ISSN 0555-2656

БИОНИКА ИНТЕЛЛЕКТА

ИНФОРМАЦИЯ, ЯЗЫК, ИНТЕЛЛЕКТ

№ 1 (92) 2019

НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

Основан в октябре 1967 г.

Учредитель и издатель
Харьковский национальный университет радиоэлектроники

Периодичность издания – *2 раза в год*

Харьков • ХНУРЭ • 2019

Научно-технический журнал

«БИОНИКА ИНТЕЛЛЕКТА»

ISSN 0555-2656

Основан Харьковским национальным университетом
радиоэлектроники в 1967 году

Реферирование и индексирование:

Google Scholar

Microsoft Academic

ACADEMIA

BASE

INDEX COPERNICUS
INTERNATIONAL

ResearchGate

Національна бібліотека України
імені В. І. Вернадського

Журнал включен в список научных специализированных изданий Украины
по техническим и физико-математическим наукам
согласно приказа Министерства образования и науки Украины № 820 от 11.07.2016

УДК 004.8: 004.032.26

Є.В. Бодянський [1] , Т.Є. Антоненко [2]

[1] Професор кафедри штучного інтелекту, ХНУРЕ, м. Харків, Україна, yevgeniy.bodyanskiy@nure.ua;

[2] Аспірант, ХНУРЕ, м. Харків, Україна, tymofii.antonenko@nure.ua

ГЛИБОКА НЕО-ФАЗЗІ НЕЙРОННА МЕРЕЖА ТА ЇЇ НАВЧАННЯ

Оптимізація швидкодії навчання глибоких нейронних мереж є надзвичайно актуальним питанням. Сучасні підходи орієнтуються на використання нейронних мереж на основі персептрону Розенблатта. Але отримувані результати не являються задовільними для індустріальних та наукових потреб в контексті швидкодії навчання нейронних мереж. Також такий підхід натикається на проблеми зникаючого та вибухаючого градієнта. Для вирішення проблеми в статті запропоновано використовувати нео-фаззі нейрон, властивості якого основані на F-перетворенні. В статті розглянуто використання нео-фаззі нейрона як основного компонента нейронної мережі. Показана архітектура глибокої нео-фаззі нейронної мережі а також алгоритм зворотнього поширення похибки для цієї архітектури з трикутною функцією принадлежності для нео-фаззі нейрона. Приведені основні переваги щодо застосування нео-фаззі нейрона як основного компоненту нейронної мережі. В статті описано за рахунок яких властивостей нео-фаззі нейрона вирішуються питання покращення швидкодії та зникаючого чи вибухаючого градієнта. Порівняно запровоновану архітектуру нео-фаззі глибокої нейронної мережі зі стандартними глибокими мережами на основі персептрону Розенблатта.

НЕО-ФАЗЗІ НЕЙРОН, БАГАТОШАРОВА НЕЙРОННА МЕРЕЖА, F-ПЕРЕТВОРЕННЯ

Е.В. Бодянский, Т.Е. Антоненко. Глубокая нео-фаззи нейронная сеть и ее обучение. Оптимизация быстродействия обучения глубоких нейронных сетей является чрезвычайно актуальным вопросом. Современные подходы ориентируются на использование нейронных сетей на основе персептрона Розенблатта. Но получаемые результаты не являются удовлетворительными для индустриальных и научных потребностей в контексте быстродействия обучения нейронных сетей. Также такой подход натыкается на проблемы исчезающего и взрывающегося градиента. Для решения проблемы в статье предложено использовать нео-фаззи нейрон, свойства которого основаны на F-преобразовании. В статье рассмотрено использование нео-фаззи нейрона как основного компонента нейронной сети. Показана архитектура глубокой нео-фаззи нейронной сети, а также алгоритм обратного распространения ошибки для этой архитектуры с треугольной функцией принадлежности для нео-фаззи нейрона. Приведены основные преимущества по применению нео-фаззи нейрона как основного компонента нейронной сети. В статье описано за счет каких свойств нео-фаззи нейрона решаются вопросы улучшения быстродействия и исчезающего или взрывающейся градиента. Проведено сравнение предложенной архитектуры нео-фаззи глубокой нейронной сети со стандартными глубокими сетями на основе персептрона Розенблатта.

НЕО-ФАЗЗИ НЕЙРОН, МНОГОСЛОЙНАЯ НЕЙРОННАЯ СЕТЬ, F-ПРЕОБРАЗОВАНИЕ

Ye. Bodyankiy. T. Antonenko. Deep neo-fuzzy neural network and its learning. Optimizing the learning speed of deep neural networks is an extremely important issue. Modern approaches focus on the use of neural networks based on the Rosenblatt perceptron. But the results obtained are not satisfactory for industrial and scientific needs in the context of the speed of learning neural networks. Also, this approach stumbles upon the problems of a vanishing and exploding gradient. To solve the problem, the paper proposed using a neo-fuzzy neuron, whose properties are based on the F-transform. The article discusses the use of neo-fuzzy neuron as the main component of the neural network. The architecture of a deep neo-fuzzy neural network is shown, as well as a backpropagation algorithm for this architecture with a triangular membership function for neo-fuzzy neuron. The main advantages of using neo-fuzzy neuron as the main component of the neural network are given. The article describes the properties of a neo-fuzzy neuron that addresses the issues of improving speed and vanishing or exploding gradient. The proposed neo-fuzzy deep neural network architecture is compared with standard deep networks based on the Rosenblatt perceptron.

NEO-FUZZY NEURON, MULTILAYER NEURAL NETWORK, F-TRANSFORM

Вступ

В цей час штучні нейронні мережі (ШНМ) отримали широке розповсюдження для вирішення задач опрацювання інформації різної природи завдяки своїм універсальним апроксимуючим властивостям що досягаються в процесі їх навчання на основі існуючих даних спостережень. Більш високих результатів можливо досягти за допомогою глибоких нейронних мереж (ГНМ) [1-5], котрі хоча й перевершують за якістю рішення задач традиційні нейронні мережі, але їх навчання пов'язано з низкою суттєвих обчислювальних проблем та вимагає більших витрат часу. В основі більшості ШНМ лежить, так званий, елементарний персептрон Розенблатта [6], при цьому як функція активації використовується традиційна сигмоїда або гіперболічний тангенс. Трьохшаровий персептрон забезпечує високу якість апроксимації достатньо складних функцій, заданих в обмеженій області визначення [8]. Спроби покращити якість отриманого рішення шляхом збільшення кількості прихованих шарів ШНМ натикаються на проблему, пов'язану з, так званим, зникаючим та вибуховим градієнтом [9], поява котрого пов'язана з формою

сигмоїдальних активаційних функцій. У зв'язку з цим ГНМ замість сигмоїдальних функцій зазвичай використовують *linear rectified family*, характерним представником котрого є *rectified linear unit* (ReLU), який не страждає від проблем, пов'язаних з обчисленням градієнта. Використання таких функцій не викликає проблем, пов'язаних з обчисленням градієнта, а їх похідні-константи дозволяють спростити дельта-правило навчання кожного окремого нейрона. Разом з тим, використання функції *ReLU* не відповідає вимогам апроксимаційних теорем, які лежать в основі ШНМ та, поперед всього, вимогам теореми Цибенко [7]. Таким чином, ГНМ забезпечують кусково-лінійну апроксимацію, висока якість котрої досягається збільшенням числа прихованих шарів мережі. При цьому збільшення кількості шарів зменшує швидкодію процесу навчання та вимагає більших об'ємів даних для навчання.

1. Архітектура глибокої нео-фаззі нейронної мережі

Нейронна мережа, яку ми розглядаємо, має традиційну багатошарову архітектуру прямого поширення, включаючи в загальному випадку *s* шарів опрацювання інформації.

На вхідний (нульовий) шар подається $x(k) \in R^n$ вектор вхідних сигналів.

$$x(k) = (x_1(k), x_2(k), \ldots, x_n(k)),$$

де k=1, 2, ..., N – номер спостереження в навчальній виборці або індекс поточного дискретного часу. Вихідним сигналом мережі є вектор

$$\hat{y}(k) = (\hat{y}_1(k), \hat{y}_2(k), \ldots, \hat{y}_m(k))^T \in R^m.$$

Далі, для спрощення запису позначень будемо також використовувати вид

$$x(k) \equiv o^{[0]}(k) = (o_1^{[0]}(k), \ldots, o_{i_0}^{[0]}(k) \ldots, o_{n_0}^{[0]}(k))^T,$$

$$\hat{y}(k) \equiv o^{[s]}(k) = (o_1^{[s]}(k), \ldots, o_{i_s}^{[s]}(k), \ldots, o_{n_s}^{[s]}(k))^T.$$

Таким чином, вхідним сигналом p-го шару (p=1, 2..., s) є вектор

$$o^{[p-1]}(k) = (o_1^{[p-1]}(k), \ldots, o_{i_{p-1}}^{[p-1]}(k), \ldots, o_{n_{p-1}}^{[p-1]}(k))^T \in R^{n_{p-1}},$$

а вихідним - вектор

$$o^{[p]}(k) = (o_1^{[p]}(k), \ldots, o_{i_p}^{[p]}(k), \ldots, o_{n_p}^{[p]}(k))^T \in R^{n_p}.$$

При цьому нео-фаззі нейронна мережа містить $\sum_{p=1}^{s} n_p$ нейронів.

Вузлом цієї архітектури є нео-фаззі нейрон (НФН) [10-12] з n_{p-1} входами та одним виходом $o_{i_p}^{[p]}$. На рис. 1 зображено архітектуру i_p-го $NFN_{i_p}^{[p]}$ p-го шару мережі.

Кожен i_p-ий ($i_p = 1,2,\ldots, n_p$) нео-фаззі нейрон p-го ($p = 1,2,\ldots, s$) шару нео-фаззі нейронної мережі містить n_{p-1} нелінійних синапсів $NS_{i_p i_{p-1}}^{[p]}$, кожен з котрих включає h функцій належності $\mu_{i_p i_{p-1} l}^{[p]}$ ($l = 1,2,\ldots, h$) і таку ж кількість синаптичних вагових коефіцієнтів $w_{i_p i_{p-1} l}^{[p]}$, які налаштовуються в процесі навчання. Таким чином, ця архітектура має $\sum_{p=1}^{s} n_p n_{p-1}$ нелінійних синапсів та $h\sum_{p=1}^{s} n_p n_{p-1}$ функцій належності $\mu_{i_p i_{p-1} l}^{[p]}(o_{i_{p-1}}^{[p-1]})$ та таку ж кількість налаштованих синаптичних вагових коефіцієнтів $w_{i_p i_{p-1} l}^{[p]}$.

Вихідний сигнал кожного нелінійного синапсу $NS_{i_p i_{p-1}}^{[p]}$ може бути записаний у вигляді

$$f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \sum_{l=1}^{h} w_{i_p i_{p-1} l}^{[p]} \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \quad (1)$$

а вихідний сигнал нео-фаззі нейрону $NFN_{i_p}^{[p]}$ в цілому

$$o_{i_p}^{[p]} = \sum_{i_{p-1}=1}^{n_{p-1}} f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \sum_{i_{p-1}=1}^{n_{p-1}} \sum_{l=1}^{h} w_{i_p i_{p-1} l}^{[p]} \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right). \quad (2)$$

Для вихідного шару мережі сигнал (2) можна також записати у формі

$$\hat{y}_j = o_{i_s}^{[s]} = \sum_{i_{s-1}=1}^{n_{s-1}} f_{i_s i_{s-1}}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right) = \sum_{i_{s-1}=1}^{n_{s-1}} \sum_{l=1}^{h} w_{i_s i_{s-1} l}^{[s]} \mu_{i_s i_{s-1} l}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right). \quad (3)$$

Далі, вводячи в розгляд синаптичні ваги та функції належності

$$w_{i_p i_{p-1}}^{[p]} = (w_{i_p i_{p-1} 1}^{[p]}, \ldots, w_{i_p i_{p-1} l}^{[p]}, \ldots w_{i_p i_{p-1} h}^{[p]})^T,$$

$$\mu_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) =$$

$$= (\mu_{i_p i_{p-1} 1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p i_{p-1} h}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right))^T$$

розмірності ($h \times 1$), а також

$$w_{i_p}^{[p]} = (w_{i_p 1}^{[p]}, \ldots, w_{i_p i_{p-1}}^{[p]}, \ldots w_{i_p n_{p-1}}^{[p]})^T$$

та

$$\mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) =$$

$$= (\mu_{i_p 1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p n_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right))^T$$

розмірності ($n_p h \times 1$), можна записати

$$f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = w_{i_p}^{[p]T} \mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) \text{ замість (1),}$$

$$o_{i_p}^{[p]} = w_{i_p}^{[p]T} \mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) \text{ замість (2)}$$

$$\text{і } \hat{y}_j = o_{i_s}^{[s]} = w_{i_s}^{[s]T} \mu_{i_s}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right) \text{ замість (3).}$$

В якості функції належності нелінійних сигналів $NS_{i_p i_{p-1}}^{[p]}$ автори нео-фаззі нейрона [10-12] використовували традиційну трикутну функцію, яка задовольняє вимогам одиничного розбиття Руспіні:

$$\mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) =$$

$$= \begin{cases} \dfrac{o_{i_{p-1}}^{[p-1]} - c_{i_p i_{p-1} l-1}^{[p]}}{c_{i_p i_{p-1} l}^{[p]} - c_{i_p i_{p-1} l-1}^{[p]}}, \text{ if } o_{i_{p-1}}^{[p-1]} \in [c_{i_p i_{p-1} l-1}^{[p]}, c_{i_p i_{p-1} l}^{[p]}) \\ \dfrac{c_{i_p i_{p-1} l+1}^{[p]} - o_{i_{p-1}}^{[p-1]}}{c_{i_p i_{p-1} l+1}^{[p]} - c_{i_p i_{p-1} l}^{[p]}}, \text{ if } o_{i_{p-1}}^{[p-1]} \in [c_{i_p i_{p-1} l}^{[p]}, c_{i_p i_{p-1} l+1}^{[p]}) \\ 0, \text{ інакше,} \end{cases} \quad (4)$$

де $c_{i_p i_{p-1} l}^{[p]}, l = 1,2,\ldots,h$ — центри трикутних функцій належності. У випадку, коли всі нелінійні синапси

мережі $NS_{i_p i_{p-1}}^{[p]}$ мають однакове число h центрів, котрі розподілені рівномірно по осям вхідних сигналів, вираз (4) може бути переписаний у більш зручному вигляді:

$$\mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)= \\ =\begin{cases}\Delta_c^{-1}(o_{i_{p-1}}^{[p-1]}-c_{i_p i_{p-1} l-1}^{[p]}), \text{ if } o_{i_{p-1}}^{[p-1]}\in[c_{i_p i_{p-1} l-1}^{[p]}, c_{i_p i_{p-1} l}^{[p]}) \\ \Delta_c^{-1}(c_{i_p i_{p-1} l+1}^{[p]}-o_{i_{p-1}}^{[p-1]}), \text{ if } o_{i_{p-1}}^{[p-1]}\in[c_{i_p i_{p-1} l}^{[p]}, c_{i_p i_{p-1} l+1}^{[p]}) \\ 0, \text{ інакше,}\end{cases} \quad (5)$$

при цьому:

$$\mu_{i_p i_{p-1} l-1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)+\mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)= \\ \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)+\mu_{i_p i_{p-1} l+1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)=1. \quad (6)$$

Умови (4)-(6) означають, що в кожен момент часу k в кожному нелінійному синапсі тільки дві сусідні функції належності можуть спрацювати. Це приводить до того, що в цей же момент налаштовуються тільки два сусідні синаптичні вагові коефіцієнти, тобто на кожному кроці навчання уточнюється тільки $2\sum_{p=1}^{s} n_{p-1} n_p$ синаптичних ваг замість

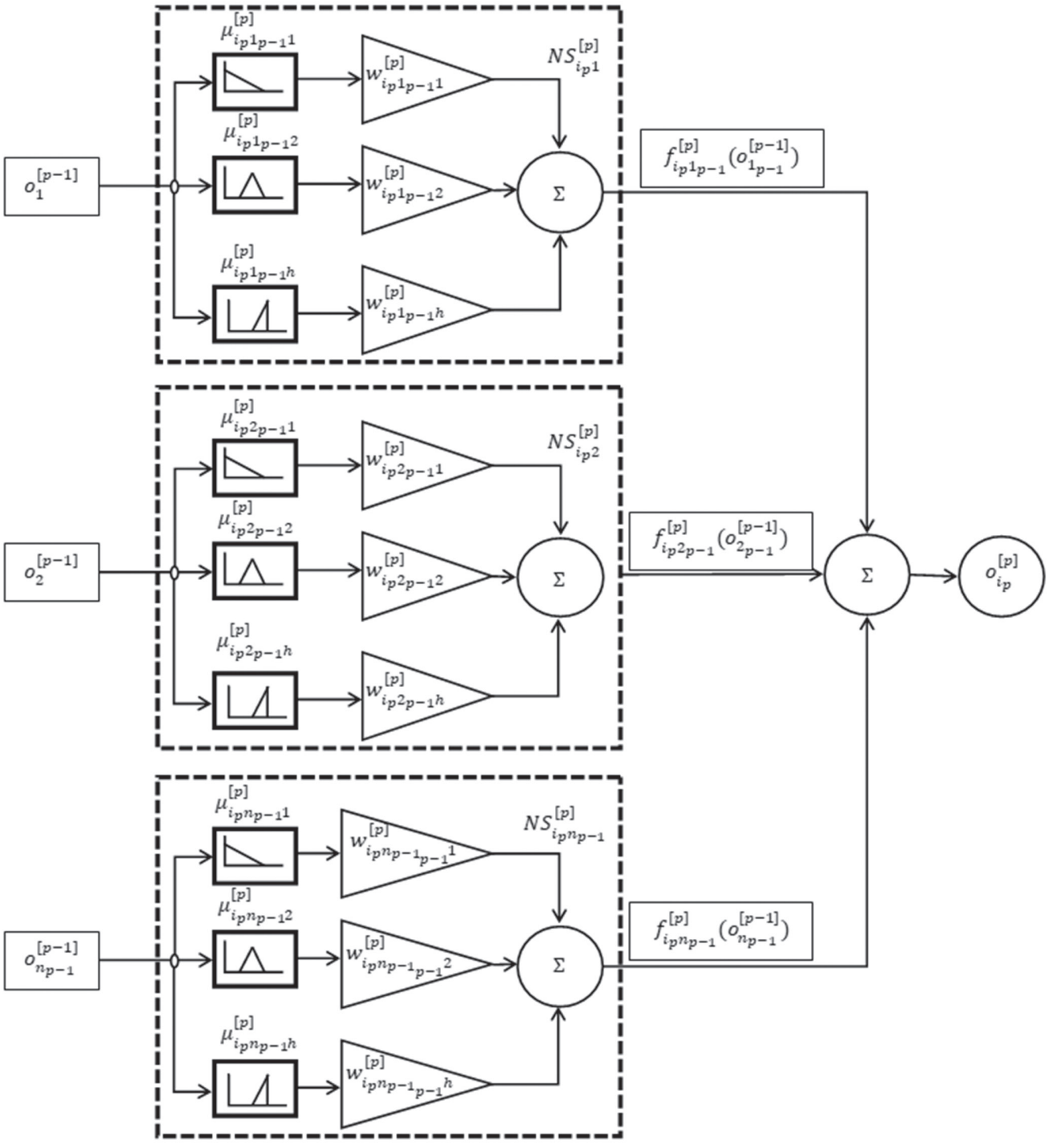

Рис. 1. *i*-ий нео-фаззі нейрон *p*-го шару нео-фаззі нейронної мережі

$h\sum_{p=1}^{s} n_{p-1} n_p$. Оскільки кожен нелінійний синапс мережі реалізує по суті F-перетворення [17], це дозволяє як завгодно точно апроксимувати будь-яку обмежену одновимірну функцію при достатньо великому h. Збільшення числа функцій належності не ускладнює навчання мережі.

