Journal of Advanced Zoology

ISSN: 0253-7214

*Volume **44** Issue **S-2 Year 2023** Page **1949:1954***

A Recognition Model and an Algorithm for Calculating Values for the Classification and Coding of Documents in Organizational Management Systems

Sultonjon A. Tishlikov

Gulistan State University, Gulistan, Uzbekistan

Article History

Received: 13 June 2023
Revised: 12 September 2023
Accepted: 21 September 2023

Abstract

In this paper, a logical-mathematical model of the task of analyzing the composition and structure of documents by quantitative and qualitative features, as well as a recognition model for classifying and coding documents by their composition and structure, has been developed. A meaningful mathematical formulation of the problem and functional characteristics of the developed automated recognition complex is presented.

Keywords. *Document, information, algorithm, model, authenticity, coding, automation, recognition, redundancy, vector, requisites, pattern recognition, structural redundancy, technological redundancy.*

1. Introduction

In solving problematic management problems in information systems, a large number of documents are transferred for processing, and in dozens of copies, which are formed in databases. Any document during storage and placement can be represented as a matrix of messages, where the rows describe the types of the object and document, and the columns describe the indicators and details of documents that carry information about quantitative and qualitative values.

It should be noted that when documents of the same type are used in solving a specific problem, it is possible to trace the number of repeated names and values of indicators and details. Such structural and technological redundancy contained in the document can be used in the construction of algorithms for logical control of the reliability of information and control of built-in expert systems and databases.

On the other way, prerequisites are being formed for solving the problems of classifying and recognizing documents according to their quantitative and qualitative characteristics.

In this article, we have developed a logical-mathematical model for the problem of analyzing the composition and structure of documents by quantitative and qualitative characteristics, as well as a recognition model for classifying and coding documents by their composition and structure. A meaningful mathematical formulation of the problem and functional characteristics of the developed automated recognition complex is presented.

The solution of a set of problems is modeled as a cybernetic system. The input of this system is formalized as a transformed matrix $Y=\{X^I_J\}$, where i=1, m-are rows, j=1, n- are columns, which describe the characteristics of the document in the form of vectors.

Grouping features are formed as vectors, $\gamma = (\gamma_1, \gamma_2, \ldots \gamma_{n_1})^T$, $n_1<n$ and $\beta = (\beta_1, \beta_2, \ldots \beta_{n_2})$, $n_2<n$.

Here x^i_j is the value of the j-th attribute of the i-th document; m- is the number of documents; n - the number of features that describe the characteristics of the rows or columns of the document;γ_q and β_r components of the vectors γ and β, which are the qth and rth grouping features..

The output of the system Y' is obtained using the operator L, which transforms the input of the system Y into the matrix Y', i.e.

L(Y, γ, β}⇒Y'.

The output of the system has the form

Y'={$\gamma_q\beta_r$}, q=1,n, r=1, n_2.

Here, each element of the matrix γ_q β_2 is the number of (document) rows or columns of a document with grouping characteristics q and r.

The elements of the Y' matrix are summarized by columns and rows. It turns out the final row vector W=(W_1,...W_{n2}), where Wr= $\sum_{r=1}^{n_1} \lambda_q \beta_r$ and the final column vector σ = (σ1 ,…, σ n1)^T.

It has been established that for solving the formulated problems, methods of pattern recognition based on the calculation of estimates [1], which classify documents according to a set of given features, are effective.

The developed recognition system includes the following functional blocks:

1. Formation of a system of indicators describing the characteristics of the document;

2. Formation of reference tables - T_1^e and T_2^e, respectively, for the recognition task, as well as the control table T^k;

3. Comparison of the elements of the control table Tk with the reference tables T^e_1 and T^e_2 and determining the degree of compliance of the characteristics of the document with the reference classes. Let the r-th system consist of the set M={μ_1,…, μ_t,… μ_k} of documents, and each μt-th document includes the set S= {S_1, ..., S_j, ... , $S_{\mu t}$ }, t= 1,k rows.

$T_{k\mu_t n} = ||x_{hi}^{(t)}||$,t=1,k, h=1,$m_i$, i=1,n.

