calculate the a priori probability of events $E3R_i, i = \overline{1,5}$ according to the binomial law, assuming independence of errors in separate characters, as following: $p(E3R_1) = (1-\pi_c)^{3L}$. $p(E3R_2) = p(E3R_3) = p(E3R_4) = (1-\pi_c)^{2L}(1-(1-\pi_c)^{L})$, $p(E3R_5) = (1-(1-\pi_c)^{L})^3$. For example, for $L = 6$ and $\pi_c = 10^{-2}$ $p(E3R_1) \approx 0,83$, $p(E3R_2) \approx 0,05$, $p(E3R_5) \approx 0,0002$.

As previously, calculation of the conditional probabilities $p(E3I \mid E3R_i)$ is based on the conditions of the equal probability of all admissible sm_{ip}, sm_{rp} and sm_{sp} and uniformity of transformation to corresponding ranges a_{ip}, a_{rp} and a_{sp}. So, $p(E3I \mid E3R_1) = 1$ due to the fact that the mapping H each value sm_{ip} sends to no more than one a_{ip} and hence H is a function. The conditional probabilities of coincidence of codes in the case of only one failed requisite are following: $p(E3I \mid E3R_2) = p(E3I \mid E3R_3) = p(E3I \mid E3R_4) =$

$= \dfrac{p(a_{ip} = a_{rp} = a_{sp}, sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} = sm_{sp})}{p(sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} = sm_{sp})} = \dfrac{2^{u-(w-u)} - 1}{2^u - 1}$. The conditional probability of indices coincidence in the case of three different requisites input

by human- operator is following: $p(E3I \mid E3R_5) = \frac{p(a_{ip} = a_{rp} = a_{sp}, sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp})}{p(sm_{ip} \neq sm_{rp}, sm_{ip} \neq sm_{sp}, sm_{rp} \neq sm_{sp})} =$

$= \frac{(2^{u-(w-u)} - 1)(2^{u-(w-u)} - 2)}{(2^u - 1)(2^u - 2)}$. Thus, the posterior probability of identity requisites $sm_{ip} = sm_{rp} = sm_{sp}$ provided that $a_{ip} = a_{rp} = a_{sp}$, can be calculated as $p(E3R_1 \mid E3I) =$

$$= \frac{(1-\pi_c)^{3L} \cdot 1}{(1-\pi_c)^{3L} \cdot 1 + 3 \cdot (1-\pi_c)^{2L} (1-(1-\pi_c)^L) \cdot \frac{2^{u-(w-u)} - 1}{2^u - 1} + (1-(1-\pi_c)^L)^3 \cdot \frac{(2^{u-(w-u)} - 1)(2^{u-(w-u)} - 2)}{(2^u - 1)(2^u - 2)}}.$$

For example, for $u = 48, w = 64, L = 6$ $p(E3R_1 \mid E3I) \approx 0,999997$.

The standard CRC-CCITT polynomial $G_1(x) = x^{16} + x^{12} + x^5 + 1$ and CRC-16 - $G_2(x) = x^{16} + x^{15} + x^2 + 1$ are commonly used to increase the reliability of information transmission in computer networks [6]. It is obvious that they satisfy the first and second requirements, as $deq(G_i(x)) > 0$ and $G_i(x) = x^{w-u} + ... + 1$, where $w - u \geq 16$. $i = \overline{1,2}$. Furthermore, they may be represented as a product of polynomials of lower degree, for example, $G_1(x) = (x+1) \cdot (x^{15} + x^{14} + x^{13} + x^{12} + x^4 + x^3 + x^2 + x + 1)$, and are not irreducible. Therefore, the codes constructed based on the $G_1(x)$ and $G_2(x)$ does not refer to cyclic, but inherit all the capabilities of error detection, the inherent cyclic codes, including the ability of uniform mapping the possible keys sm_{ip}, sm_{rp} ..., sm_{sp} to the corresponding ranges a_{ip}, a_{rp} ..., a_{sp}. Therefore, assuming that each of the elements sm_{ip}. sm_{rp} ..., sm_{sp} with equal probability takes any of the permissible values, then $G_1(x)$ and $G_2(x)$ satisfies the third and fourth requirements.

Choosing the best alternative was carried out using the method of weighted sum. Natural when forming the weighting factors will have an idea of ranking weights according densities classes most common error.

Thus, the code based on the polynomial $G_1(x) = x^{16} + x^{12} + x^5 + 1$ will have the best total controlling ability relative to the most common classes of errors in the data on the names of employees of the KhAI University

4 Diagnostic Data Model

According to the signal-parametric approach to control systems diagnostic [4,8], the diagnostic models are defined as mathematical constructions linking indirect signs with direct reasons of the fault. In our case, diagnostic data model (DMD) is named a mathematical construction that relates indirect indications of the data lines with errors, the DMD must be of the form

$$\Delta D = \tilde{D} - \hat{D}, \tag{4}$$

where ΔD is an indirect indication of the presence of failed data; $\tilde{D}$, $\hat{D}$ are direct functions of signs of error and the reference data, respectively. For any DMD, the conditions of diagnosability must also be fulfilled, i.e. the possibility of an unambiguous establishment of the fact of the presence of failed data on indirect signs.