Таким чином, кожен NFN є гібридом нео-фаззі системи Ванга-Менделя та F-перетворення, що забезпечує високі апроксимаційні властивості мережі в цілому.

2. Навчання глибокої нео-фаззі нейронної мережі

Навчання багатошарової нео-фаззі мережі реалізується на основі зворотнього поширення похибки.

Процес навчання нео-фаззі мережі зводиться до пошуку синаптичних вагових коефіцієнтів

$$w^{[p]}_{i_p i_{p-1} l}, p = 1,2,\ldots,s; l = 1,2,\ldots,h;$$
$$i_p = 1,2,\ldots,n_p; i_{p-1} = 1,2,\ldots,n_{p-1}$$

шляхом мінімізації прийнятої цільової функції навчання.

В якості критерію навчання використаємо стандартну квадратичну функцію

$$E(k) = \frac{1}{2}\sum_{i_s=1}^{n_s} e^2_{i_s}(k) = \frac{1}{2}\sum_{j=1}^{m} e^2_j(k), \tag{7}$$

де $e^2_j(k) = (y_j(k) - w^{[s]}_{i_s}(k-1)\mu^{[s]}_{i_s}(o^{[s-1]}_{i_{s-1}}(k)))^2$,

$y_j(k)$ – зовнішній навчальний сигнал.

Градієнтна мінімізація (7) для кожного синаптичного вагового коефіцієнту вихідного шару $w^{[s]}_{i_s i_{s-1} l}$ може бути описана рекурентною процедурою

$$\begin{aligned} w^{[s]}_{i_s i_{s-1} l}(k) &= w^{[s]}_{i_s i_{s-1} l}(k-1) - \eta(k)\frac{\partial e^2_{i_s}(k)}{\partial w^{[s]}_{i_s i_{s-1} l}} = \\ &= w^{[s]}_{i_s i_{s-1} l}(k-1) - \eta(k)\mu^{[s]}_{i_s}(O^{[s-1]}_{i_{s-1}}(k)) \end{aligned} \tag{8}$$

(тут $\eta(k)$ – параметр кроку навчання), або у векторній формі

$$\begin{aligned} w^{[s]}_{i_s}(k) = w^{[s]}_{i_s}(k-1) - \eta(k)(y_j(k) - \\ - w^{[s]}_{i_s}(k-1)\mu^{[s]}_{i_s}\left(o^{[s-1]}_{i_{s-1}}(k)\right). \end{aligned} \tag{9}$$

Для вибору $\eta(k)$ може бути використана процедура, яка є гібридом оптимального алгоритму Kaczmarz-Widrow-Hoff та стохастичної апроксимації [18, 19].

$$\begin{cases} w^{[s]}_{i_s}(k) = w^{[s]}_{i_s}(k-1) - r^{-1}_{i_s}(k) e^2_{i_s}(k)\mu^{[s]}_{i_s}\left(o^{[s-1]}_{i_{s-1}}(k)\right), \\ r^{-1}_{i_s}(k) = \alpha r_{i_s}(k-1) + \mu^{[s]}_{i_s}(o^{[s-1]}_{i_{s-1}}(k))^2, \end{cases}$$

де $0 \le \alpha \le 1$ – фактор забування.

Налаштування синаптичних вагових коефіцієнтів прихованих шарів реалізується за допомогою алгоритму зворотного поширення похибки, для чого може бути використана процедура типу (8) у вигляді

$$\begin{aligned} w^{[s-1]}_{i_{s-1} i_{s-2} l}(k) &= w^{[s-1]}_{i_{s-1} i_{s-2} l}(k-1) - \eta(k)\frac{\partial E(k)}{\partial w^{[s-1]}_{i_{s-1} i_{s-2} l}} = \\ &= w^{[s-1]}_{i_{s-1} i_{s-2} l}(k-1) - \eta(k)\frac{\partial e^2_{i_s}(k)}{\partial w^{[s]}_{i_s i_{s-1} l}}\frac{\partial o^{[s]}_{i_s}(k)}{\partial o^{[s-1]}_{i_{s-1}}(k)}\frac{\partial o^{[s-1]}_{i_{s-1}}(k)}{\partial w^{[s]}_{i_{s-1} i_{s-2} l}}. \end{aligned} \tag{10}$$

Оскільки

$$\frac{\partial o^{[s]}_{i_s}}{\partial o^{[s-1]}_{i_{s-1}}} = \frac{\partial f^{[s]}_{i_s i_{s-1} l}(o^{[s-1]}_{i_{s-1}})}{\partial o^{[s-1]}_{i_{s-1}}} = \sum_{l=1}^{h} w^{[s]}_{i_s i_{s-1} l}\frac{\partial \mu^{[s]}_{i_s i_{s-1} l}(o^{[s-1]}_{i_{s-1}})}{\partial o^{[s-1]}_{i_{s-1}}},$$

де

$$\frac{\partial \mu^{[s]}_{i_s i_{s-1} l}(o^{[s-1]}_{i_{s-1}})}{\partial o^{[s-1]}_{i_{s-1}}} = \begin{cases} \Delta_c^{-1}, & \text{if } o^{[p-1]}_{i_{p-1}} \in \left[c^{[p]}_{i_p i_{p-1} l-1}, c^{[p]}_{i_p i_{p-1} l}\right), \\ -\Delta_c^{-1}, & \text{if } o^{[p-1]}_{i_{p-1}} \in \left[c^{[p]}_{i_p i_{p-1} l}, c^{[p]}_{i_p i_{p-1} l+1}\right), \\ 0, & \text{інакше}, \end{cases} \tag{11}$$

алгоритм (10) остаточно може бути записаний у формі

$$\begin{aligned} w^{[s-1]}_{i_{s-1} i_{s-2} l}(k) &= w^{[s-1]}_{i_{s-1} i_{s-2} l}(k-1) - \\ &- \eta(k)\sum_{i_s=1}^{n_s}\partial e^2_{i_s}(k)\sum_{l=1}^{h} w^{[s]}_{i_s i_{s-1} l}(k) * \\ &* \frac{\partial \mu^{[s]}_{i_s i_{s-1} l}\left(o^{[s-1]}_{i_{s-1}}\right)}{\partial o^{[s-1]}_{i_{s-1}}}\mu^{[s-1]}_{i_{s-1} i_{s-2} l}(o^{[s-2]}_{i_{s-2}}(k)) \end{aligned} \tag{12}$$

Аналогічно (12) може бути записаний алгоритм налаштування для p-го нейронного шару, при чому похибка нео-фаззі нейрона p-го шару має вигляди:

$$\frac{\partial E}{\partial o^{[p]}_{i_p}} = \sum_{i=1}^{n_{p+1}}\frac{\partial E}{\partial o^{[p+1]}_{i_{p+1}}}\frac{\partial o^{[p+1]}_{i_{p+1}}}{\partial o^{[p]}_{i_p}},$$

де дельта-похибка має вигляд

$$\delta^{[p]}_j = \frac{\partial E}{\partial o^{[p]}_j} = \sum_{i=1}^{n_{p+1}}\delta^{[p+1]}_j\frac{\partial o^{[p+1]}_{i_{p+1}}}{\partial o^{[p]}_{i_p}} \tag{13}$$

Тоді для всіх вагових коефіцієнтів можна записати

$$\frac{\partial E(k)}{\partial w^{[p]}_{i_p i_{p-1} l}} = \sum_{i=1}^{n_{p+1}}\delta^{[p+1]}_j\frac{\partial o^{[p+1]}_i}{\partial o^{[p]}_j}\frac{\partial o^{[p]}_i}{\partial w^{[p]}_{i_p i_{p-1} l}},$$

де дельта похибки береться з наступного шару, а проміжні похідні обчислюються наступним чином:

$$\frac{\partial o^{[p+1]}_{i_{p+1}}}{\partial o^{[p]}_{i_p}} = \frac{\partial f^{[p+1]}_{i_s i_{s-1} l}(o^{[p]}_{i_p})}{\partial o^{[p]}_{i_p}} = \sum_{l=1}^{h} w^{[p+1]}_{i_{p+1} i_p l}\frac{\partial \mu^{[p+1]}_{i_{p+1} i_p l}(o^{[p]}_{i_p})}{\partial o^{[p]}_{i_p}},$$

де права частина отримується за допомогою (11).

Похідна вихідного сигналу нео-фаззі нейрона по ваговому коефіцієнту має вигляд:

$$\frac{\partial o^{[p]}_i}{\partial w^{[p]}_{i_p i_{p-1} l}} = \frac{f^{[p]}_{i_p i_{p-1}}\left(o^{[p]}_i\right)}{\partial w^{[p]}_{i_p i_{p-1} l}} = \mu^{[p]}_{i_p i_{p-1} l}\left(o^{[p-1]}_{i_{p-1}}\right),$$

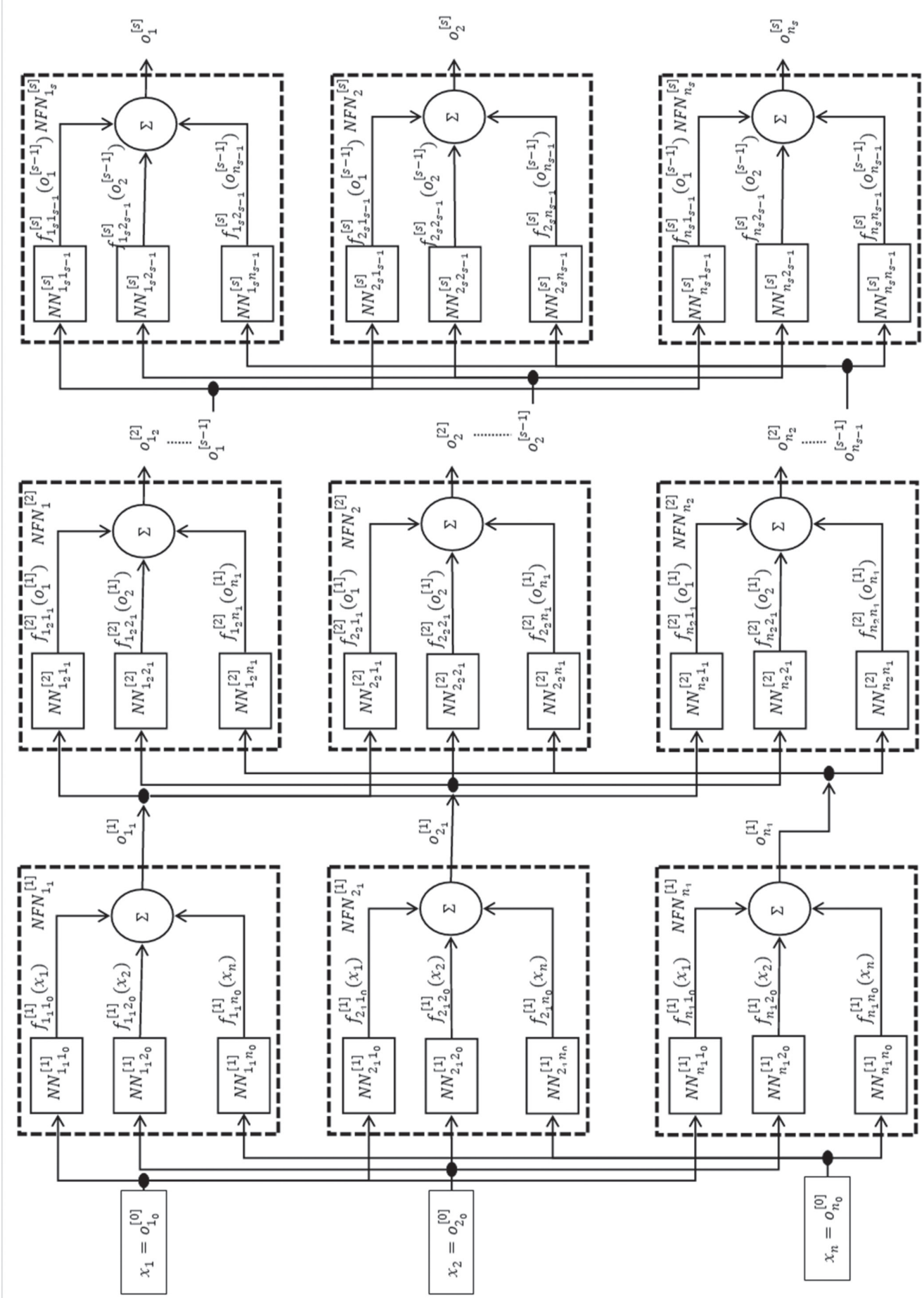

Рис. 2. Архітектура багатошарової нео-фаззі нейронної мережі

де остаточний вираз

$$w^{[p]}_{i_p i_{p-1} l}(k) = w^{[p]}_{i_p i_{p-1} l}(k-1) - \eta(k)\frac{\partial E(k)}{\partial w^{[p]}_{i_p i_{p-1} l}}$$

може бути записано у вигляді:

$$w^{[p]}_{i_p i_{p-1} l}(k) = w^{[p]}_{i_p i_{p-1} l}(k-1) - \\ -\eta(k)\sum_{i=1}^{n_{p+1}}\delta^{[p+1]}_j \frac{\partial o^{[p+1]}_i}{\partial o^{[p]}_j}\mu^{[p]}_{i_p i_{p-1} l}(o^{[p-1]}_{i_{p-1}}) \qquad (14)$$

Нескладно бачити, що навчання нео-фаззі нейронної мережі, архітектура якої в цілому наведена на рис. 2, з обчислювальної точки зору простіше в порівнянні з ШНМ та ГНМ побудованими на основі традиційних персептронів Розенблата, оскільки похідні трикутних функцій належності є константами.

3. Експериментальні дослідження

Порівнювалися між собою результати навчання на якість апроксимації багатошарового персептрону з сигмоїдальними та ReLU-функціями активації та запропонованої нео-фаззі мережі. Кожна модель мала 4 прихованих шари та по 50 нейронів на кожному прихованому шарі. Результати експериментів показали, що нео-фаззі мережа досягає тієї ж якості що й багатошаровий персептрон за меншу кількість епох навчання.

Висновки

Запропоновано архітектуру та алгоритм навчання глибокої нео-фаззі нейронної мережі, основною відмінністю якої від традиційних багатошарових нейронних мереж з прямим поширенням інформації є використання в якості вузлів нео-фаззі нейронів замість традиційних елементарних персептронів Розенблата. Запропонована нео-фаззі нейронна мережа має високі апроксимаційні властивості та характеризується підвищеною швидкістю навчання своїх синаптичних ваг, завдяки використанню трикутних функцій належності в нелінійних синапсах нео-фаззі нейронів. Крім того, запропонована глибока нео-фаззі нейронна мережа є досить простою з точки зору обчислювальної реалізації.

Список літератури:

[1] *Bengio Y, LeCun Y, Hinton G.* Deep Learning – Nature – 2015-521 – p.436-444.

[2] *Schmidhuber J.* Deep learning in neural networks: An overview – Neural Networks – 2015–01 – p.85-117.

[3] *Googfellow I., Bengio Y., Courville A.* Deep Learning – MIT Press, 2016-787p.

[4] *Graupe D.* Deep Learning Neural Networks: Design and Case Studies- New York: World Scientific, 2016 – 260p.

[5] *Caterini A.L., Chang D.E.* Deep Neural Networks in a Mathematical Framework – Springer, 2018 –79p.

[6] *Cichocki A., Unbehauen R.* Neural Networks for Optimization and Signal Processing – Stuttgart: Teubner, 1993-526p.

[7] *Cybenko G.* Approximation by superpositions of a sigmoidal function – Math. Control Signals Systems. – 1985 – 2 – p.303-314.

[8] *Hornik K.* Approximation capabilities of multilayer feedforward networks – 1994 – 4 – p.251-257.

[9] *Aggarwal Ch.C.* Neural Networks and Deep Learning – Cham: Springer, 2018-512p.

[10] *Yamakawa T, Uchino E, Miki T., Kusabagi H.* A neo fuzzy neuron and its applications to system identification and predictions to system behavior. – Proc. 2nd Int. Conf. on Fuzzy Logic and Neural Networks, pp. 477-483, 1992.

[11] *Uchino E, Yamakava T.* Neo-fuzzy neuron based new approach to system modeling with application to actual system - Proceedings Sixth International Conference on Tools with Artificial Intelligence – New Orlean, LA, USA, 1994 – p.564-570.

[12] *Miki T., Yamakawa T.,* "Analog implementation of neo-fuzzy neuron and its on-board learning," In Computational Intelligence and Applications, Piraeus: WSES Press, 1999, pp. 144-149.

[13] *Kolodyazhniy V, Bodyanskiy Ye.* Fuzzy Kolmogorov's network – Lecture Notes in Computer Science. – 3214 – Heidelberg: Springer Verlag, 2004. – p.764-771.

[14] *Bodyanskiy Ye., Kolodyazhniy V., Otto P.* Neuro-fuzzy Kolmogorov's network for time series prediction and pattern classification – Lecture Notes in Artificial Intelligence – 3698 – Heidelberg: Springer Verlog, 2005. – p.191-202.

[15] *Bodyanskiy Ye., Popov S., Rybalchenko T.* Multilayer neuro-fuzzy network for short term electric load forecasting – Lecture Notes in Computer Science. – 5010 – Berlin-Heidelberg: Springer Verlag, 2008. – p.339-348.

[16] *Bodyanskiy Ye., Vynokurova O., Setlak G., Peleshko D., Mulesa P.* Adaptive multivariate hybrid neuro-fuzzy system and its on-board fast learning – Neurocomputing – 2017 – 230-p.409-416.

[17] *Perfilieva T.* Fuzzy transforms: Theory and applications – Fuzzy Sets and Systems – 2006 – 157 – p.993-1023.

[18] *Bodyanskiy Ye., Kolodyazhniy V., Stephan A.* An adaptive learning algorithm for a neuro-fuzzy network – Ed. by B.Reush "Computitional Intelligence. Theory and Application" – Berlin-Heidelberg: Ney York: Springer, 2001. – p.68-75.

[19] *Otto P., Bodyanskiy Ye., Kolodyazhniy V.* A new learning algorithm for a forecasting neuro-fuzzy network - Integrated Computer Aided Engineering – 2003 – 10(4) – p.399-409.

Надійшла до редколегії 10.04.2019

UDK 519.62

D.S. Nazarenko[1], I.V. Afanasieva[2], N.V. Golian[3]

[1]Student, Software engineering,NURE, Kharkiv, Ukraine, dmytro.nazarenko@nure.ua

[2]PhD, assistant professor, Software engineeringNURE, Kharkiv, Ukraine, iryna.afanasieva@nure.ua

[3]PhD, assistant professor, Software engineeringNURE, Kharkiv, Ukraine, nataliia.golian@nure.ua

NEURAL NETWORK APPROACH FOR EMOTIONAL RECOGNITION IN TEXT

The article is devoted to one of the most popular trends in the field of IT today – natural language processing, in particular, the extraction of emotions from the text using the neural network approach. The main task was to solve the problem of the high costs of time and human resources for companies to receive feedback from users and process emotional reactions of the second one. That to decide the task it was necessary to make modelling and learn neural network using own architecture based on the backpropagation algorithm that to recognize the emotional component in the text.The emotional component of reviews was used as a metric for evaluating user reactions. It was decided to work with five types of emotions that will help to provide better results. The neural network architecture consists of interconnected layers: embedding, bidirectional LSTM, pooling, dropout layers and two dense layers.