Each document included in the system μt is characterized by the values of the feature set X= {x_1,...,x_i…,x_n}. The values of features will be denoted by x_{ji}^t, j=1,μt; i=1,n. Then the sequence of rows $S^{(t)}_1$,…, $S^{(t)}_j$, ... $S^{(t)}_{\mu t}$formsacontroltable

$T_{k\mu_t n} = ||x_{hi}^{(t)}||$,t=1,k, h=1,$m_i$, i=1,n.

Let a table of standards be compiled on the basis of a priori given information

$T_{hq_u n} = ||x_{hi}^{(u)}||$. U=1,L, h=1,$q_u$, i=1,n

And each set of lines of the form

{$x_{h1}^{(1)}$,…, $x_{hi}^{(1)}$,…, $x_{hn}^{(1)}$} h=1,q.

{ $x_{h1}^{(2)}$,…, $x_{hi}^{(2)}$,…, $x_{hn}^{(2)}$} h=1,q_2

{ $x_{h1}^{(l)}$,…, $x_{hi}^{(2)}$,…, $x_{hn}^{(2)}$} h=1,q_l

describes the qualitative characteristics of an artificially idealized document.

It is required to determine the class K_u, U=1,I to which the document $S^{(t)}_j$ belongs for the assessment task, and the system of features μ_t, $\mu_t \in$ M for comparative analysis.

Therefore, the following problem is posed. Find an algorithm A that provides recognition (assignment to one of the Ki classes) of the object S_i and the system of objects {S_1, ...S_j, ..., S_{mt}} == μ_t .

A modified version of the algorithm for calculating estimates has been developed, consisting of seven stages.

The first stage of setting Algorithm A is to identify subsets from the set {1,2,..., n}, i.e. identification of system Ω_A. The elements of system Ω_A are called support sets, and Ω_A itself is called the system of support sets of algorithm A.

The set Ω_A is finite. Its cardinality is equal to C^n_k, where n - is the number of columns of the table $T_{lq_u n}$, and k - is the length of the voting set, consisting of an integer equal to the number of columns contained in all possible support sets Ω. The value k- is the first parameter of the driving algorithm.

The second stage is the determination of the proximity function r(ωs, ωs_h,) between the strings ωs and ωs_h, where corresponds to $\Omega \subset \Omega_A$. Here, ωs and ωs_h are part of the allowed strings S and S_h, where S_h is the reference description string of the h-th document (h=1, q, U=1,l), and S is an input vector that carries information about the estimated row ((or column)).

Different models consider different proximity functions. We have used the following two types:

1) features take values from the binary alphabet, then

$$r(\omega s, \omega s_h) = \begin{cases} 1 & если\ \omega s = \omega s_h \\ 0 & если\ \omega s \neq \omega s_h \end{cases}$$

2) given $\omega s = (\alpha_1, \ldots, \alpha_n)$, $\omega s_h == (\beta_1, \ldots, \beta_n)$, i.e. features take values from an arbitrary alphabet, are positive numbers, ε_i and ε_i p(s,s_h) is the number of failed inequalities of the form $|\alpha_i - \beta_i| \leq \varepsilon_i$, i= 1, 2, ... ,n. Then

$$r(\omega s, \omega s_h) = \begin{cases} 1 & если\ \rho(s, s_h) \leq \varepsilon, \\ 0 & если\ \rho(s, s_h) > \varepsilon. \end{cases}$$

At the third stage, the row score for the fixed reference set $T\omega(s, s_h)$. is calculated. The estimate is determined by the value of the proximity function r(ωs, ωs_h). At the stage of calculating estimates, the "external parameters" of the S_h line can be taken into account.

This model uses parameters γ_h - the degree of importance or representativeness of rows S_h of the reference table. Then the desired estimate can be written in the form
$\Gamma\omega(s,s_n) = \gamma(s_h), r(\omega s, \omega s_h)$.

At the fourth stage, an estimate is calculated for a fixed reference set. $\Gamma^u\omega(s)$:

$$\Gamma^u\omega(s) = \varphi[\Gamma\omega(s,s_1), \Gamma\omega(s,s_2), \ldots, \Gamma\omega(s,s_h), \ldots, \Gamma\omega(s, s_{q_u})].$$

The function φ is given $\Gamma^u_w = \sum_{h=1}^{q_U} \Gamma_\omega(s, s_h)$. where q_u is the number of objects of the U-th class.