Let's create the DMD to identify and search for a place of failed requisites in the multiset M_ρ. Let $A_\rho = \{a_{1\rho}, a_{2\rho}, ..., a_{q\rho}\}$ be multiset indices calculated for the initial requisites, and $G(x) = x^{16} + x^{12} + x^5 + 1$, let D be row vector of dimension $[0,..., 2^{16} - 1]$ such that $D[a_{i\rho}] = |A_\rho \cap A_{i\rho}|$ where $A_{i\rho} = \{\underbrace{a_{i\rho}, a_{i\rho}, ..., a_{i\rho}}_{q-раз}\}$. Then the equation, characterized by the absence of failed requisites in M_ρ will have the form $\hat{D}[a_{1\rho}] = q$, i.e. all indexes are the same. If, however, M_ρ contains failed requisites, the $\tilde{D}[a_{1\rho}] = |A_\rho \cap A_{1\rho}|$. Thus, the DMD to detect failed data in M_ρ looks as:

$$\Delta_{\det} D_\rho = \tilde{D}[a_{1\rho}] - \hat{D}[a_{1\rho}] = |A_\rho \cap A_{1\rho}| - q, \tag{5}$$

where $\Delta_{\det} D_\rho$ is an indirect indication of the presence of failed data in M_ρ. If $\Delta_{\det} D_\rho \equiv 0$ then M_ρ is error-free, or M_ρ contains failed information.

To find a place in the wrong requisite M_ρ DMD will be as follows:

$$\Delta_{pl} D_{i\rho} = \tilde{D}[a_{i\rho}] - \hat{D}[a_{i\rho}] = |A_\rho \cap A_{i\rho}| - \frac{q}{2} - \alpha, \tag{6}$$

where $\Delta_{pl} D_{i\rho}$ is an indirect indication of the presence of failed data in the requisite $sm_{i\rho}$. $\alpha \in [\frac{1}{2}; \frac{q}{2} - 1]$, and if $\Delta_{pl} D_{i\rho} < 0$, then $sm_{i\rho}$ is faulty requisite, otherwise $sm_{i\rho}$ is a reference requisite.

The performance of the method of index-requisite diagnosis was evaluated. In this case the first stage is filling a row vector D. It can be assumed to be proportional to the value q. It is assumed that the calculation of indices occur before data cleaning process. As for the time of the second stage, it coincides with the time of the second stage in the case of pairwise comparisons requisites. Maximum wait time for diagnostic procedures on the basis of the method of index-requisite diagnosis - $T_{общ.инд.рекв} \approx 2 * q * t_{ср.в}$. The overall performance of the method of index-requisite diagnosis of redundant information in $\frac{(q-1) * L + 2}{4}$ times higher than the performance of the method based on pairwise comparison requisites. For example, when $q = 3$ and $L = 8$ $\frac{T_{общ.поп.ср}}{T_{общ.инд.рекв}} \approx 4.5$.

5 Conclusion

Deep diagnostics data is the basis for the following problem solution of data recovery. Determining, based on the principle of majority, error and reference values for each attribute it is possible automatic replacement of standard errors. In addition, the failed attributes should be corrected in the source subsystem. Since the change in the original data in the data warehouse is technically impossible, human-operator should be informed about the error occurred to ensure the quality of subsystem data. Such notification must include the failed attribute, reference attribute, as well as the record ID, for example, last name, first name, etc. If the source subsystem allows working with a clipboard, the failed value could be replaced by correct one automatically.

If it is impossible to find the reference and failed values for the attribute, for example, if there are two different requisites and diagnostic model cannot detect the place, it is concluded that both requisites are incorrect. Decision-making is entrusted to the system administrator, which can redirect the problem to the operators.

References

1. Chukhray, A., Havrylenko, O.: Proximate Objects Probabilistic Searching Method. Advances in Intelligent Systems and Computing, 1113 AISC, 219-227 (2020).
2. Cormen, T., Leiserson, C., Rivest, R., Stein, C.: Introduction to Algorithms. 3rd edn. The MIT Press, 1292 p., (2009).
3. Borda, M.: Fundamentals in Information Theory and Coding. 2011th edn, Springer, 485 p. (2011).
4. Kulik, A.: Rational intellectualization of the aircraft control: Resources-saving safety improvement. Studies in Systems, Decision and Control, 173-192 (2017).
5. Martínez Bastida, J.P., Havrylenko, O., Chukhray, A.: Developing a self-regulation environment in an open learning model with higher fidelity assessment. Communications in Computer and Information Science, 826, 112-131 (2018).
6. Tanenbaum, A., Wetherall, D.: Computer Networks, 5 edn, Pearson, 960 p. (2012).
7. Ghahramani, S.: Fundamentals of Probability. With Stochastic Processes. 4th edn, CRC Press, (2018).
8. Martínez Bastida, J.P., Gavrilenko, E.V., Chukhray, A.G.: Developing a pedagogical intervention support based on Bayesian networks. CEUR Workshop Proceedings, 1844, 265-272 (2017).