For the neural network learning was selected an open dataset consisted of 47,288-tagged posts from Twitter. As a result, the F-measure on the test dataset was 0.62 and which is a worthy indicator in comparison with large business solutionsю.

CLASSIFICATION, DATA, EMOTIONS, NEURAL NETWORKS, RESPONSES, TEXT, TONALITY

Introduction

In the modern world, automation has affected all spheres of human life. It is impossible to imagine life without computer technology: everything around us is connected with the IT industry. Humanity is evolving and the IT industry is developing too.

People generates huge number of data every day. It is can not be unnoticed that finally led to the emergence of a science about big data [1].

It is important to understand that the data cannot be benefit by themselves. In addition, it is important to use them in real problems. For example, all posts and photos in social networks, queries in search engine, and the benefits of choosing videos and music, and this is not an empty data set, but an important digital resource. It was the big data pushed to a revolution in the field of computer sciences – the development of artificial intelligence and machine learning [2].

A new round of software evolution is just beginning, but today you can feel its impact, which lead to the development of products built based on relevant data usage approaches will soon.

1. Analysis domain

Information technology is changing the nature of modern business. Companies are trying to automate internal processes to reduce the cost of maintaining a large staff and speed up work. Today, the number of spheres that non-automated are decreasing.

Huge amount of text data forces business representatives to hire staff for data processing. Companies that are not capable of holding a large staff who will make market analysis cannot cope with the problem of text data analysis.

The problem is actual for both small and large businesses, because nobody is interested in additional costs.

However, not all business processes are able to automation, especially in areas of data processing and analysis where human skills are needed. These areas include face recognitions in photos and real-time video [3], text processing [4].

For such tasks, decisions should use actual approaches in the world of information technologies – neural networks.

Neural networks are built on the principle of the human brain that allows them to be trained on the provided data [5].

So feedback from users to business is needed.

2. Problem Statement

The main task is to solve the problem of the high costs of time and human resources for companies to receive feedback from users and process emotional reactions of the second one using the neural network approach.

It is necessary to make modelling and learn neural network using own architecture based on the backpropagation algorithm [6] that to recognize the emotional component in the text.

That to prepare training data it is necessary to make data collection. First, to choose the text data that will be used for learning neural network, as well as its validation.

Also it is important to improve the quality of the learning process to make comprehensive data pre-processing. The dataset should be divided into a training one, which will be used during the training process and a test one, it is necessary to make a conclusion about the ability of the neural network generalization [7].

The key criterion of the neural network operation is the accuracy of emotion classification [8]. To do this, it is important to choose a metric that will help to evaluate the classification accuracy. The F1 score metric will allow to get a more realistic measure of the accuracy classifier.

Potential customers of this application will be companies of medium and large sizes, which are engaged in sales and provision of services. An important advantage for companies will be to reduce staffing in order to solve these types of problems and in particular to analyze the reaction of consumers of a product or company's services on the Internet.

3. Decision making

It is necessary to use various methods for text analysis that to receive user reviews on the certain product. It is required to understand how products satisfy the users and what emotions they feel.

The emotional component of reviews is used as a metric for evaluating user reactions. This metric allows you to determine the overall customer attitude to the product. Usually reviews are divided into 3 categories: positive, negative and neutral [9,10-13].

If to divide all reviews into positive and negative, we will lose information that is useful for companies. In real life, comments and reviews are rarely absolutely positive or completely negative. The emotional colour of a comment has a whole palette of emotions. According to the spectrum of emotions, we can study the reaction of the market in detail and make conclusions [8].

However, some hidden emotions are difficult to detect even for person. Specific and unpopular emotions do not have a big impact on the overall user reaction, because of business is interested in indicators of the consumer market as a whole, and not of the user individually. Therefore, it is necessary to choose the basic emotions that it makes sense to extract but the result of the analysis will have business value.

Selection of emotions from the text should not be limited by binary classification [9]. If we will divide text data only to positive and negative classes, then the results of the analysis will lose flexibility and will not sufficiently characterize what emotions the users had exactly.

Nevertheless, during research the number of emotions should not exceed seven [8], otherwise, there will be an information that does not have an impact on the user reaction.

Thus, several classes should be extracted for classification, since binary classification entails a lack of informativeness in the results. In addition, an information overload should be avoided by optimizing the learning process of neural network on some layers.

As the investigated emotions were chosen: 1 – happiness; 2 – sadness; 3 – anger; 4 – aggression. They have good informational content in tasks of comments and feedback processing from users. Types of emotions are common and well expressed in text messages. It is suitable characteristic for user reactions analysis. Also, we decided add fifth emotion – neutral, because of not brightly expressed emotions (unemotional colouring of the text).

It was decided to work with five types of emotions that will help to provide better results.

4. Neural network architecture

The neural network architecture was created. Our interconnected layers are consist of embedding, bidirectional LSTM, pooling, dropout layers and two dense layers, so let us describe each layer more detail.

In general, embedding is a form of words representation [14]. It helps to connect human and machine in language understanding. This type of layer provides distributed representations of text in n-dimensional space. Word embedding is a family of natural language processing methods aimed at mapping semantic meaning into geometric space. It is implemented by associating a numerical vector with each word in the dictionary, so that the distance between any two vectors encompasses part of the semantic relations between two related words. For example, "coconut" and "polar bear" are words that semantically rather different, but embedding will present them as vectors that will be located at a great distance from each other; but the words "kitchen" and "dinner" are related words, so they will be located closer to each other.

Word embedding is computed by applying dimension reduction methods to co-occurrence datasets between words in the body of the text and implemented using the GloVe approach.

The next layer is bidirectional long short-term memory (bidirectional LSTM). LSTM is a type of recurrent neural network capable of learning long-term dependencies [15]. LSTMs are designed to eliminate long-term dependency problems and the specialty is to memorize of information for long periods of time.

A bidirectional recurrent neural network (BRNN) [16] connects in opposite directions two hidden layers with the same input. The output layer of the recurrent neural network can receive information from past (previous) and future states (next) simultaneously due to the form of generative learning. BRNN has been created to increase the amount of incoming information available to the network, as shown in Figure 1.

For example, multilayer perceptron (MLP) [17] and time delay neural network (TDNN) [16] have limitations on the input data flexibility, because of requirements of the input data recording. The standard recurrent neural network (RNN) [15] also has a limit, since future input information from the current state can not be reached. On the contrary, the BRNN does not

require the recording of input data. Moreover, future input information is available from the current state. It allows future data to be taken into consideration during training and improve the train ability of the model. BRNN is especially useful when the presence of the context of the input data improves the result. For example, accuracy can be enhanced by the words that placed consecutive in the sentence when we are recognizing word by context.

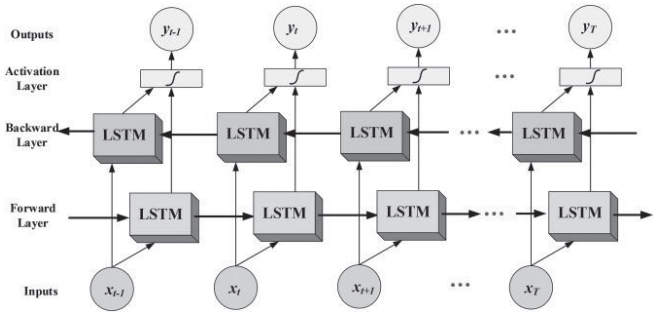

Fig. 1. BRNN scheme

The next layer is pooling [18] is a non-linear compaction of matrices with numbers. Numbers in matrices are words that pass non-linear transformation. The most used function is the max function in GlobalMaxPooling [18]. Therefore, transformations affect non-intersecting rectangles or squares, where each is compressed into a number that has a maximum value. The pooling can significantly reduce the spatial data volume and is interpreted as follows: if at the previous operation some data attributes have already been identified, then for further processing such detailed data are not use, and they are condense to less detailed. In addition, unnecessary parts filtering helps the neural network not to be retrain.

Than the training data are combined and transferred to a dense layer. Moreover, further layers do not have spatial structure, but have a relatively small dimension to return the final result of the classification.

To improve the learning quality of the model, the dropout layer was included in front of the output layer [19]. The dropout is a method for regularization of artificial neural networks, and the goal is to prevent overfitting. The essence of the method is that in the process of learning from the neural network subnet is randomly allocated and training for subnet is provided. The training subnet comes from excluding of neurons from the full original network (dropping out) with probability p, thus the probability that the neuron will stay in the network is $q = 1–p$. In the process of learning the excluded neurons do not contribute to any stages of chosen backpropagation algorithm, therefore excluding at least one of the neurons is equivalent to learning new neural network.

It was decided to make the output layer as a fully layer with the five neurons equal to the number of classified emotions. The Softmax function was chosen as the neuron activation function [20]. The Softmax function is a generalization of the logistic function for the multidimensional case. The function converts a vector z of dimension K into a vector σ of the same dimension, where each coordinate σ_i of the obtained vector is represented by a real number in the interval [0,1] and the sum of the coordinates equal 1. Softmax is given by the following formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}},$$

where z_i is the value at the output of the i-th neuron before activation, but N is the total number of neurons in the layer. The Softmax function is used in machine learning for classification problems when the number of possible classes are more than two. The coordinates σ_i of the resulting vector are interpreted as the probabilities that the text belongs to class i.

As a loss function for this model, it was decided to choose categorical cross entropy [21]. Categorical cross entropy is used to categorize one label and it means that only one category for each data point is applicable. In other words, a data sample can belong to just one class. Categorical cross entropy compares the distribution of predictions, that is, activations in the output layer, one for each class, with the true distribution, where the probability of the true class is set to 1 and 0 for other classes. In other words, the true class as a one-hot encoded vector is represented, and the closer the model outputs to this vector, the less the loss. Categorical cross entropy is used in conjunction with the Softmax activation function.

5. The neural network training

For the neural network learning, an open dataset from the github repository was selected [22]. The dataset consist of 47,288-tagged posts from Twitter in English. The data set was chosen due to the large number of textual data samples with already labelled classes of emotions. The mark-up was made in five classes: joy, sadness, anger, aggression and a neutral class of emotions to designate text samples where the emotional component is not brightly expressed. The number of examples of each class are neutral – 9643, happy – 16297, sad – 15938, hate – 4301 and anger – 1109.

The dataset was pre-processed where we used normalization and lemmatization [23]. During normalization- the texts were cleaned of punctuation marks, all letters were switched to lowercase. Then lemmatization was performed – the words were reduced to their normal form, and the stop words were deleted.

The next step of data preparation for training was to bring data samples to a form that would be convenient

for using them as input parameters of a neural network. So using the tools of the Keras library [24] text was vectorized. Each word in the text was associated with a numeric index in the dictionary. As a result, each data sample represented by a vector of numbers. Each vector was supplemented with zeros to a constant length that the length of the text does not affect the final ability of the neural network to generalize.

The vectorized data representation then was split into two subsets – train and test datasets, in a ratio of 3 to 1. The train dataset was used at the entire training stage, and the test dataset was used to evaluate the quality of the model prediction on data that in the training were not involved.

So implementing the model, the Keras library was used [24] and own neural network architecture was created and used.

The vector of tokens of constant length was fed to the input of the model. The sequence of tokens was transferred to the Embedding layer. As a layer of embedding, it was decided to use the pre-trained word vectors GloVe [25]. GloVe is an unsupervised learning algorithm for generating vector representations for words. The training is performed on aggregated global word-match statistics from the corpus, and the resulting representations demonstrate linear substructures of the vector word space. The GloVe model is trained on non-zero elements of the global word match matrix and shows how often the words occur with each other in a given corpus. This matrix requires one pass through the entire corpus to collect statistics. Subsequent training iterations are much faster, because of number of non-zero elements of the matrix is usually much smaller than the total number of words in the body. However, in this model have already used filled matrix with ready-made word vectors, since the pre-trained GloVe model was used.

The output data of the embedding layer is a matrix of 30 by 200, where each word is associated with its vector representation and, the data passes through the layer of a bidirectional recurrent neural network. The input sequence is served in the usual order of time for one network and in the reverse order of time for another. The outputs of the two networks at each time step are combined. This structure allows the network to have both reverse and direct information about the sequence at each time step. At the result the output, we have less matrix of 30 to 64.

The GlobalMaxPooling1D approach [18] for time data takes the maximum vector for measuring steps.

Thus, a multidimensional table with a form [10, 4,10] becomes a global multidimensional table with a form [10, 10] after merging.

Suppose we have a simple sentence with 3 words, and some vector representation of these words. In the case of GlobalMaxPooling1D, the maximum vector of this sentence is taken.

The pooling layer helps get rid of data redundancy, which allows the neural network to take up less memory and learn faster. Than output data passes through two fully connected Dense layers, between which we disable 5% of random neurons and it helps the model to avoid overfitting. At the same time, the last Dense layer in the neural network contains 5 neurons that equal to the number of classes of emotions. Each neuron has a Softmax activation function. Figure 2 shows the final architecture of the designed neural network.

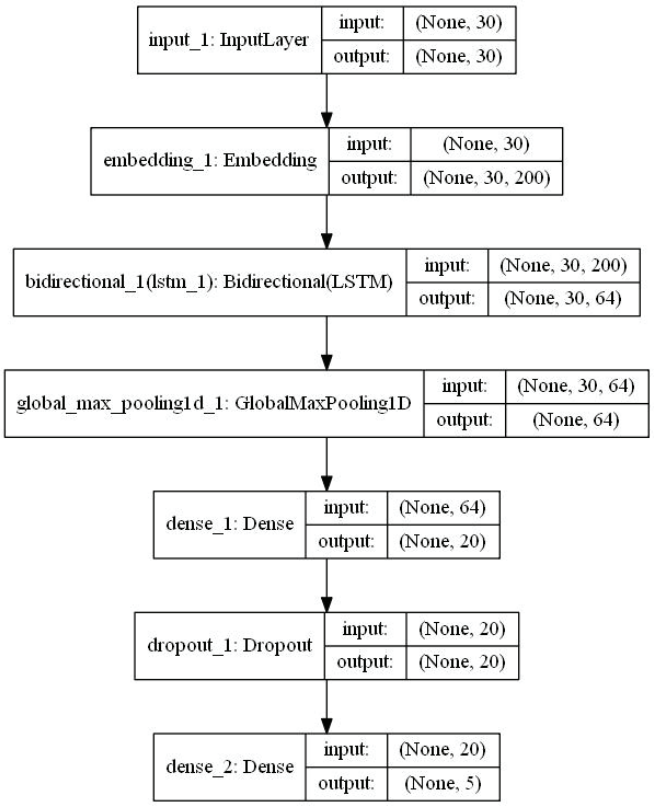

Fig. 2. Neural network architecture

As a result, the neural network is trained using the backpropagation algorithm [6]. Training by this method involves two passes through all layers of the network: direct and reverse. The input vector is fed to the input layer of the neural network with a direct pass. Than input vector propagates through the network layer by layer. As a result, a set of output signals are generated and gives the actual response of the neural network to this input image. During the direct pass, all synaptic weights of the network are fixed. During the back pass, all synaptic weights are adjusted in accordance with the error correction rule: the actual output of the neural network is subtracted from the desired and as a result is an error signal. This signal subsequently propagates through the network in the direction opposite to the direction of synaptic connections, hence the name – algorithm of

backpropagation. Synaptic weights are adjusted to maximize the output signal of the network to the desired.

The backpropagation algorithm is as follows:

1. Initialize synaptic weights with small random values.

2. Choose the next training pair from the training set; submit the input vector to the network input.

3. Calculate network output.

4. Calculate the difference between the network output and the required output (the target vector of the training pair).

5. Adjust network weights to minimize errors.

6. Repeat steps 2-5 for each vector of the training set until the error on the entire set reaches an acceptable level.

In the process of training to assess the accuracy of the neural network the accuracy metric was used. It is interprets as the proportion of correct answers.

6. Results

As a result, at the testing stage, the resulting model was evaluated on a test dataset. It is consisted of 9457 text records from the Twitter microblogging service. As a metric, the F-measure was used – average harmonic of precision and recall. Precision is the proportion of correctly predicted instances among all found, and recall – the proportion of correctly predicted instances relative to the total number of relevants. The F-measure on the test dataset was 0.62 and which is a worthy indicator in comparison with large business solutions and other solutions of the problem.

References

[1] *Майер-Шенбергер В.et al.* Big Data. A Revolution That Will Transform How We Live, Work, and Think / пер. сангл. Инны Гайдюк.–М.: Манн, Иванов, Фербер. –2014. – 240 с.

[2] *Флах П.* Машинное обучение. – М.: ДМК Пресс.– 2015.– 400 с.

[3] *Друки А.* Система поиска, выделения и распознавания лиц на изображениях// Известия Томского политехнического университета [Известия ТПУ]. – 2011. – Т. 318, № 5. – С. 64 -70.

[4] *Zhang L., Wang S., Liu B.* Deep Learning for Sentiment Analysis: A Survey// Wiley Online Library.– 2018. –34 p.

[5] *Lawrence, Jeanette.*Introduction to neural networks: design, theory and applications // Nevada City: California Scientific Software.–1994. – 324 p.

[6] *Goodfellow I., Bengio Y., Courville A.* Deep Learning. // MIT Press. –2016. – 196 p.

[7] Bousquet O., Elisseef A. Stability and Generalization // Journal of Machine Learning Research.–1992. – P. 499-526.

[8] *Назаренко Д. Афанасьєва І.* Аналіз сервісів щодо розпізнавання емоційної складової відгуків користувачів. Актуальные вызовы современной науки// Сб. научных трудов – Переяслав-Хмельницкий.– 2019. – Вип.4(36), ч.1 – С. 85-90.

[9] *Read J. et al.* Classifier Chains for Multi-label Classification// Machine Learning Journal. Springer. –2011.–Vol. 85(3). –27 p.

[10] *ChoiY., Lee H.* Data properties and the performance of sentiment classification for electronic commerce applications//Information Systems Frontiers. Springer. – 2017. – Vol. 19, Issue 5. – P. 993–1012

[11] *Dashtipour K. et al.* Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques// Cognitive Computation. Springer. – 2016. –Volume 8, Issue 4. –P. 757–771

[12] *Shuhaida Mohamed Shuhidan et al.*Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm//International Conference on Kansei Engineering & Emotion Research. KEER. – 2018.–P. 64-72.

[13] *Souma W.et al.* Enhanced news sentiment analysis using deep learning methods//Journal of Computational Social Science. –2019. – Volume 2, Issue 1. P. 33–46.

[14] *Goldberg Y., Levy O.et al.*Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. – 2014. – 5 p.

[15] *Hochreiter S., Schmidhuber J.* Long Short-Term Memory// Neural Computation.–1997. –№9 (8). – P. 1735–1780.

[16] *Graves A. et al.* Bidirectional LSTM networks for improved phoneme classification and recognition. Artificial Neural Networks: Formal Models and Their Applications// ICANN. Springer Berlin Heidelberg. –2005. –P.799-804.

[17] *Rosenblatt F.* Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms// Spartan Books, Washington DC. – 1961.–

[18] *Graham B.* Fractional Max-Pooling.– 2015. –10p.

[19] *Hinton G. et al.* Improving neural networks by preventing co-adaptation of feature detectors– 2012. –18p.

[20] *Gao Bolin, Pavel Lacra.* On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. –2017–10p.

[21] *Murphy K.* Machine Learning: A Probabilistic Perspective// MIT Press. – 2012. –P. 57-571

[22] *Liu T.*Text emotion classification dataset. –2018.

[23] *Airio Eija.* Word Normalization and Decompounding in Mono- and Bilingual IR//Information Retrieval. – 2006.– Vol. 9, Issue 3. P. 249–271.