At the fifth stage, the estimate $\Gamma_u(s)$ for the class K_u is calculated using the system of support sets. It is defined as a function of scores $\Gamma^u\omega$ (s) for various subsets Ω:

$$\Gamma_u(s) = \sum_{\{\omega\Omega \subset \Omega_A\}} \Gamma^u\omega(s) = \sum_{\{\omega\Omega \subset \Omega_A\}} \sum_{h=1}^{q_u} \Gamma_\omega(s, s_h).$$

The recognition operator RA, using these five stages of the algorithm, calculates the evaluation matrix $\{\Gamma_{ju}\}_{m*l}$ for solving evaluation problems.

At the sixth stage, the total estimate $\Gamma_U(\mu_t)$ for the class K_U is calculated for the fixed sets of rows

$$\mu_t = (s_1^{(t)}, \ldots, s_i^{(t)}, \ldots, s_{\mu_t}^{(t)}), \quad t=1,k.$$

Let the quantities be calculated.
$\Gamma_U(s_1^{(t)},\ldots,)\ldots$, $\Gamma_U(s_j^{(t)}),\ldots$, $\Gamma_U(s\,\mu_t^{(t)})$, those. the matrix of estimates is given $\{\Gamma_{ju}\}_{m*t}$. Then the calculated total score is defined as a function

$$\Gamma_U(\mu_t)=\varphi[\Gamma_U(s_1^{(t)}),\ldots,\Gamma_U(s_j^{(t)}),\ldots,\Gamma_U(s\,\mu_t^{(t)})]=\varphi[\{\Gamma_{ju}\}\,\mu_{t*l}],$$

and the function φ looks like

$$\Gamma_U(\mu_t)=\sum_{j=1}^{m_t}\Gamma_U\left(s_i^{(t)}\right),$$

or, according to the fifth step,

$$\Gamma_U(\mu_t)=\sum_{j=1}^{m_t}\sum_{\{\omega:\Omega\subset\Omega_A\}}\sum_{h=1}^{q_U}\Gamma_\omega\left(s_j^{(t)},s_h\right)=\sum_{\{\omega:\Omega\subset\Omega_A\}}\sum_{j=1}^{m_t}\sum_{h=1}^{q_U}\Gamma_\omega\left(s_j^{(t)},s_h\right).$$

Thus, at the sixth stage, the matrix of the total assessment $\{\Gamma_{tu}\}_{kxl}$ is calculated to solve the problems of comparative analysis.

At the seventh stage, the decision rule r_A of algorithm A is developed.

Let us assume that the quantities $\Gamma_1(s)$, $\Gamma_2(s),\ldots,\Gamma_l(s)$, $\Gamma_2(\mu_t,\ldots,\Gamma_l(\mu_t)$, i.e. estimates of the input vector S, respectively, have been calculated.

If $r(\Gamma_1,\ldots,\Gamma_l)=U$, $1\le U\le l$.l. then the string S or the set of strings are μ_t classified as belonging to the class K_U.

The decision rule rA in the case of using a combination of rows and a set of rows is represented as follows:

$$r\left(\{\Gamma_{ju}\}_{\mu\times l}\right)=\{\alpha^A{}_{ju}\}_{\mu\times l};\ r\left(\{\Gamma_{tu}\}_{k\times l}\right)=\{\alpha^A{}_{tu}\}_{k\times l}$$

Here $\{\alpha^A{}_{ju}\}_{\mu\times l}$ *and* $\{\alpha^A{}_{ju}\}_{k\times l}$ *are matrices of information vectors of objects* $s_1,\ldots,s_2,\ldots,s_j,\ldots$, *respectively, and systems of objects* $\mu_1,\mu_2,\ldots,\mu_k$, *built by algorithm A.* To solve a complex of problems of comparative analysis, a modified version of the pattern recognition algorithm based on the calculation of estimates, and an algorithm for calculating the total estimate of the characteristics of a document, have been developed.