[24] *Ketkar N.* Introduction to Keras. In: Deep Learning with Python//Apress, Berkeley, CA – 2017. – P. 97-111

[25] *Pennington J. et al.* GloVe: Global Vectors for Word Representation// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) – 2014. – P.1532-1543.

The article was delivered to your editory stuff on the 23.05.2019

УДК 004.891.3

В.О. Лещинський[1], **І.О. Лещинська**[2]

[1]доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Україна, volodymyr.leshchynskyi@nure.ua

[2]доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Україна, iryna.leshchynska@nure.ua

МОДЕЛЮВАННЯ ВИБОРУ КОРИСТУВАЧА В УМОВАХ ОБМЕЖЕНЬ ХОЛОДНОГО СТАРТУ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ

Розглянуто проблему підтримки вибору користувача в рекомендаційних системах з урахуванням обмежень, що виникають в умовах холодного старту. Виконано структуризацію даної проблеми та виділено такі аспекти холодного старту, як поява нового користувача, поява нового об'єкту інтересу споживача, зміна контексту вибору об'єктів користувачем, зміна інтересів споживачів з часом. Запропоновано орієнтовану на систему обмежень модель вибору об'єктів у нормальному режимі роботи рекомендаційної системи, а також орієнтовану на обмеження модель вибору об'єктів в умовах холодного старту. Обмеження у запропонованих моделях представлені у вигляді предикатів на змінних, що характеризують властивості споживачів та об'єктів їх інтересу, а також контекст вибору споживача. Перевага запропонованих моделей полягає у можливості обмежити вхідні дані, таким чином, щоб вони відповідали найбільш суттєвим закономірностям вибору споживачів у даному контексті на даному інтервалі часу, що дає можливість спростити побудову рекомендацій для нових споживачів і нових об'єктів. Запропоновано підхід до побудови рекомендацій в умовах обмежень холодного старту. Підхід передбачає формування обмежень на основі інтелектуального аналізу вхідних даних рекомендаційної системи, а також подальше використання цих обмежень при побудові рекомендацій в умовах холодного старту.

РЕКОМЕНДАЦІЙНА СИСТЕМА, ОБМЕЖЕННЯ, ПЕРСОНАЛІЗАЦІЯ РЕКОМЕНДАЦІЙ, ХОЛОДНИЙ СТАРТ, КОНТЕКСТ ВИБОРУ КОРИСТУВАЧА

Лещинский В.А., Лещинская И.А. Моделирование выбора пользователя в условиях ограничений холодного старта рекомендательной системы. Рассмотрена проблема поддержки выбора пользователя в рекомендательных системах с учетом ограничений, возникающих в условиях холодного старта. Выполнено структурирование данной проблемы и выделены такие аспекты холодного старта, как появление нового пользователя, появление нового объекта интереса потребителя, изменение контекста выбора объектов пользователем, изменение интересов потребителей со временем. Предложена ориентированная на систему ограничений модель выбора объектов в нормальном режиме работы рекомендательной системы, а также ориентированная на ограничения модель выбора объектов в условиях холодного старта. Ограничения в предложенных моделях представлены в виде предикатов на переменных, характеризующих свойства потребителей и объектов их интереса, а также контекст выбора потребителя. Преимущество предложенных моделей заключается в возможности ограничить входные данные, таким образом, чтобы они соответствовали наиболее существенным закономерностям выбора потребителей в данном контексте на данном интервале времени, что позволяет упростить построение рекомендаций для новых потребителей и новых объектов. Предложен подход к построению рекомендаций в условиях ограничений холодного старта. Подход предполагает формирование ограничений на основе интеллектуального анализа входных данных рекомендательной системы, а также дальнейшее использование этих ограничений при построении рекомендаций в условиях холодного старта.

РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА, ОГРАНИЧЕНИЯ, ПЕРСОНАЛИЗАЦИЯ РЕКОМЕНДАЦИЙ, ХОЛОДНЫЙ СТАРТ, КОНТЕКСТ ВЫБОРА ПОЛЬЗОВАТЕЛЯ

Leshchynskyi V., Leshchynska I. Modeling the user's choice in the constraints of the cold start of the recommender system. The problem of supporting user choice in recommender systems is considered, taking into account the limitations that arise when solving a cold start problem. Structuring of this problem was carried out and such aspects of a cold start were highlighted as the emergence of a new user, the emergence of a new consumer interest object, a change in the user selection context, a change in consumer interests over time. A system-oriented model of object selection in the normal operation mode of the recommender system was proposed, as well as a model-oriented model of object selection under cold start conditions. Restrictions in the proposed models are presented in the form of predicates on variables that characterize the properties of consumers and objects of their interest, as well as the context of consumer choice. The advantage of the proposed models is the ability to limit the input data, so that they correspond to the most significant laws of consumer choice in this context at a given time interval, which allows us to simplify the construction of recommendations for new consumers and new objects. An approach to building recommendations in the context of cold start restrictions is proposed. The approach assumes the formation of constraints based on the intellectual analysis of the input data of the recommender system, as well as the further use of these constraints in constructing recommendations in cold start conditions.

RECOMMENDATION SYSTEM, CONSTRAINTS, PERSONALIZATION OF RECOMMENDATIONS, COLD START, CONTEXT OF USER CHOICE

Вступ

Рекомендаційні системи призначені для підтримки вибору споживача в умовах відсутності детальних знань споживача про властивості множини аналогічних товарів та послуг. Такі системи знайшли широке застосування в сфері електронної комерції, наприклад, у сайтах з продажу товарів у мережі Інтернет, сайтах бронювання та продажу квитків, бронювання готелів, у сфері розваг, для продажу фільмів та музики [1].

Функціонування таких систем базується на прогнозуванні потреб конкретного споживача. За результатами прогнозування формується персоналізований перелік товарів та послуг, який може бути цікавим цьому споживачеві.

Формування рекомендованого переліку товарів та послуг відбувається на основі аналізу вибору схожих споживачів або встановлення схожості характеристик товарів. Також використовуються додаткові знання про предметну область. Зазвичай такі знання описують спосіб та умови використання товарів та послуг, що їх вибирає користувач рекомендаційної системи [2].

Однією із важливих проблем, яка виникає при розробці та впровадженні таких систем, є проблема холодного старту. Ця проблема пов'язана із відсутністю або неповнотою інформації про користувачів, товари або зв'язки між ними. Зазвичай вона виникає при появі в рекомендаційній системі нових користувачів, товарів, або при суттєвій зміні інтересів існуючих користувачів [3].

Традиційні підходи до побудови рекомендованого переліку об'єктів базуються на використанні існуючих даних про вибір користувача. У разі реєстрації у рекомендаційній системі нового користувача інформація про його вибір відсутня. Він вважається «холодним», тобто виникає проблем холодного старту.

Існуючі підходи до вирішення проблеми холодного старту орієнтовані на розв'язання задач, що відображають окремі аспекти цієї проблеми, пов'язані із відсутністю інформації про користувачів та об'єкти їх інтересу. Зокрема, моделюється поведінка користувача, його поточні та стратегічні інтереси, а також контекст прийняття рішень з вибору товарів або послуг. В роботі [4] модель динаміки змін інтересів користувача (темпоральної динаміки за визначенням автора) будується за допомогою методу градієнтного спуску. В роботі [5] запропоновано графове представлення зміни інтересів користувача. При використанні моделі застосовується випадковий пошук. Більш складні аспекти, пов'язані із описом поведінки декількох користувачів, моделюються за допомогою нейронної мережі [6]. Демографічні характеристики користувача використовуються при побудові рекомендацій в роботі [7]. Використання контекстних даних розглянуто в роботі [8].

Ряд робіт використовує фільтрацію вхідних даних на основі активного навчання [9, 10]. Головна ідея таких робіт полягає в тому, щоб до початку побудови рекомендацій підібрати релевантні вхідні дані.

Також використовується модель на базі багатошарового графу, що відображає зрізи по вибору користувачів у різні періоди часу [11].

Однак в цілому наведені підходи не використовують системне представлення ситуації холодного старту з урахуванням всіх її елементів та їх взаємодії.

Сукупність взаємодіючих споживачів та об'єктів задає гнучку систему обмежень для вибору нового споживача у ситуації холодного старту. Така система обмежень змінюється з часом, адаптуючись до змін уподобань споживачів. Вона дає можливість зменшити множину варіантів вибору нового споживача навіть за відсутності повної інформації про його інтереси.

Тому системне представлення вибору нового користувача рекомендаційної системи з формалізацією обмежень, пов'язаних із можливими інтересами існуючих споживачів, властивостями об'єктів, що є цікавими для споживачів, а також можливими контекстно-орієнтованими способами використання цих об'єктів є актуальною задачею.

Побудова такого представлення дає можливість реалізувати комплексний підхід до побудови рекомендацій, що заснований на виділенні обмежень методами інтелектуального аналізу даних та подальшому використанні отриманих обмежень для відбору релевантних вхідних даних, призначених для побудови рекомендацій.

2. Постановка задачі

Метою статті є розробка орієнтованих на обмеження моделей ситуації вибору користувача рекомендаційної системи у нормальному режимі роботи та у умовах холодного старту.

Для досягнення цієї мети вирішуються такі задачі:

– структуризація ситуації холодного старту з урахуванням інформації про користувача, об'єкти, контекст та час вибору;

– розробка орієнтованих на обмеження моделей вибору користувача у нормальному режимі роботи рекомендаційної системи та в умовах холодного старту;

– розробка узагальненого підходу до побудови орієнтованої на обмеження моделі вибору користувача рекомендаційної системи.

3. Формування рекомендацій в умовах обмежень холодного старту рекомендаційної системи

Проблема холодного старту виникає в результаті алгоритму взаємодії між користувачем та об'єктами його інтересу. Така послідовність змінюється внаслідок появи нових споживачів та об'єктів, а також внаслідок зміни контексту вибору цих об'єктів у статиці або динаміці.

Тому доцільно виділити такі випадки виникнення холодного старту:

– поява нового користувача;

– поява нового об'єкту;

– зміна контексту вибору об'єктів користувачем;

– зміни інтересів споживачів з часом, що призводять до динамічної зміни контексту вибору товарів та послуг.

Проблема появи нового користувача рекомендаційної системи характеризується неповнотою інформації про нього та утруднює реалізацію user-based підходу до побудови рекомендацій. Зазвичай потрібна інформація про нового споживача має два аспекти:

– персональні дані користувача;

– дані про його інтереси споживача.

Мінімальний набір персональних даних у більшості випадків вводиться до системи електронної комерції при реєстрації користувача. Рекомендаційна підсистема може їх отримати із бази даних відповідного сайту електронної комерції. Однак цих даних недостатньо для побудови точних рекомендацій.

Дані про інтереси споживача за замовчуванням представлені історією його пошуків товарів та послуг, а також покупок цих товарів й виставлених ним рейтингів. Історія пошуку відображається у вигляді послідовності переглянутих сторінок сайту інтернет-магазину, вибраних груп товарів, встановлених фільтрів на характеристики товарів, тощо.

Історія покупок зберігається в базі даних системи електронної комерції і разом із історією пошуку задають неявний зворотний зв'язок від споживача.

Рейтинг відображає явний зворотний зв'язок, оскільки в даному випадку використовується шкала оцінки товарів або послуг, що були вибрані споживачем.

У випадку нового користувача наведені дані відсутні в рекомендаційній системі, що не дає можливості сформувати точні рекомендації.

Додаткові дані про користувача зазвичай отримують із соціальних мереж. Слід зазначити, що при аналізі соціальних мереж використовуються два підходи.

Згідно першого підходу з мережі вилучаються додаткові персональні дані користувача, наприклад про його вподобання, освіту, регіон проживання, тощо.

Другий підхід передбачає аналіз повідомлень користувачів соціальної мережі. Використовуються методи інтелектуального аналізу тексту для вилучення понять, об'єктів, та зв'язків між ними. На основі аналізу отриманих знань формується опис контексту прийняття рішень користувачем. Наприклад, аналіз повідомлень про плани на уикенд, про персональні інтереси використовується мережами магазинів для прогнозування та підбору відповідного асортименту товарів для споживача.

Проблема холодного старту при появі нового об'єкту, тобто товару або послуги пов'язана із відсутністю інформації про його характеристики. Такі характеристики також доцільно розділити на дві групи.

Перша група відображує інтерес користувачів до цього об'єкту. Дана інформація в цілому є неявною і може бути отримана лише на основі аналізу рейтингів товару, виставленого декількома користувачами або історії покупок.

Проблема холодного старту при появі нового товару в даному випадку пов'язана із відсутністю інформації про його вибір, що не дозволяє сформувати рекомендації згідно user –based підходу.

Друга група характеристик відображає властивості товару, що є цікавими для користувача рекомендаційної системи. На практиці така інформація купується у спеціалізованих фірм, що ведуть бази даних із технічними характеристиками широкого спектру товарів. Відсутність цієї інформації не дає можливості сформувати рекомендації користувачеві за допомогою item-based підходу.

Проблема холодного старту при зміні контексту вибору користувача виникає у трьох випадках:

– зміна персональних вподобань користувача при зміні глобального контексту, тобто зміні місця проживання, навчання, роботи, тощо.

– в результаті поліпшення або серйозних змін у характеристиках об'єктів (товарів, послуг): появи товарів та послуг із новими можливостями, що впливають на спосіб їх використання, наприклад електровелосипеди замість велосипедів, смартфони замість телефонів, тощо;

– змін у організаційній системі, в рамках якої оперує споживач; прикладами таких контекстних перемін є зміна корпоративної культури підприємства; зміна організаційної взаємодії у державі (наприклад онлайн-банкінг, електронне урядування).

Проблема холодного старту на основі динамічної зміни контексту вибору товарів та послуг зазвичай розглядається як проблема постійного або циклічного холодного старту [3]. Вона пов'язана із циклічними змінами інтересів споживачів. Така

проблема є характерною для споживачів, що нерегулярно взаємодіють із рекомендаційною системою. Відповідно, їх історія пошуку й вибору товарів та послуг з часом перестає бути релевантною. Такі споживачі мають розглядатись як «холодні» в рамках рекомендаційної системи.

Структуризацію проблеми холодного старту згідно наведеного опису представлено у табл. 1.

Таблиця 1

Характеристики ситуацій холодного старту

Ситуація	Неповні дані	Додаткові дані
Новий користувач	Персональні дані; інформація про інтереси користувача	Неявний зворотний зв'язок: історія пошуку; історія покупок. Явний зворотний зв'язок: рейтинг. Персональні дані, що можуть бути отримані із соціальних мереж.
Новий об'єкт (товар, послуга)	Інформація про характеристики об'єкту	Неявні характеристики з точки зору користувача: рейтинг. Властивості об'єктів, можуть бут отримані від спеціалізованих фірм, що ведуть базу специфікацій.
Зміна контексту вибору користувача	Спосіб використання об'єкту споживачем	Персональні дані, що відображають глобальні зміни у контексті використання об'єктів споживачем. Додаткові знання про товари, що визначають сферу їх застосування. Загальні знання про підходи до використання об'єктів у предметній області
Циклічні зміни потреб користувача	Залежність зміни інтересів від часу	Інтервал часу, на якому відбувається цикл зміни інтересів споживача

Виконана структуризація проблеми холодного старту показує, що в усіх чотирьох випадках ця проблема пов'язана із виникненням неявних обмежень при побудові рекомендацій. Кожне із таких обмежень характеризується набором змінних, значення яких відомі в нормальному режимі функціонування рекомендаційної системи та є відсутніми у ситуації холодного старту.

Таким чином, модель ситуації холодного старту доцільно розглядати в парадигмі моделювання на основі обмежень. Ці обмеження задаються через можливі значення змінних, які характеризують вибір споживача. Іншими словами, можливі інтереси споживача визначаються шляхом обмежень на значення змінних, що характеризують товари та послуги, контекст, час вибору, тощо.

Наприклад, при виборі користувачем монітору в якості обмежень можуть виступати значення змінних, що характеризують розмір діагоналі екрану, тип матриці, тип роз'єму для підключення сигналу, тощо. Відповідно, вибір декількох схожих за інтересами користувачів також має спільні обмеження пов'язані, наприклад із категорією об'єктів, їх вартістю, тощо.

Для подальшої формалізації ситуації холодного старту розглянемо спочатку представлення вибору споживача у обмеженнях.

Обмеження, що впливають на формування рекомендацій, можна розглядати як кон'юнкцію предикатів, Q_i над змінними, що характеризують користувача, об'єкти та їх взаємодію у часі:

$$C = \{Q_i(X_i)\}, X_i = \{x_1,...,x_k,...\}, \quad (1)$$

де C – множина обмежень; x_k – змінна, що визначає властивість користувача, об'єкту, контексту або часу.

Тоді, на основі результатів роботи [12], ситуація вибору користувача M може бути представлена множиною змінних $X = \{x_j\}$, що характеризують різні аспекти формування рекомендацій, множиною обмежень C, описом стану предметної області D, а також зв'язками у предметній області R:

$$M = (X, C, D, R). \quad (2)$$

Опис предметної області задає відношення між змінними із предикатів у вигляді множини предикатів Q_i, тобто визначає такі значення змінних, для яких предикат Q_i буде мати значення «істина» у даній предметній області.

$$D = \{v_j : \forall j\ x_j = v_j | x_j \in X\}, \quad (3)$$

де v_j – значення, яке приймає змінна x_j у предметній області D.

r задає відображення змінних на їх значення у даній предметній області:

$$r : X \to D \quad (4)$$

Зв'язки між властивостями споживачів, об'єктів, контексту та часу у предметній області задаються на основі n-арних відношень на значеннях змінних x_j:

$$R = \{R_i(v_1,...,v_k,...)\}. \quad (5)$$

Дані відношення дають можливість визначити істинність предиката Q_i таким чином:

$$Q_i = true | \exists r \wedge \exists R_i. \quad (6)$$

Згідно (6) для моделювання ситуації вибору споживача необхідно спочатку для кожної змінної визначити множину її допустимих значень, а також встановити можливі зв'язки для значень цих змінних у предметній області.

Наприклад, при побудові обмежень на опис монітору необхідно задати допустимі значення його

параметрів. Тобто ситуація вибору користувача має обмеження виду:

$$Q_i = \textit{Роздільна\_здатність} \wedge \wedge \textit{ частота} \wedge \textit{ яскравість} \wedge \wedge \textit{ контрастність} \wedge ... \quad (7)$$

Реальний монітор в інтернет-магазині буде мати комбінацію значень виду:

$$R_i = 2560 * 1440 \wedge \wedge 60\text{Гц} \wedge 300\text{Кд} \wedge \wedge 3000:1 \wedge ... \quad (8)$$

Відношення між Q_i та R_i встановлюються через відображення r, тобто для даного прикладу:

$$\begin{aligned} \textit{Роздільна\_здатність} &\rightarrow 2560*1440, \\ \textit{частота} &\rightarrow 60\ \text{Гц}, \\ \textit{яскравість} &\rightarrow 300\ \text{Кд}, \\ \textit{контрастність} &\rightarrow 3000 \end{aligned} \quad (9)$$

Однак у більш загальному випадку у інтернет – магазині продається набір моніторів, тобто:

$$\begin{aligned} \textit{Роздільна\_здатність} &\rightarrow 2560*1440, \\ \textit{Роздільна\_здатність} &\rightarrow 1660*1200, \\ \textit{Роздільна\_здатність} &\rightarrow 1920*1080... \end{aligned} \quad (10)$$

Представлений приклад показує, що орієнтований на обмеження опис ситуації вибору користувача дає можливість зменшити кількість об'єктів, що пропонуються споживачеві. У даному прикладі зменшення відбувається шляхом врахування допустимих комбінацій значень властивостей товарів.

Також обмеження можуть враховувати комбінацію властивостей споживача: вік, регіон проживання, тощо.

Контекст враховується у обмеженнях на основі додаткової інформації про властивості предметної області, наприклад про спосіб застосування рекомендованого товару.

Приклад контексту використання монітору із сайту elmir.ua наведено на рис. 1.

Дополнительно применена продвинутая технология многозонного вертикального совмещения, которая обеспечивает сверхвысокий коэффициент статического контраста, формируя более яркую, живую картинку. На этом дисплее без труда можно работать в стандартных офисных программах, но особенно эффективен такой экран для просмотра фотографий, веб-страниц, фильмов и игр, и работы с мощными графическими приложениями. Технология оптимизированной

Рис. 1. Приклад контексту використання об'єкту

Із прикладу представленого на рис. 1 контексту використання монітору видно, що контекст задає множину способів його застосування і передбачає можливості роботи як з офісними та графічними програмами, так і з ігровими додатками. Очевидно, що контекст впливає на вибір користувача, в тому числі і в ситуації холодного старту.

Розглянемо опис контекстних обмежень у предметній області згідно наведеного прикладу. Фрагмент контексту характеризується такими змінними:

$x_1 = \textit{«вертикальне суміщення»}$;
$x_2 = \textit{«офісне застосування»}$;
$x_2 = \textit{«робота з графікою»}$.

Відповідно, формальне представлення контекстного обмеження Q_i для даного спрощеного прикладу має вигляд:

$$Q_i = x_1 \wedge x_2 \wedge x_3. \quad (11)$$

Зв'язки між властивостями контексту в предметній області мають вигляд:

$$R_i = (v_1 = true) \wedge (v_2 = true) \wedge (v_3 = true). \quad (12)$$

Вибір користувача з формальної точки зору означає, що його вимога(важливі для його властивості товарів та послуг) задовольняються в рамках обмежень моделі (2). Тому для потреба споживача β має бути істинною в моделі M ситуації вибору, тобто має місце умова :

$$(X, C, D, R) \models \beta. \quad (13)$$

Оскільки обмеження C задають зв'язок між елементами моделі M, то умова (13) може бути переписана у вигляді $C \models \beta$.

Умови холодного старту задаються підмножиною обмежень Cl. Модель вибору користувача в умовах обмежень холодного старту має вигляд:

$$M_{Cl} = (X, Cl, D, R), Cl \subset C. \quad (14)$$

Такі обмеження залежать від того, яка ситуація холодного старту виникла. Наприклад, для нового користувача така підмножина має вигляд:

$$Cl = C_{Item} \bigcup C_{Context} \bigcup C_{Temporal}, \quad (15)$$

де C_{Item} – обмеження по об'єктам; $C_{Context}$ – контекстні обмеження; $C_{Temporal}$ – темпоральні обмеження.

Запропонований підхід до побудови рекомендацій в умовах холодного старту рекомендаційної системи використовує розроблену модель M_{Cl}.

Ключова концепція, яку положено в основу підходу, полягає у відборі релевантних вхідних даних на основі формування предикатних обмежень, що характеризують контекстно-орієнтовані закономірності вибору споживачів. Підхід складається з таких фаз.

Фаза 1. Визначення попередньої підмножини обмежень Cl в залежності від причин виникнення ситуації холодного старту.

Фаза 2. Уточнення підмножини обмежень з урахуванням інтервалу часу актуальності значень змінних у предметній області.

Фаза 3. Побудова залежностей на основі аналізу вхідних даних. На даній фазі використовуються

підходи інтелектуального аналізу даних для побудови причинно-наслідкових темпоральних залежностей.

Фаза 4. Використання отриманих залежностей для підготовки набору вхідних даних рекомендаційної системи.

На даній фазі може бути виконана як фільтрація, так і коригування вхідних даних. Мета фільтрації полягає в тому, щоб відібрати дані, що відповідають інтересам споживача на визначеному інтервалі часу. Мета коригування полягає в тому, що деталізувати дані, що відповідають базовим закономірностям вибору у даній предметній області.

Фаза 5 Використання традиційних методів побудови рекомендацій, наприклад методу колаборативної фільтрації.

4. Висновки

Розглянуто проблему підтримки вибору користувача при холодному старті рекомендаційної системи. Дана проблема виникає внаслідок недостатньої інформації про користувачів, об'єкти їх інтересу або зв'язки між ними. На практиці холодний старт рекомендаційної системи виникає в результатів появи нових користувачів, товарів, або суттєвої зміни інтересів існуючих споживачів.

Виконана структуризація даної проблеми дала можливість виділити різні аспекти холодного старту: поява нового користувача; поява нового об'єкту; зміна контексту вибору об'єктів користувачем; циклічні зміни інтересів споживачів з часом.

Запропоновано модель вибору об'єктів у нормальному режимі роботи рекомендаційної системи у вигляді системи предикатних обмежень, що визначають допустимі комбінації властивостей користувачів, об'єктів, контексту, а також динаміки змін інтересів користувачів.

Запропоновано модель вибору об'єктів в умовах холодного старту. Дана модель використовує підмножину обмежень для формування рекомендацій у нормальному режимі роботи.

Запропоновані моделі забезпечують можливість відфільтрувати вхідні дані з урахуванням найбільш суттєвих закономірностей вибору споживачів у даному контексті на даному інтервалі часу, що дає можливість спростити побудову рекомендацій для нових споживачів і нових об'єктів.

Запропоновано підхід до побудови рекомендацій в умовах обмежень холодного старту. Підхід передбачає формування підмножини причинно-наслідкових та темпоральних обмежень на основі інтелектуального аналізу вхідних даних рекомендаційної системи, а також подальше використання цих обмежень при побудові рекомендацій в умовах холодного старту.

Список літератури:

[1] *Ricci F., Rokach L., Shapira B.* Recommender systems. Handbook. - Second Edition. – 2015. – 1008 p.

[2] *Aggarwal C.* Recommender systems: The Textbook. – New York: Springer. – 2017. – 498 p.

[3] *Bernardi L. et al.* The Continuous cold start problem in e-commerce recommender systems // The Computing Research Repository (CoRR) in arXiv. – Vol. 1508.01177. – 2015. – P. 1-6.

[4] *Koren Y.* Collaborative Filtering with Temporal Dynamics // International conference on knowledge discovery and Data Mining (ACM SIGKDD). – 2009. – P. 447-456.

[5] *Xiang L., Yuan Q.* Temporal recommendation on graphs via long-and short-term prefence fusion // International Conference on Knowledge Discovery and Data Mining. – 2010. – P. 723-732.

[6] *Elahi M., Ricci F., Rubens N.* A survey of active learning in collaborative filtering recommender systems // Computer Science Review. – 2016. – Vol. 20. – P. 29-50.

[7] *Lika B., Kolomvatsos K., Hadjiefthymiades S.* Facing the cold start problem in recommender systems. // Expert Systems with Applications. – 2014. – Vol. 41(4). – P. 2065-2073.

[8] *Adomavicius G. et al.* Incorporating contextual information in recommender systems using a multidimensional approach // ACM Transactions on Information Systems. – 2005. – Vol. 23(1). – P. 103–145.

[9] *Luo C., Cai X.* Self-training Temporal Dynamics Collaborative Filtering // The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'14). – 2014. – P. 461-472.

[10] *Zhu Y. et al.* Addressing the item cold-start problem by attribute-driven active learning // IEEE Transactions on Knowledge and Data Engineering. – 2019. – P. 1-14.

[11] *Chalyi S., Pribylnova I.* The method of constructing recommendations online on the temporal dynamics of user interests using multilayer graph // EUREKA: Physics and Engineering. – 2019. – Vol. 3. – P. 13-19.

[12] *Junker U.* QUICKXPLAIN: Preferred explanations and relaxations for overconstrained problems// Proceedings of the 19th National Conference on Artificial Intelligence (AAAI '04). – 2004. – P. 167–172.

Надійшла до редколегії 3.04.2019

УДК 519.62

Д. О. Володін, І. В. Афанасьєва
Програмна інженерія, ХНУРЕ, м. Харків, Україна, volodindmitriy121@gmail.com
Програмна інженерія, ХНУРЕ, м. Харків, Україна, iryna.afanasieva@nure.ua

АНАЛІЗ МЕТОДІВ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ АВТОМОБІЛЬНИХ РЕЄСТРАЦІЙНИХ НОМЕРІВ

Проведено аналіз існуючих методів сегментації зображень. Було розроблено модифікований алгоритм, який сегментує зображення реєстраційних номерів автомобіля. Запропонована модифікація призначена для сегментації нетекстурованих або слабко текстурованих зображень. Мета – виділення об'єктів (в тому числі декількох) на зображенні і видалення фону.

Для аналізу методів обрано зображення, типові для поставленої задачі. Реєстраційні знаки автомобілей було розділено на 3 групи.

Результати машинних експериментів показали, що комбінація методів дає хороші результати як для одного, так і для декількох об'єктів, що мають навіть незначні відмінності від фону за яскравістю. У великих об'єктів, що складаються з декількох частин різної яскравості, вдалося виділити ключові частини.

Таким чином, досягнуто певну універсальність розробленого модифікованого алгоритму щодо різних типів зображень реєстраційного знаку автомобіля, що представлено порівнянням.

СЕГМЕНТАЦІЯ, ЗОБРАЖЕННЯ, РЕЄСТРАЦІЙ НОМЕР, РОЗПІЗНАВАННЯ, АЛГОРИТМ, МЕТОД, АНАЛІЗ

Д.О. Володин, И.В. Афанасьева. Анализ методов сегментации изображений автомобильных регистрационных номеров. Проведен анализ существующих методов сегментации изображений. Был разработан модифицированный алгоритм, который сегментирует изображение регистрационного номера автомобиля. Предложенная модификация предназначена для сегментации нетекстурированных или слабо текстурированных изображений. Цель – выделение объектов (в том числе нескольких) на изображении и удаление фона.

Для анализа методов избраны изображения, типичные для поставленной задачи. Регистрационные знаки автомобилей были разделены на 3 группы.

Результаты машинних экспериментов показали, что комбинация методов дает хорошие результаты как для одного, так и нескольких объектов, имеющих даже незначительные отличия от фона по яркости. В крупных объектах, состоящих из нескольких частей различной яркости, удалось выделить ключевые части.

Таким образом, достигнута определенная универсальность разработанного модифицированного алгоритма по различным типам изображений регистрационного знака автомобиля, что представлено сравнением.

СЕГМЕНТАЦИЯ, ИЗОБРАЖЕНИЕ, РЕГИСТРАЦИОННЫЙ НОМЕР, РАСПОЗНАВАНИЕ, АЛГОРИТМ, МЕТОД, АНАЛИЗ

D.O. Volodin, I.V. Afanasieva. Analysis of images segmentation methods for car registration numbers. The analysis of existing methods of image segmentation was held. A modified algorithm was developed, which segments the images of the car registration number.

The proposed modification is intended for segmentation of non-textured or poorly textured images. Purpose – to select objects (including several) in the image and delete the background.

For the analysis of methods selected images that are typical for the task.The registration marks of the cars were divided into 3 groups.

The results of machine experiments showed that the combination of methods gives good results for both one and also for several objects, which have even slight differences from the background on brightness. In large objects, which consist of several parts of different brightness, it was possible to identify the key parts.

Thus, a certain versatility of the developed modified algorithm has been achieved for different types of images of the registration mark of the car, which is represented by comparison.

SEGMENTATION, IMAGE, LICENCE PLATE, RECOGNITION, ALGORITHM, METHOD, ANALYSIS

Вступ

Розпізнавання об'єктів на зображеннях одна з галузей інформаційних технологій що найбільш інтенсивно розвивається. Необхідність в такому розпізнаванні виникає в самих різних галузях – від військової справи і систем безпеки домедичної діагностики та контролю дорожнього руху.

Метою роботи є дослідження існуючих методів і розробка модифікованого алгоритму розпізнавання символів, що забезпечують аналіз та обробку інформації на зображенні з метою ідентифікації реєстраційного знаку автомобіля.

Розпізнавання складається з трьох основних етапів: первинної обробки, сегментації та розпізнавання. Первинна обробка складається з бінарізації [1] вхідного зображення, що скорочує обсяг інформації для подальшого аналізу. Процес

бінарізації – це перетворення кольорового (або в градаціях сірого) зображення в двокольорове чорно-біле. Після бінарізації відбувається сегментація номерних знаків і розпізнавання окремих символів, що зображено на номерному знаку.

1. Загальна математична модель сегментації

Нехай $D(m \times n)$ – растр або область поля зору, на якому задано зображення $B(i, j)$; $D_k \subset D$ – область k-го об'єкта $k = 1, 2, ..., s$; $D_\text{ф} \subset D$ – область фону. Вважаємо $D_1 \cup D_2 \cup ... \cup D_S \cup D_\text{ф} = D$, $D_i \cap D_j = \varnothing$, $i \neq j$.

Розглядаємо дискретне зображення $B(i, j)$, $i = 0, ..., m$, $j = 0, ..., n$. Зображення є сукупність зображень окремих об'єктів і фону.

Завдання сегментації зображень полягає в побудові предиката виду:

$$\varpi(i, j) = \begin{cases} k, \text{ якщо } (I, j) \in D_k \\ 0, \text{ якщо } (i, j) \in D_\text{ф} \end{cases}$$

На змістовному рівні це означає, що кожна точка (i, j) OD зображення $B(i, j)$ отримує смислову мітку з номером $p(i, j)$. Строго кажучи, в ідеалі точки з одною міткою утворюють область одного окремого об'єкта, міткою 0 розмічається область фону. Математична модель передбачає розмітку непересічних об'єктів або видимих їх частин при закритті одного об'єкта іншим. В результаті декомпозиція загальної задачі перетворюється на підзадачі: сегментація, розпізнавання, поліпшення (фільтрація, усунення шумів, підкреслення границь та ін.). При такій постановці, інтерпретація і розуміння частково заслонених об'єктів відносяться вже до області розпізнавання, а не до сегментації.

2. Класифікація

Процес кластеризації зображень – пошуку в них однорідних областей, називається сегментацією. Вона вважається другим етапом аналізу зображень, це – базова процедура практично у всіх завданнях обробки зображень за допомогою систем комп'ютерного зору.

Сегментація включає в себе елементи фільтрації шумів і виділення зображень. Класифікація методів сегментації описується в різних роботах [2, 3], присвячених даному питанню. Часто їх підрозділяють на ті, які виділяють області однорідної яскравості або кольору, і ті, які визначають однорідності інших властивостей, найчастіше текстури. Інший важливий критерій, за яким можна класифікувати методи сегментації, – це характеристики областей. Вони, в одному випадку, можуть бути задані заздалегідь (наприклад, бібліотека еталонів, текстур), а в іншому – їх необхідно отримати в процесі сегментації.

Для сегментації зображень було розроблено кілька універсальних алгоритмів і методів. Оскільки спільного рішення для задачі сегментації зображень не існує, часто ці методи доводиться поєднувати зі знаннями в предметній області, щоб ефективно вирішувати певну задачу в її предметної галузі.

Основні методи сегментації, порівняння яких розглядається в даній роботі, наведено на рис. 1.

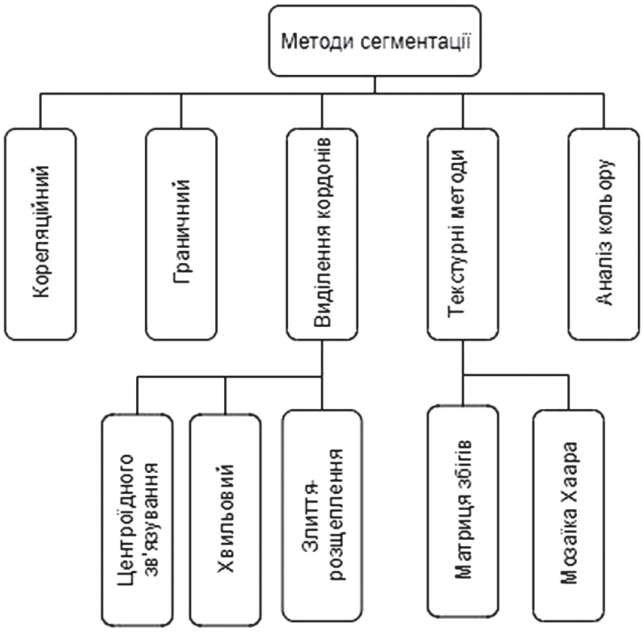

Рис. 1. Класифікація методів сегментації

Кореляційні методи застосовуються в разі, якщо відомі зразки об'єктів. Вони ефективні в системах прикладного телебачення і відносяться більше до галузі розпізнавання зображень [4]. Порогові методи застосовуються при існуванні стабільних відмінностей в яскравості окремих областей. Методи нарощування кордонів ефективні при наявності стійкої зв'язності всередині окремих сегментів. Метод виділення кордонів слід використовувати, якщо межі досить чіткі і стабільні. Для опису та сегментації властивостей зображень (однорідності, регулярності) використовують текстурні методи, які умовно поділяються на дві категорії: статистичні та структурні. Прикладом статистичного підходу є використання матриць збігів, що формуються шляхом вихідних зображень; структурного – мозаїка Хаара. Методи, засновані на аналізі кольору, по своїй суті є комбінованими.

3. Методи аналізу різниці яскравостей

3.1. Граничний метод

Припустимо задано зображення $B(i, j)$, $s = 1$ (один об'єкт), яскравість точок знаходиться в межах $[T_1, T_2]$, а яскравість точок фону не входять. Якщо $B(i, j) \in [T_1, T_2]$, то точку (i, j) вважаємо належною до області об'єкта, в іншому випадку – області фону. У разі $s>1$ повинні бути відомі відрізки $[T^k_1, T^k_2]$, в межах яких знаходяться яскравості

k-x об'єктів. Ці відрізки не повинні перетинатися. Розмітка точок здійснюється за допомогою відображення.

Проблемною частиною цього метода є визначення порогових величин. Для цього проводиться аналіз гістограми яскравостей. У випадку з одним об'єктом ($s = 1$) на гістограмі має бути два максимуми та поріг обирається між цими двома максимумами.

3.2. Виділення кордонів

При такому способі сегментації об'єкти представляються їх межами. Граничними прийнято вважати точки різкого перепаду функції яскравості. Для знаходження граничних точок використовується чисельне диференціювання [5]. Найпоширенішим є градієнтний метод, відомий також як метод контрастування або просторового диференціювання [6]. Застосовуючи маску (фільтр) до зображення, отримують так зване зображення градієнтів. Воно відрізняється від вихідного підкресленими перепадами яскравості. Точка (i, j) належить контуру, якщо яскравість зображення градієнтів перевищує певний поріг, який може визначатися загістограмами.

3.3. Аналіз кольору

Сегментація шляхом аналізу кольору заснована по суті на його впізнаванні. Ознаками служать, наприклад, три компоненти (координати) від функції випромінювання [7] в кожній точці (i, j):

$$C_1 = \int_{\lambda} B(\lambda) K_1(\lambda) d\lambda \, , \; l = 1, 2, 3.$$

Спектральні кривічутливості $K_1(\lambda)$, $K_2(\lambda)$, $K_3(\lambda)$, можуть відповідати функціям трьох видів колбочок людського ока, однак на практиці використовуються багато інших систем кодування колірних компонент, включаючи двох- і чотирьох- компонентні системи [8]. Сегментація може проводитися методами, викладеними в цій статті покомпонентно, а впізнавання – шляхом дешифрування значень координат кольору.