2. References

1. Путькина Л.В. Особенности использования электронного документооборота для эффективной работы современного предприятия // Nauka-rastudent.ru. – 2016. – No. 01 (25) / [Электронный ресурс] – Режим доступа. – URL: http://nauka-rastudent.ru/25/3173/
2. Бобылева М.П. Управление документальными коммуникациями и информационными ресурсами организации: концептуальные подходы // Проблемы теории и практики управления. 2011. № 5. С. 116-126.
3. Монахов, М. Ю., Семенова, И. И., Полянский, Д. А., Монахов, Ю. М. Особенности среды обеспечения достоверности информации в информационно-телекоммуникационных системах // Фундаментальные исследования. - 2014. - № 9.
4. Келдыш Н. В.Обеспечение заданного уровня безопасности при решении функциональных задач ведомственного электронного документооборота. // Научно-технический сборник НИИ ТП № 4. – М. 2013. – С. 118-123.
5. Бессонов С.В.Оптимизация электронного документооборота в корпоративных системах: Дис. . канд. экон. наук. Москва. 2001. 187с.

6. Гудов А. М.Об одной модели оптимизации документопотоков, реализуемой при создании системы электронного документооборота // Вычислительные технологии. -2006. - том 11, специальный выпуск. - С. 53 - 65
7. Jumanov I.I., Karshiev Kh. B Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management systems // «Chemical technology. Control and management» (WCIS-2018)– P.146-150
8. Jumanov I.I., Tishlikov S. A. Method of stochastic search in the system to monitoring and correcting of spelling in electronic texts // 2013 International Conference in Central asia on Internet (ICI), Tashkent, 8-10 october 2013, Section 7, IEEE. – Tashkent, 2013.
9. Jumanov I.I., Akhatov A.R. Fuzzy Semantic Hypernet for Information Authenticity Controlling in Electronic Document Circulation Systems // 4-th International Conference on Application of Information and Communication Technologies, 12-14 october 2010, Section 2, IEEE. - Tashkent, 2010. - p.21-25.
10. Jumanov I.I., Tishlikov S. A. Control of integrity and authenticity of electronic documents on the basis of genetic principles of tests formation and generation // In proceedings of the Eight World Conference on Intelligent Systems for Industrial Automation, 25-27 November, 2014. – Tashkent, Uzbekistan, 2014. – P.242-246
11. Jumanov I.I., Karshiev Kh. B., Tishlikov S.A Examination of the Efficiency of Algorithms for Increasing the Reliability of Information on Criteria of Harness and the Cost of Processing Electronic Documents. International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019, ISSN: 2277-3878
12. Жуманов И.И., Ахатов А.Р. Оценка эффективности программного комплекса контроля достоверности текстовой информации систем электронного документооборота // «Химическая технология. Контроль и управление». № 2, С. 46-52
13. Jumanov I.I., Karshiev Kh. B Expanding the possibilities of instruments to improve the information reliability of electronic documents of industrial management systems // «Chemical technology. Control and management» (WCIS-2018)– P.146-150
14. Akhatov A.R., Jumanov I.I. Improvement of text information processing quality in documents processing systems // 2nd IEEE/IFIP International Conference In Central Asia On Internet ICI-2006, September 19-21, International Hotel Tashkent, Uzbekistan.
15. Akhatov A.R., Jumanov I.I., Djumanov O.I. An Effective Quality Control of Textual Information on the Basis of Statistical Redundancy in Distributed Mobile IT Systems and e-Applications //3-d International Conference in Central Asia on Internet, Tashkent, 2007.
16. JumanovI.I., KarshievKh. B. Оптимизация достоверности информации на основе базы электронных документов и особенностей правил контроля базы знаний // Проблемы вычислительной и прикладной математики, 2019 - № 3.С. 57-74.
17. Ахатов А.Р., Тишликов С.А. Методы повышения достоверности передачи и обработки информации в системах электронного документооборота на основе нечеткой семантической гиперсети // Научный журнал «Проблемы вычислительной и прикладной математики», Ташкент. - №3(5), 2016. – с. 10-20
18. Ney H. and Kneser R. Improved clustering techniques for class-based statistical language modelling. In European Conference on Speech Communication and Technology (Eurospeech), 1993. pages 973–976, Berlin.
19. Kemal Oflazer, Sergei Nirenburg and Marjorie McSchane.Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning // Computational Linguistics, 2001. Vol: 27, No: 1.59-85.

20. Dilek Hakkani-Tür, Kemal Oflazer, Gökhan Tür. Statistical Morphological Disambiguation for Agglutinative Languages. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), August, 2000.