4. Методи нарощування кордонів

4.1. Центроїдного зв'язування

Для застосування цього методу необхідна апріорна інформація про об'єкти, а саме, одна або кілька стартових точок. Задаються стартові точки $a_1, \ldots, a_k$, яким відповідно присвоюються мітки $\lambda_1, \ldots, \lambda_k$. Точки, що мають мітку λ_i, утворюють безліч S_i. Після вибору стартових точок проводиться процес розмітки, в ході якого розглядаються всі точки A множин S_i. Якщо сусідня з A точка N така, що

$$|B(A) - B(N)| < T,$$

і не має мітки, то точці N присвоюється мітка λ_i. Після розбиття на області можливе проведення злиття областей, тобто присвоєння точкам з мітками λ_x і λ_y єдиної мітки $\min(\lambda_x, \lambda_y)$. При довільному порядку розмітки метод найбільш придатний для простих зображень (з одним об'єктом). Для більш складних зображень застосовується так званий хвильовий спосіб перегляду точок.

4.2. Хвильовий метод

Після вибору стартових точок проводиться процес, що складається з ітерацій. На кожній з ітерацій розглядаються точки множин S_i, крім тих, що були включені в S_i на даній ітерації. Для точки (i, j) розглядаються її сусідні точки. Однією з них може бути присвоєна мітка λ_i, якщо виконується умова, що описана в попередньому пункті. Після того як аналіз виконано для всіх точок множин S_i, крім тих, що були додані на даній ітерації, проводиться аналіз точок з S_{i+1}. Точки множин S_i, додані на k-й ітерації, називаються фронтом $F_k(\lambda_i)$, з'єднання, що представлено формулою 1.

$$\bigcup_i F_k(\lambda_i), \qquad (1)$$

називається хвилею.

4.3. Метод вододілів

Метод вододілів є модифікацією хвильового методу. Вводиться множина точок $S = \{i, j, B(i, j)\}$, що називається поверхнею. На поверхні вводиться поняття шляху від точки S_m до точки S_n. Шляхом називається послідовність $\{S_m, S_m + 1, \ldots, S_n - 1, S_n\}$, де S_i є сусідньою до S_i+1. Незростаюча шляхом називається така послідовність $\{S_i\}$, що

$$\vee S_m(i_m, j_m, B(i_m, j_m)), S_n(i_n, j_n, B(i_n, j_n)):$$
$$: m \geq n \leftrightarrow B(i_m, j_m) \leq B(i_n, j_n).$$

Точка $s \in S$ називається локальним мінімумом, якщо не існує незростаючого шляху з початком в точці s. Після визначення локальних мінімумів переходять до так званого процесу заповнення басейнів. Знаходять яскравість $B = \min_{i\,j} (B(i, j))$. Проводять ітерації, на кожній з яких збільшують яскравість B на одиницю, поки не досягнуть максимального рівня яскравості. На кожній ітерації проводять розмітку точок з яскравістю B методом хвиль. В результаті отримуємо розбиття зображення на басейни. Іноді виділяють так звані точки вододілу, тобто точки, що мають сусідів, що належать двом або більше басейнам. Залежно від завдань яскравість B можна збільшувати не до максимального значення Bmax, а до будь-якого порогового значення.

4.4. Злиття-розщеплення

Метод полягає в розбитті зображення на квадрати. Потім проводиться аналіз однорідності квадратів, найчастіше аналізується однорідність яскравості. Якщо квадрат не задовольняє умові однорідності, то він замінюється чотирма

підквадратами. Якщо ж ділянка з чотирьох сусідніх квадратів виявляється такою, що для неї виконується умова однорідності, то ці чотири квадрати об'єднаються в один.

Результатом злиття-розщеплення може служити деяка структура з інформацією про квадрати, найчастіше – граф, а може бути і зображення, в якому всі пікселі всередині однорідної області мають однакову яскравість.

5. Текстурні методи

5.1. Матриці збігів

Цей метод входить до групи статистичних методів. Шляхом обчислення для кожної ділянки так званої матриці збіжностей він дозволяє визначити, чи містять ділянки зображення текстури одного класу.

Розглядається ділянка $N \times N$. Маємо множину яскравостей $\{B(i, j),\ i = 1, ..., N;\ j = 1, ..., N\}$ з G градаціями сірого. Визначається вектор зміщення $d=(dx, dy)$. Вводиться матриця збігів $G \times G$, що відповідає числу випадків матриці P_d. Елемент (a, b) матриці P_d дорівнює числу випадків, коли від точки з яскравістю a на відстані, що визначається вектором d, знаходиться точка з яскравістю b. Формально слід представити (2):

$$P_d(a,b) = \sum p\big(a,b,(r,s)-(t,v)\big), r, s, \qquad (2)$$

де $p\big(a,b,(r,s)-(t,v)\big) = 1$ або 0,
та де 1, якщо $B(r,s) = a, B(t,v) = b$, а 0, в іншому випадку, та $(t,v) = (r + dx, s + dy)$.

На основі матриці P_d можуть бути обчислені різні характеристики $\sum_a \sum_b P_d^2(a,b)$, (Енергія), $-\sum_a \sum_b P_d(a,b) log P_d(a,b)$, (Ентропія).

5.2. Метод мозаїки Хаара

Сегментація з використанням цього методу складається з трьох кроків: побудови примітивів, складання мозаїки, аналізуелементів мозаїки. Найчастіше для отримання примітивів до зображення застосовують фільтри, як при виділенні меж. Потім обирають точки локальних максимумів, до яких застосовують метод нарощування, в результаті чого отримують компоненти з 8-ми зв'язаних елементів. Отримані таким чином компоненти або точки локальних максимумів визначають як примітиви. Далі будується мозаїчне розбиття Хаара для примітивів. Розглянемо побудову для випадку, коли примітиви є точками.

Припустимо дана множина S з трьох і більше примітивів та не всі точки лежать на одній прямій. Роздивимося пару точок P і Q. Побудуємо пряму – геометричне місце точок, рівновіддалених від P і Q.

Отримаємо дві півплощини H_Q^P і H_P^Q .Для будь-якої точки P можна провести таке розбиття з усіма $Q \in S$. Перетин визначає багатокутник, всі крапки якого ближче до P, ніж до будь-якої іншої точки. Такий багатокутник називають багатокутником Хаара для даної точки. Розглядають безліч багатокутників, назване каскадом Хаара. Багатокутники із загальними властивостями об'єднують в області. Для обчислення властивостей часто використовують центр ваги і момент площі багатокутника. Момент площі $(p + q)$-го порядку для багатокутника щодо примітиву з координатами (x, y) визначають так:

$$m_{pq} = \iint_R (x - \bar{x})^P (y - \bar{y})^q \, dxdy,$$

де $(p + q) = 0, 1, 2, ...;\ \bar{x}, \bar{y}$ – координати центру ваги багатокутника.

Часто використовуються ознаки:

$$f_1 = m_{00}; f_2 = \sqrt{(\bar{x})^2 + (\bar{y})^2}\ ,\ f_3 = \operatorname{arctg}\left(\frac{\bar{y}}{\bar{x}}\right)$$

6. Модифікований алгоритм

В результаті аналізу існуючих методів сегментації для поставленої задачі розпізнавання номерних знаків автомобіля, де символи потрібно відокремлювати від можливої тіні зверху чи знизу, шумів або інших пошкоджень номерного знаку, було вирішено розробити евристичний алгоритм сегментації на основі методу нарощування кордонів.

Запропонована модифікація призначена для сегментації нетекстурованих або слабкотекстурованих зображень. Мета – виділення об'єктів (в тому числі декількох) на зображенні і видалення фону. Виділяють об'єкти, що цілком потрапляють в межі зображення.

Процедури сегментації і розпізнавання працюють з бінарним зображенням, тобто тільки з чорними і білими пікселями. Тому перш ніж виконувати роботу з даними процедурами, потрібно вихідне кольорове зображення привести до бінарного виду. Це завдання вирішується в два етапи. На першому етапі кольорове зображення перетворюється в чорно-біле і представляється в градаціях сірого. Для кожного пікселя обчислюється його яскравість в межах від нуля до 255. Рівню яскравості 0 відповідає чорний колір, рівню 255 – білий.

Другим етапом є, власне, бінарізація. Результат бінарізації залежить від заздалегідь заданого параметра – співвідношення чорних пікселів і загальної їх кількості на зображенні.

Алгоритм бінарізації зображення складається з наступних кроків:

– створюємо одновимірний масив I з 256 елементів (від 0 до 255). Заповнити його нулями;

–проходемо піксель за пікселем усі зображення. Збільшити на одиницю значення в осередку масиву I, відповідної яскравості пікселя $i(I[i]++)$.

У підсумку, значення кожного осередку масиву $I[i]$ буде дорівнювати кількості пікселів яскравості рівня i на цілому зображенні;

– на цьому кроці визначається поріг яскравості a. Припустимо, N – загальна кількість пікселів (висота помножена на ширину), k – коефіцієнт, що визначає кількість чорних пікселів. Тоді kN дорівнюватиме бажаної кількості чорних пікселів на бінарному зображенні. Підсумовуємо значення осередків масиву, починаючи з нульової до тих пір, поки значення цієї суми не перевищуватиме kN. Індекс останньої суми осередку і буде порогом a;

– повторно проходимо піксель за пікселем усі зображення. Порівнюємо рівень яскравості кожного пікселя з порогом a. Якщо цей рівень менше або дорівнює a, то піксель стає чорним, інакше – білим.

Отримавши бінарне зображення переходимо до етапу сегментації.

Сегментація складається з трьох етапів, кожен з яких описаний вище, як її самостійний метод. На першому етапі до зображення застосовується метод злиття-розщеплення. В результаті видаляються незначні неоднорідності. Наступний етап – виділення кордонів. При дослідуванні використовується оператор Лапласа [9] як найбільш універсальний і який рівномірно відображає перепади яскравостей в усіх напрямках. Виходить нове зображення, що складається з контурів. У програмному модулі, що також було розроблено, застосовується нормалізація яскравостей, щоб діапазон їх зміни був фіксованим. Третій етап – застосування до нового зображення розроблене доопрацювання для методу нарощування кордонів. В результаті виділяється фон і деяке число областей, не помічених як фон, які вважаємо виділеними об'єктами.

Сегментація проходить зліва направо. Відшукуємо точку, що підозріла на приналежність символу. Зображення сканується вертикальними смугами. Маючи на увазі ймовірність тіні, пропускаємо точки пов'язані з верхом і низом растра. Точка, яка не має такого зв'язку, передається алгоритму виділення символу.

Алгоритм пошуку точки, підозрілої на приналежність символу:

1. Початковою стає лівіша вертикальна лінія пікселів.

2. Починаючи від верхнього пікселя, знижуємся вниз, поки не зустрінемо білий піксель. Так ми отримуємо верхню межу області пошуку.

3. Починаючи від нижнього пікселя, піднімаємося вгору, поки не зустрінемо білий піксель. Так ми отримуємо нижню межу області пошуку.

4. У виділеній області пошуку шукаємо першу ліпшу чорну крапку, якщо такої не знайдено, то поточною стає наступна правіша лінія. Переходимо на крок 2.

5. Знайдена точка передається алгоритму виділення символу.

Від обраної точки будується чотиризв'язна область. Таким чином, виділяється символ. Накладаються обмеження на ширину області, з огляду на можливі зливання символів тінню і шумом.

Модифікований метод доопрацьований таким чином, що в ньому враховується ймовірність появи тіні на номерному знаку зверху чи знизу.

Рекурсивний алгоритм виділення символу складається з таких кроків:

1. Точка позначається належною символу.

2. Якщо верхня сусідня точка є чорною, то переходимо на крок 1.

3. Якщо нижня сусідня точка є чорною, то переходимо на крок 1.

4. Визначаємо можливість розгляду сусідніх точок зліва і справа. Для цього досліджується вертикальна лінія пікселів, в якій знаходиться поточна точка.

а) Якщо верхній піксель є білим, то переходимо на крок 5.

б) Знижуємося вниз по лінії, поки не зустрінемо білу точку, якщо такої не знайдено, переходимо на крок 5.

в) Продовжуємо рух вниз, поки не зустрінемо чорну точку, якщо такої не знайдено, то переходимо на крок 5.

г) Якщо нижній піксель є білим, то переходимо на крок 5.

д) Піднімаємося вгору по лінії, поки не зустрінемо білу точку, якщо такої не знайдено, то переходимо на крок 5.

е) Продовжуємо рух вгору, поки не зустрінемо чорну точку, якщо такої не знайдено, то переходимо на крок 5.

є) Якщо ліва сусідня точка є чорною, то переходимо на крок 1.

ж) Якщо права сусідня точка є чорною, то переходимо на крок 1.

5. Вихід з рекурсії.

Кроки 4.а-4.ж є доопрацюванням, і несуть в собі наступний сенс. Бічні сусідні точки розглядаються тільки в тому випадку, якщо лінія, в якій знаходиться поточна точка, складається більш ніж з двох відрізків різного кольору. У переважній більшості випадків інша структура лінії вказує на те, що чорні точки належать тіні, і розглядати їх немає сенсу.

На виході отримуємо обрамлення символу. Наступна підозріла точка шукається від правої межі раніше виділеного символу. У підсумку отримуємо набір обрамлень, не всі з яких виділяють символ. Дрібні обрамлення видаляються.

7. Результати машинних експериментів

Оцінювати методи з точки зору їх застосування в системах комп'ютерного зору можна за якістю приглушення фону, тіней та шуму і виділення об'єктів у вигляді зв'язних областей символів на номері автомобілю. Оскільки у поставленій задачі об'єктомрозпізнавання є номерний знак автомобіля, то основними кроками сегментації є:

– знаходження автомобіля;

– знаходження плити реєстраційного номеру автомобіля;

– виділення символів номерного знаку.

У даній задачі поняттям «об'єкт» є номерний знак, тому використаємо інформацію з ГОСТу [10] для формалізації і можливості вимагати найточніше виділення символів номерного знаку без виділення тіней, шумів або інших пошкоджень. Отжевимогу до завдання можна спростити: повинні бути виділені принаймні ключові частини об'єкта, необхідні для розпізнавання його окремих символів.

Для аналізу методів обрано зображення, типові для поставленої задачі. Номери були розділені на 3 групи: нормальні – номера, де нахил менше 30 градусів чітко видно цифри і букви; під кутом – номера, де кут нахилу більше 30 градусів і з сильним спотворенням; з дефектом – розфокусовані, змащені, брудні, з низьким дозволом.

Результати обробки зображень кожним з описаних методів наведені в таблиці.

Таблиця

Ефективність алгоритмів сегментації

Алгоритм	Тип розпізнаннязнаків автомобілів		
	Нормальний (кут <30)	Під нахилом (кут >30)	З дефектом
	Точність/помилка		
Граничний метод	0,85 / 0,2	0,69 / 0,38	0,78 / 0,27
Нарощування кордонів	0,89 / 0,42	0,73 / 0,4	0,75 / 0,43
Аналіз кольору	0,83 / 0,4	0,72 / 0,38	0,71 / 0,43
Центроідного зв'язування	0,85 / 0,33	0,66 / 0,45	0,74 / 0,48
Хвильовий метод	0,88 / 0,27	0,76 / 0,35	0,8 / 0,38
Метод вододілів	0,9 / 0,26	0,82 / 0,36	0,79 / 0,35
Злиття – розщеплення	0,83 / 0,38	0,79 / 0,41	0,78 / 0,41
Матриці збігів	0,85 / 0,29	0,70 / 0,57	0,77 / 0,37
Метод мозаїки Хаара	0,87 / 0,38	0,77 / 0,43	0,81 / 0,32
Модифікований алгоритм	0,93 / 0,24	0,87 / 0,32	0,86 / 0,27

Де із використанням модифікованого алгоритму отримано найкращі результати за запропонованими типами розпізнавання.

8. Висновки

Аналіз відомих методів сегментації з точки зору практичного застосування дозволив виявити їх основні характеристики і на основі розглянутих методів запропонувати комбінований алгоритм, найбільш ефективновиконує сегментацію з точки зору систем комп'ютерного зору. Алгоритм автоматично виділяє об'єкти або, для складних об'єктів, їх ключові частини.

Експериментальна перевірка показала, що комбінація методів дає хороші результати як для одного, так і для декількох об'єктів, що мають навіть незначні відмінності від фону по яскравості. У великих об'єктів, що складаються з декількох частин різної яскравості, вдалося виділити ключові частини. Таким чином, досягнута певна універсальність розробленого модифікованого алгоритму щодо різних типів зображень реєстраційного знаку автомобіля.

Список літератури

[5] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений: Пер. с англ. – М.: Техносфера,– 2005. – 1072 с.

[6] *Niblack W.* An introduction to digital image processing. – Prentice Hall, Englewood Cliffs.–1986. – 231 p

[7] *Navon E. et al.* Color image segmentation based on adaptive local thresholds // Image and Vision Computing. – 2012. – № 23. – P. 69–85.

[8] *Ciresan D., Meier U., Schmidhuber J.* Multi-column deep neural networks for image classification //Computer Vision and Pattern Recognition (CVPR). –IEEE Conference on. –2012. –P. 3642-3649.

[9] *Фихтенгольц Г.* Курс дифференциального и интегрального исчисления. – Т.3. – М.:Наука. –1969. – 2064 р.

[10] *LeCun Y. et al.* Learning algorithms for classification: A comparison on handwritten digit recognition.//Neural networks: the statistical mechanics perspective. – 1995. –Т. 261. –Р. 276.

[11] *Кей С., Марпл С.* Современные методы спектрального анализа // ТИИЭР.– 1981.– Т.69. –№ 11. –С. 5-51.

[12] *Rogers D.*Algorithmic bases of compute rgraphics.– M. Mir. –1989.– 288 p.

[13] *Helgason S.* Differential Geometry and Symmetric Spaces – M.: Mir 1964 – 534p.

[14] *Душник В., Макаренко Є., Савченко І.* Державний Стандарт України ДСТУ 4278:2006 Дорожній транспорт. Знаки номерні транспортних засобів. Загальні вимоги. Правила застосування.– 2007. – 34 р.

Надійшла до редколегії 25.04.2019

UDK 004.89

Chetverykov G.[1], Tereshchenko G.[2], Konarieva I.[3]

[1] Doctor of Technical Sciences, Professor of Software Engineering Department, Kharkiv National University of Radio Electronics, grirorij.chetverykov@nure.ua, ORCID iD: 0000-0001-5293-5842

[2] Graduate students of the Department of Software Engineering, Kharkiv National University of Radio Electronics, hlib.tereshchenko@nure.ua, ORCID iD: 0000-0001-8731-2135

[3] Graduate students of the University Complutense, Madrid (UCM), Spain, iulikona@ucm.es, ORCID iD: 0000-0001-9266-9877

DETECTION OF BLOOD CELLS

The structure of the medical image analysis system is considered. The algorithm of the blood cell recognition system is given. Formulated the main tasks to be solved during the morphological analysis of blood. The requirements for the algorithm in determining the leukocyte formula and the detection of blood corpuscles on a smear were determined. A model of color-brightness characteristics is proposed for describing typical images of a blood smear. The threshold values of the sizes of objects are determined when searching for cells. A histogram of the brightness of a typical field of view was investigated. A two-step algorithm for detecting blood cells is described, as well as an algorithm for constructing a dividing line on the plane of relative colors. The results of experiments on real preparations are given. The causes of detection errors are considered.

CELL COUNTING, DIGITAL MICROSCOPY, IMAGE SEGMENTATION.

Introduction

In the tasks of analyzing images obtained with a microscope, in the framework of cytological studies, it is often necessary to count the number of cells of a certain type. When examining blood products, an important task is to count the number of red blood cells, based on the indices of which it is possible to diagnose disorders in blood formation or damage to red blood cells due to various factors.

Among the blood cells distinguish erythrocytes, leukocytes, platelets. The erythrocyte is a nuclear-free cell of pink color, having the form of a somewhat flattened ellipsoid with a depression in the center with an average size of 8 microns. Leukocytes differ from erythrocytes in their larger size, amounting to 9–20 μm, in the presence of the nucleus and in the nature of their color, which can be violet, pinkish, or bright red. Platelets are nuclear-free formations of a round or oval shape with a size of 1-3 microns, with a red-violet center and a pinkish-blue periphery.

There are various methods for counting red blood cells in the blood, some of them use the already modeled base of images of blood cells and their characteristics [1–4], some use threshold decomposition [5] or segmentation using the method of a controlled watershed [6]. There are approaches in which the color characteristics of the image [7, 8] or texture characteristics [8] are used for segmentation. In [9], it was proposed to use the algorithm of the active contour model for the selection of cell contours.

The main problem with counting cells is that they can overlap with each other, as well as change their shape in a certain range. The presence of extraneous noise, foreign objects in the field of view of the microscope further complicates image analysis.

In this work, in order to reduce the effect of noise, it is proposed to use median filtering of images [10] with the subsequent extraction of cell contours by the Canny boundary detector [11]. To improve the recognition of borders, the image is additionally contrasted.

1. General scheme of the algorithm

When leukocytes are detected, two methods of segmentation of "primary objects" are used. At the beginning of the screening, information on leukocyte and erythrocyte colors in this preparation is assumed to be unknown. Therefore, a one-dimensional iterative method is used, based on the study of the peaks of the brightness histogram. After it has found several nuclei of leukocytes and accumulated information about the colors of erythrocytes, a straight line separating the colors of the nuclei and the color of erythrocytes is built on the plane of relative colors fR, fB. If the border between the colors of erythrocytes and nuclei is drawn "with a margin" (errors of the first and second kind are small), it begins to be used for segmentation. So, when selecting primary objects (possibly, nuclei), the transition from the one-dimensional method to the two-dimensional method takes place: instead of segmentation of the brightness histogram, the segmentation of the color plane is performed. The latter option is more stable: there is no danger of missing light nuclei, it is easier to work with poorly focused frames, it is not necessary to accurately determine the position of the peak of the background [4].

The algorithm for extracting fragments consists of four steps.

1. The selection of primary objects - possibly nuclei. In this case, two different segmentation methods are used — according to a histogram of brightness or on a color plane. The second method is preferable, but at the initial stage of sample accumulation the first one is used.

2. Verification of primary objects for compliance with an already accumulated sample of nuclei, which is possible if the number of accumulated objects more than 5. Primary objects are classified into nuclear fragments and artifacts.

3. Fragments of nuclei are combined with each other, and an attempt is made to build cytoplasm around them. As a result, we obtain a rectangle, inside which there is one leukocyte.

4. If a brightness histogram was used for the segmentation of the primary objects, and there were several peaks on it, possibly corresponding to the nuclei, then the number of found nuclei pixels is compared with the estimated (number of pixels in the peak). As a result of testing the hypothesis of which peak should be considered the peak of the nuclei, it can change and the algorithm can be started from the first step anew.

As a rule, there is no need for iterations; the algorithm consists of three steps: primary objects → nuclear fragments → leukocytes.

2. A segmentation algorithm based on a brightness histogram

The proposed leukocyte detection algorithm consists of two stages, which can be repeated several times for the same frame. At the first stage (based on the study of the histogram of the brightness of the frame and the history of the search) are selected threshold values for brightness G and share blue fB. At the second stage, the sets of pixels that meet these conditions (primary objects) are examined to determine whether they can be considered as leukocyte nuclei. If the total number of pixels in these fragments is significantly less than the previously estimated number, then the selection of threshold values is considered unsatisfactory and the algorithm is launched again, etc.

The first stage is the study of the histogram, the choice of threshold values. The first step is to localize the peak associated with the image background, which will later be the reference, both when calculating the optical density and when determining the relative colors for the remaining pixels. Absolute values have to be used only if the background peak is not localized. An extreme right peak with a rather small dispersion is chosen as the peak of the background pixels: the standard deviation is less than 10 digits [5].

Next, a list of maxima (peaks) is compiled that could correspond to leukocyte nuclei. To do this, their optical density must be sufficiently large (the empirically found boundary > 0.6) and the average blue fraction $fB = B/(B + G + R)$ for pixels at this maximum should exceed the same value for the background by 0.03 (empirically found border). If there are several such suspicious maxima, they are selected sequentially one after the other (in this case the second stage of the algorithm is called), starting with the brightest one. Held threshold segmentation by brightness and relative proportion of blue $G < G\max$; $fB = B/(B + G + R) < fB\max$.

The obtained primary objects are compared with already existing nuclei. If they are not qualified as nuclei, and this is possible in the presence of optically dense and bluish red blood cells, stains of paint, large platelets, then the next maximum will be selected. If the leukocyte peak is not distinguished at all, then threshold values are used, which are no longer based on the current histogram, but on the prehistory of the search, and if there is none, then on a priori values.

The second stage is the study of the obtained fragments. This part is independent of the method by which the primary objects were obtained. The algorithm for checking selected objects consists of three cycles. In the first cycle, too large ($A > 2000$ μm^2) and too small ($A < 11$ μm^2) objects are discarded.

Further, the optical density and color characteristics are measured. If there is a prehistory of the search, then by the criterion of "three sigma" excessively light objects with a low optical density are discarded and the procedure for checking colors is called. The remaining objects are placed in the class of conditional nuclei of leukocytes.

In the second cycle, the completion of the cytoplasm around the nucleus takes place on the remaining objects. To the cytoplasm include a coherent set of nearby pixels, which with a high probability (more than 0.95) are not pixels of red blood cells or background. The constructed set is rejected if it is too large (more than 2000 μm^2) or the form factor of its external border (square of perimeter / area) exceeds a sufficiently large value (more than 50). It is often enough, with a close diligence of the white blood cells of similarly colored erythrocytes, the cytoplasm cannot be completed in such a simple way. Then the core or its fragment is placed inside a rectangle with added frames of 15 μm.

After completing the cytoplasm, agglutination of nuclear fragments is performed. This is necessary since the nuclei of neutrophils are detected in the form of several fragments. In this case, the separation of cells that are close to each other. This is possible if the cells lie in islands that are not connected with each other, surrounded by background.

After the second cycle, the objects obtained are considered as separate leukocytes. In the third cycle, the sizes of these objects are checked again, and too large objects are discarded.

If the segmentation used a brightness histogram, then at the end a check is made for the consistency of the assumptions and the results obtained. For this, the found number of pixels of leukocyte nuclei is compared with the number of pixels at the peak of the histogram, which was assumed to be the corresponding nuclei. If the differences are significant (exceed> 50%) and to the left of the peak of the supposed leukocytes there was another one, then the above algorithm runs again. In fig. 5 shows the scheme of the detection of secondary objects.

Algorithm for checking the primary object for belonging to the leukocyte nucleus group. Leukocyte nuclei do not constitute a homogeneous group. Therefore, it is not necessary to relate strictly to checking for the belonging of a new object to a two-dimensional, normal distribution, even for average values of relative colors. In addition, in the smear screening process, it is desirable to use the verification algorithm as early as possible when the number of objects accumulated is small. Therefore, the algorithm proposed below is heuristic. It is based on the following provisions.

1. Each new object is compared with two groups: a group of nuclei and a group of red blood cells.

2. When determining the probability of a new point belonging to an existing group, the probability is calculated twice. In the first case, the probability P_1 is calculated before the point is added to the existing statistics, and in the second case — P_2, after such an addition. Obviously, $P_2 > P_1$. Such a calculation of two probabilities at once is necessary if decisions are made on the basis of small samples, and the number of objects accumulated without checking is usually just a little — about 5. If the number of objects in a group exceeds several dozen, the difference between P_1 and P_2 almost disappears.

3. If the object under study is more "blue-red" than already accumulated nuclei, then it is accepted in any case. In other words, the possibility of emissions in this "blue-red" side is not taken into account. Conversely, if the object is more "green" than red blood cells, then it is rejected in any case.

4. If the probability P_2 calculated relative to the group of erythrocytes is greater than the corresponding probability calculated relative to the nuclei, then the object is rejected.

5. If a decision is not taken on the basis of the preceding paragraphs, then it is made taking into account three probabilities. The probabilities P_1 and P_2 estimate the deviation of the average value, and the probability P_3 — the deviation of the area of the ellipse of scattering from the characteristic for a group of nuclei. The object is rejected if $P_2 < 0.01$ or $P_3 < 0.01$.

For a given primary object, the corresponding probabilities $P_1 = 0.53$, $P_2 = 0.65$, $P_3 = 0.45$ are high, so it will be correctly qualified as a core and added to the statistics of nuclei.

3. Erythrocyte isolation and counting method

The one-dimensional median filter is a "sliding window" with a length of N samples, in which the central element is replaced by the median (ie, the middle element of the sequence, ordered in ascending order of the signal values in the "window"). Thus, the operation of the median filtering of a K-dimensional sequence of signal values $s(k) = s(x_k)$, $k = 1, \ldots K$ characterized by the ratio

$$\underset{1\le k\le N}{\text{med}}\{S_k\} = \begin{cases} 0{,}5(s_n + s_{n+1}), N = 2n \\ s_n, N = 2n-1, \end{cases}$$

where the fixed value $n = 1, 2, \ldots$ determines the filter aperture.

The next stage consists in the selection of boundaries, after which the method of connected components with a connectivity criterion along eight neighbors is separated into individual contours. For each connected region, the area of a convex polygon describing the contour is calculated. For given thresholds, sections that are too large or too small are cut off. Thresholds are selected based on the estimated real cell area.

Then, points are selected from each individual contour at equal intervals along the contour length. The points are connected in pairs with each other, and a perpendicular is drawn through the middle of the obtained segment. It can be described by the equation

$$y = -\frac{x_2 - x_1}{y_2 - y_1}x + \frac{y_2^2 - y_1^2 + x_2^2 - x_1^2}{2(y_2 - y_1)},$$

where (x_1, y_1) – coordinates of the first point;
(x_2, y_2) – coordinates of the second point.

The location of the intersection point of adjacent perpendiculars is preserved. The point of intersection of two perpendiculars is calculated as

$$x = \frac{b_2 - b_1}{a_1 - a_2},$$
$$y = a_1 x + b_1 = a_2 x + b_2,$$

where $a = \frac{x_2 - x_1}{y_2 - y_1}$ – coefficient of inclination of the perpendicular; $b = \frac{y_2^2 - y_1^2 + x_2^2 - x_1^2}{2(y_2 - y_1)}$ – the coefficient of perpendicular displacement.

The operation is performed for all pairs of perpendiculars for the various steps of taking points. As a result, a cloud of points is formed, which are located more densely in areas that are the centers of the radius of the contour curves.

Cells stuck together with one another or superimposed one upon the other are rather difficult to segmentation by methods based on analyzing the size or shape of the areas inside the contours. Gaps in the contours complicate the contour segmentation of cells. The proposed method is a single image for all areas of

the contours of their centers, thus allowing to solve the problems indicated above.

The picture, composed of the obtained points of intersection of perpendiculars, undergoes morpho logical processing, as a result of which only dense and rather large clusters of points remain. These clusters correspond to the putative cell centers. Using the method of connected components, clusters are counted, which should correspond to the number of cells in the image.

4. Experimental results

In an experimental study of the proposed method, an image of a blood sample was taken using a microscope (Fig. 1).

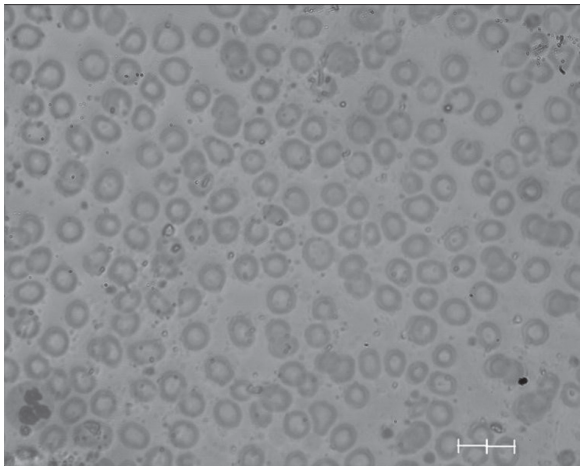

Fig. 1. Test picture

First of all, the image was converted from color to black and white. The window size in the median filtering was selected based on the average cell size and was 16x16 pixels, which corresponds to 20% of the cell diameter. The result of the median filtering of the test image is shown in Fig. 2.

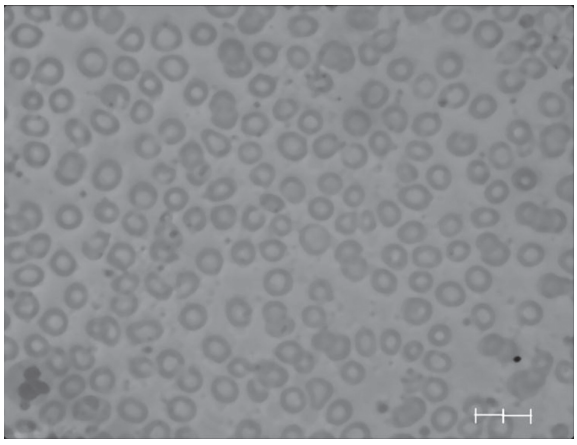

Fig. 2. The result of median filtering

After linear image contrasting, a Canny boundary detector was applied (Fig. 3). The Canny algorithm first smoothes the image to remove noise. Borders are then selected where the gradient of the image acquires the maximum value, with only local maxima marked as borders. The next step is to determine the potential boundaries of the double threshold filtering. Total boundaries are determined by suppressing all edges that are not associated with specific boundaries.

As thresholds for removing too large or too small contours in the image, two values were chosen: 0.05S as the lower threshold and 4S as the upper threshold, where S is the approximate area of the cell image calculated on the basis of its diameter.

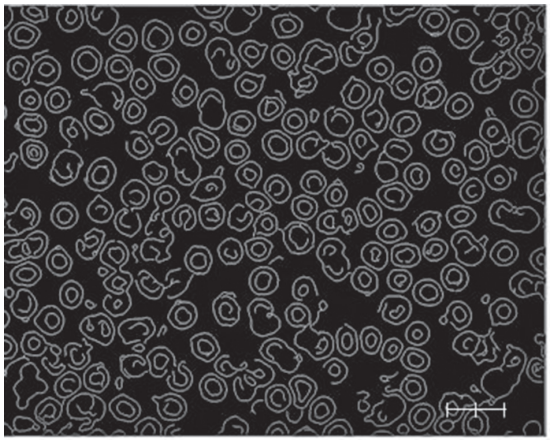

Fig. 3. Image of edges with large and small contours removed

As a result of the construction of perpendiculars (Fig. 4), for the segments between points taken with 3–40 pixel intervals, a cloud of intersection points was obtained for each contour (Fig. 5).

Fig. 4. Perpendiculars; spacing between points = 20

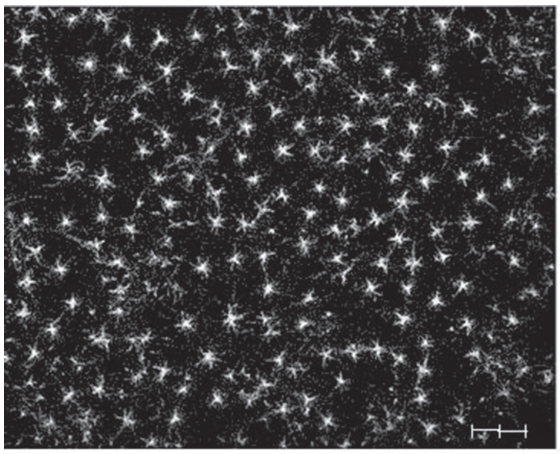

Fig. 5. Point clouds

The number of connected components was 220. The real number of cells in the image was obtained equal to 209. The number of false positives was 23, the number of unrecognized cells was 12. The method was tested on four different images containing blood cells. The average probability of correctly counting the number of erythrocyte cells was 86%. Compared with the methods using threshold decomposition [5] or segmentation by the method of controlled watershed [6], the proposed method gave the best results. However, in comparison with the methods proposed in [1–4, 9], the probability of correctly counting the number of cells turned out to be less, since the border detector incorrectly selected edges on test images due to strong noise and the presence of fuzzy boundaries. For such images, additional preprocessing methods are needed to improve the efficiency of edge extraction.

Plots with an area of less than 5 pixels were removed from the image of a cloud of points, and then the operation of closing a binary image with a mask of 8x8 pixels was performed. The result of the morphological processing of the image of intersection points superimposed on the original image is shown in Fig. 6.

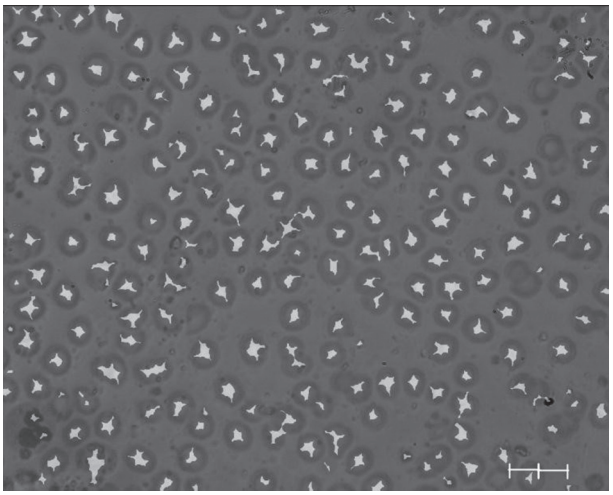

Fig. 6. Morphologically processed image of intersection points superimposed on the original image

Conclusions

Due to the fact that a border detector is used to identify cells, the segmentation results do not depend on the color of the cells, their texture and internal structure. The method with rather high accuracy segments the cells stuck together with each other or superimposed on each other. In conditions of noisy source image method showed good results.

The proposed algorithm allowed segmentation and counting of blood cells with an accuracy of 86%. The number of false cell detections is on average higher than that of other methods, which can be explained by the presence of a large amount of noise on the test images, as well as by fuzzy cell boundaries. A higher probability of correctly counting the number of cells can be achieved if, in parallel with the proposed method, you use others, specifying the result of the segmentation of one method by the results of another, as well as using other algorithms for preliminary processing of the original image. In the future we plan to develop an algorithm for image preprocessing to increase the efficiency of the proposed method, as well as a combination of the proposed method with others.

Reference*s*

[1] *Dahmen J., Hektor J., Perrey R., Ney H.* Automatic Classification of Red Blood Cells Using Gaussian Mix-ture Densities // Proc. Bildverarbeitung für die Medizin. – 2000. – P. 331–335.

[2] *Costrarido L.* Medical Image Analysis Methods: Evaluation Strategies for Medical-image Analysis. – Taylor & Francis, United Stated of America, 2005. – P. 433–471.

[3] *Kumar B.R., Joseph D.K., Teager T.V.S.* Energy Based Blood Cell Segmentation // 14th International Conference on Digital Signal Processing. – DSP, 2002. – 1–3 July. – Santorini, Greece. – V. 2. – P. 619–622.

[4] *Bamford P.* Empirical Comparison of Cell Segmentation Algorithms Using an Annotated Dataset // Proc. IEEE International Conference on Image Processing. – 2003. – V. 2. – P. 1073–1077.

[5] *Mukherjee D.P., Ray N., Acton S.T.* Level Set Analysis for Leukocyte Detection and Tracking // IP. – 2004. – V. 13. – № 4. – P. 562–572.

[6] *Park J., Keller J.M.* Snakes on the Watershed // IEEE Transactions on Pattern Analysis and Machine Intelligence. – PAMI, 2001. – V. 23. – № 10. – P. 1201–1205.

[7] *Sinha N., Ramakrishnan A.G.* Blood Cell Segmentation Using EM Algorithm // Proc. Third Indian Confer-ence on Computer Vision. Graphics Image Processing (ICVGIP), 2002. – Ahmadabad, India, 2002, Decem-ber 16–18. – P. 376–382.

[8] *Kumar R.S., Verma A., Singh J.* Color Image Segmentation and Multi-Level Thresholding by Maximization of Conditional Entrophy // International Journal of Signal Processing. – 2006. – V. 3. – № 1. – P. 121–125.

[9] *Mcinerney T., Terzopoulos D.* Deformable models in medical image analysis: A survey // Med Image Anal. – 1996. – P. 91–108.

[10] *Cormen T.H., Leisorson C.E., Rivest R.L.* Introduction to Algorithms. – MIT Press, 1990. – P. 185–191.

[11] *Canny J.F.* A computational approach to edge detection // IEEE Trans. Pattern Analysis and Machine Intelli-gence. – 1986. – P. 679–698.

The article was delivered to your editory stuff on the 24.05.2019

UDK 519.62

Shafagat Mahmudova
Institute of Information Technology of ANAS,
Software engineering, Azerbaijan, shafagat_57@mail.ru

BIOMIMETICS: NOTIONS, PROBLEMS AND TECHNOLOGIES

Biomimetics is an imitation model of systems and elements in the nature to solve complex human problems. Living organisms have well-adapted structures and materials for natural selection and have evolved over many years. The study of biomimetics technologies and their application in different areas can play an important role in the perfect economic development. This article touches upon various aspects of biomimetics and analyzes its technologies. The further development of these technologies in the future is intended.

BIOMIMETICS, TECHNOLOGY, BIOMIMETIC DESIGN, FLIGHT

Introduction

In the mid-20th century, a new scientific area began to form and was called "pattern recognition". The main purpose of this scientific area was to determine the class which the recognized object belongs to. Objects close to each other for their features were classified [1]. As the "recognition of images" developed, other related areas began to emerge and develop. One of them is biomimetics.

Biomimetics driven from Greek word (bios-life, mimetis -imitation) is an imitation model of systems and elements in the nature to solve complex human problems [2]. Living organisms have emerged and developed from a well-adapted structure and materials for natural selection. Biomimetics has developed new technologies based on biological solutions of macro and nano scale. For example, in the early stages of the development of biomimetics, the structure of birds was well-studied for man to fly.

Biomimetics studies the biological systems and processes to apply the knowledge obtained from nature to solve technical issues. Biomimetics enables people to create original technical systems based on the ideas found and obtained in nature. Biomimetics proves that people's inventions exist in living things, for example, hook and sticky fabric are invented based on bird's feathers.

Biomimetics is closely related to biology, physics, chemistry, cybernetics, engineering sciences, electronics, and so forth.

Biomimetics studies the work of human brain and explores the mechanism of memory. It intensively explores the sense organs of animals and their responses to the environment (figure 1).

Fig. 1. Human brain

The main fields of Biomimetics studies primarily cover the following problems.

- Study of neural networks through human nervous system;
- Study of sense organs and perception system of living beings for the development of new sensors and detection systems;
- Study of orientation, location and navigation principles of various animals for their use in technical fields;
- Study of morphological, physiological and biochemical features of living organisms for the development of new technical and scientific ideas.

Living organisms have emerged and developed from a well-adapted structure and materials for natural selection. Biomimetics has developed new technologies based on biological solutions of macro and nano scale. For example, in the early stages of the development of biomimetry, the structure of birds was well-studied for man to fly.

One of the most famous examples of biomimetics is a human flight (figure 2).

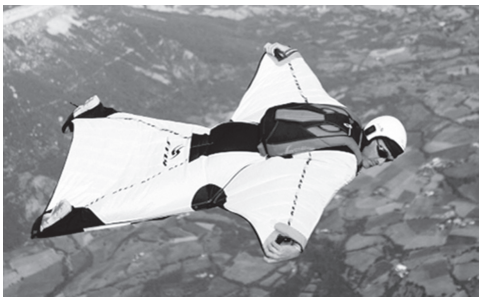

Fig. 2. Flying man

One of the most famous examples of biomimetics is a human flight. Leonardo da Vinci is known as the main challenger to design to carry out the first real research of birds' and human flights in the 1480s. His famous original design, known as Ornithopter, had never been created, but was instructed to show the human potential to fly. His famous original design, known as Ornithopter, had never been created, but was instructed to show the human potential to fly.

Leonardo da Vinci (1452-1519) repeatedly observed birds' flight, describing them in his works (Figure 2), however he did not deal with that area. Although he was not able to create an "airplane," he was an observer interested in anatomy and bird's flight, and left numerous notes and sketches, as well as the sketches of "flying machines" [3] (figure 3).

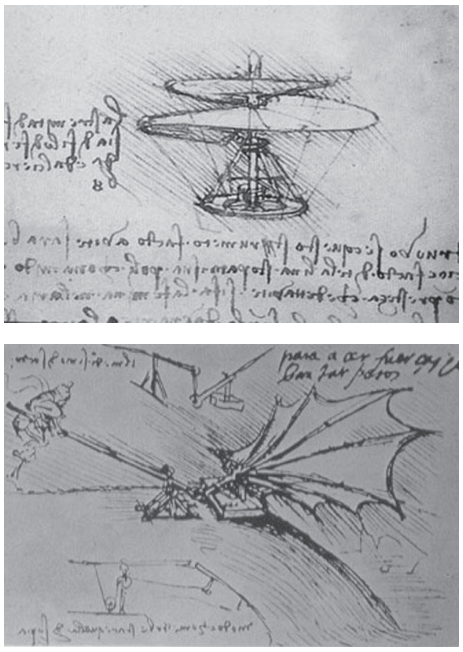

Fig. 3. Sketch of three-dimensional apparatus by Leonardo Da Vinci

In 1903, American engineers the Wright Brothers, who were able to take off the first heavy aircraft, were inspired by the pigeons in flight [4].

In 1950, American biophysics and polymath **Francis Otto Schmitt** developed the concept of biomimetics. He studied squid nerves in his doctoral thesis and attempted to develop a compatible biological nerve proliferation device [5].

1. Biomimetic design

Due to the consistent development of nature over the millennium, everything has its own solutions, and consequently, their use in solving modern human problems is being studied.

Despite the incredible inventive and engineering skills the humanity showed in the past millennium, as Pyramids, Skyscrapers, Supersonic Flight, people are constantly looking for the ways to develop new projects. Given the evolution in nature and millions of years of experiments and errors, taking advantage of the opportunities of nature is logical.

The areas of the biomimetics are shown in Figure 4.

Biological biomimetics
Theoretical biomimetics
Technical biomimetics

Fig. 4. The areas of the biomimetics

Biological biomimetics studies the processes related to biological systems.

Theoretical biomimetics builds the mathematical models of processes.

Technical biomimetics applies the theoretical biomimetic models to solve engineering problems.

2. Nature inspired biomimetics technology

One of the first examples of biomimetic materials is the invention of a widespread "sticky fabric". In recent years, the development of nanotechnology has stimulated the development of biomimetics.

Researchers have used different methods to imitate nature at the nanometric level. The goal was to create unique materials inspired by the natural samples. For example, a small lizard called gecko can adhere practically on any surface. To imitate the features of gecko, it was necessary to first understand the mechanism of the work of its pads. It was studied at the Nanotechnology Center in Manchester.

Some nature inspired technologies in biomimetics are given below (Figure 5).

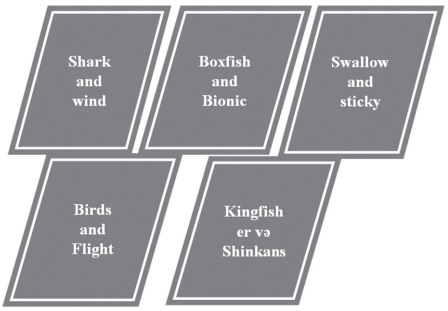

Fig.5. Some nature inspired technologies in biomimetics

Shark and wind turbines - Although the shark's weight is about 36 tones, it is one of the weakest swimmers in the sea. Bio-mechanic Frank Fish related the aerodynamic abilities of the bumpy protrusions on the front of its fins, called tubercles [6] (figure 6).

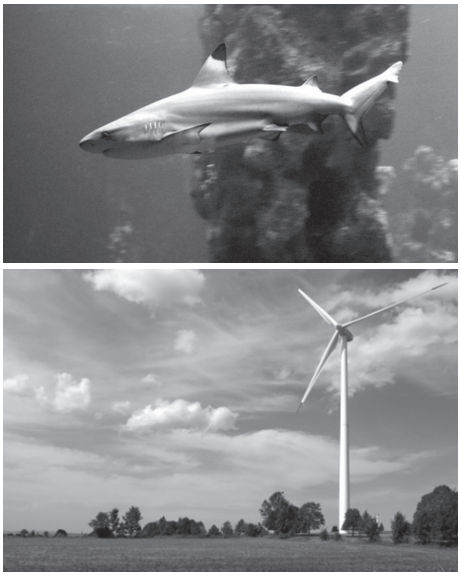

Fig. 6. Wind turbines and sharks

Boxfish and Bionic Car - Despite the huge appearance of the cube shaped boxfish, its resistance coefficient is approximately 0.06. For comparison, the coefficient of swimming penguins equals to 0.19. In 2005, Mercedes Benz developed the Bionic Car, inspired by the structure and gravity of cube shaped fish. It reduces the car's resistance, it has great rigidity and low weight, and uses less fuel than conventional cars (figure 7).

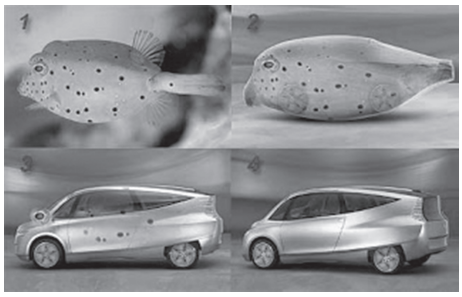

Fig. 7. Boxfish and Bionic Car

Swallow and sticky fabric - George de Mestral was inspired to invent the sticky fabric noticing how easy it was for swallow to stick to the dog's hair. Under the microscope, he realized the simplest design of small hooks at the end of the swallow's spines (figure 8).

Fig. 8. George de Mestral stick and swallow

Birds and Flight - One of the most popular examples of Biomimetics is a human flight. Several designers and engineers have been inspired by this concept, for example, German engineer Otto Lilienthal realized flights on over 2500 planes, nevertheless by 1903, the Wright brothers flew the first powered, heavier-than-air machine in a controlled and sustainable flight. This technology has led to the development of the 20th century and air industry technologies [6]:

Lotus inspired hydrophobia;
Water collecting beetles;
Biomimetic architecture;
Birds-safe glass;
Shark skin coat;
And so forth.

Kingfisher and Shinkansen. Japanese trains are famous for its incredible speed and efficiency. However, fast bullet trains driving out of tunnel at the speed of 300 km/h resulted in a strong sonic problem. The unfavorable outcome of sonic pollution caused by the change in air pressure was very alarming to the local population and attracted engineers to address this problem. Inspired by a kingfisher they solved this problem. Kingfisher are masters in traveling at a very high altitude both in air and water. Like Kingfisher, Shinkansen, a fast passenger train, is equipped with a long beak-shaped nose. This significantly

reduces the train noise, at the same time uses less than 15% of electricity and travels 10% faster than before (figure 9).

Fig. 9. Japanese fast train (Shinkansen) and Kingfisher

Bionic Bird - Drone - Bird. The biomimetic technology is used for drone to take off and fly faster without a pilot. Bionic Bird flies at a speed of 20 km/h and is controlled by a smartphone. Flights can be performed both indoors and outdoors within 12 minutes (figure 10).

Fig. 10. Bionic Bird drone - bird

The study of biomimetic robots and animal behavior is interrelated and inseparable. Through long-term evolutionary processes, animals have achieved natural advantages in movement, cognition, processing and control. Inspired by their development, the biomimetic robots, unlike others, have biological features that provide more powerful motion and cognitive ability and more sensitive control process. Simultaneously, the development of biomimetic technology and the mutual features of biomimetic robots also encourage studying animal behavior. This is a common representation of the relationship between biomimetic robots and animals' behavior. On the one hand, the role of the imitation of animals' behavior for the promotion of the development of biomimetic robots is illustrated in three aspects [7] (figure 11).

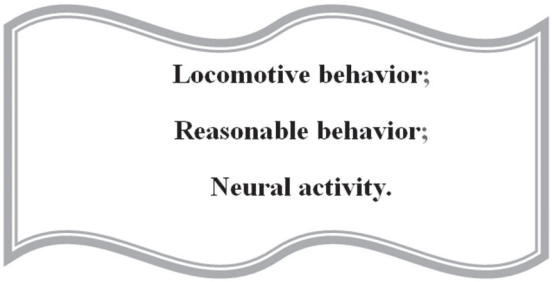

Fig. 11. Three aspects of the role of imitation of animals' behavior

On the other hand, the positive role of biomimetic robots in the study of animals' behavior is described in terms of behavioral responses, group behavioral mechanisms and cognitive-neurological activity of animals. In addition, the future development of biomimetic robots and the study of animals' behavior are discussed.

3. Biomimetics and software

[8] focuses on a preliminary assessment of the biomimetic diagram of a new form to simulate the function of human ear-hearing system (ESQ). ESQ consists of three parts: a pars-tensa and pars-flaccida, and its dynamic behavior, which, obviously, differs from other ordinary thin membranes. The developed membrane has a curved conical shape with an apex pointing medially, and with an initially bucked form. A small body also closely adheres to the medial surface of the membrane at its center. Additionally, TM is associated with a ring connection (mouse). Ultimately, the TM does not move as a straight flat or delicate diaphragm. In this study, bilinear nonlinear elliptical and conical shape, similar to the actual TM of the human hearing system, provided good vibration properties. When the Sound Pressure Level (SPL) is high, the adaptive diaphragm structure developed using the 3D printing technology, which can lead to 3D response frequency, may perform bilinear non-linearity [8].

When building the biomimetic neuronal structures, the topological features of biological neural networks are imitated on various scales. The optical technology platform is used for the reproduction of topological features of biological neural networks [9] (figure 12).

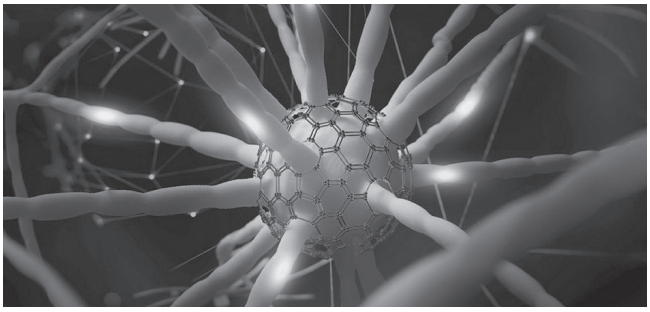

Fig. 12. Biomimetic neural structure

Autonomous submarines are completely or partly dependent on human decisions. The submarines should be equipped with special software to be independent. The main purpose of the program is to prevent the collision of submarines. In addition, the application should control various devices of the camera, such as the performance and interruption of cameras and so forth. The program is installed to the submarines' panel by the operator. Its task is to identify the submarines, disable the work, send emergency commands and remotely control the parameters. The goal of the software is to support the development and testing of other software components. In this regard, specific software is required to visualize all major facilities, the environment under the water and the submarines [10].

The project is proposed for the development of an interactive program that supports the test sequence and sensitivity of the program, which is based on the development models of the biomimetic system. Here, the prototypes are applied and evaluated. The system software specifies it and performs tests that detect errors [11].

Horseshoe bats (Rhinolophidae) differ from each other for the incoming and outgoing sound waves. In some studies, Horseshoe bats techniques are used to improve the coding of peripheral dynamic sensor data. The software architecture is based on MATLAB, which is a part of the flexible interface for experimental design and data analysis, while the server part is based on Python, LabVIEW [12].

Conclusion

Biomimetics introduces the principles and strategies interpreted from biological systems to engineering and technological designs [13].

Biomimetics is a field of research of strategy transfer from biology to technology, and has led to the emergence of important concepts in recent decades. The development of these technologies was illustrated by biomimetic processes consisting of several stages [14]. Some studies explored general descriptions and classifications of more than 40 technologies with quality criteria. The classification showed that certain stages of the process and their problems were finely solved by means of tools, while others were not solved. It can be concluded that the level of technology can be further enhanced, and the future theoretical and practical analysis is intended. These results can promote the widespread use of biomimetics [15].

References

[1] *Горелик А.Л., Скрипкин В.А.* Методы распознавания Москва: Высшая школа. – 2004. – С. 261.

[2] Vincent Julian F. V. Biomimetics: its practice and theory // Journal of the Royal Society Interface. – 2006. – № 3(9). – P. 471–482.

[3] *Francesca R.* Leonardo Da Vinci. The Oliver Press. – 2008. – P. 56.

[4] *Wright B.* The Invention of the Aerial Age. Washington: National Geographic Society. – 2003. – P. 257.

[5] *Vincent J. F.V., Bogatyreva O. A., Bogatyrev N. R., Bowyer A., Pahl A.* Biomimetics: its practice and theory // Journal of the Royal Society Interface. – 2006. –№ 3(9). – P. 471–482.

[6] Gertie G. Biomimetic design: 10 examples of nature inspiring technology, https://www.sciencefocus.com/future-technology/biomimetic-design-10-examples-of-nature-inspiring-technology/

[7] *Gao Z., Shi Q., Fukuda T., Li C., Huang Q.* An overview of biomimetic robots with animal behaviors // Neurocomputing. – 2019. – № (332). – P. 339-350.

[8] *Yoon J.Y., Kim G.W.* Harnessing the bilinear nonlinearity of a 3D printed biomimetic diaphragm for acoustic sensor applications // Mechanical systems and signal processing. –2018. – № (116). – P. 710-724.

[9] *Yu H., Zhang Q., Gu M.* Three-dimensional direct laser writing of biomimetic neuron structures // Optics express. – 2018. – № 26 (24). – P. 32111-32117.

[10] *Tomasz P., Piotr S.* Software architecture of biomimetic underwater vehicle, Conference: SPIE Defense + Security, Baltimore, Maryland, United States, May –2016. – № 9831.

[11] *Feldt R.* Biomimetic software engineering techniques for dependability. – 2002. 206 – P.

[12] *Rolf M.* System integration for a biomimetic dynamic sonar head // The Journal of the Acoustical Society of America. –2018. – № 143 (3). – P.1727-1727.

[13] *Robert F.* An Interactive Software Development Workbench based on Biomimetic Algorithms, Vasa Bokbinderi: Goteborg, Sweden. – 2002. – P. 42.

[14] *Fayemi P. E., Wanieck K., Zollfrank C., Maranzana N., Aoussat A.* Biomimetics: process, tools and practice // BIOINSPIRATION & BIOMIMETICS. –2017. – № 12 (1). –P. 53-67.

[15] *Wanieck K., Fayemi P.E., Maranzana N., Zollfrank C., Jacobs S.* Biomimetics and its tools // Bioinspired biomimetic and nanobiomaterials. –2017. –№ 6(2). – P. 53-66.

The article was delivered to your editory stuff on the 05.06.2019

УДК 004(4’22+054)

Чайников С.И.[1], Солодовников А.С.[2]

[1] К.т.н., доцент, каф. системотехники,
ХНУРЭ, г. Харьков, Украина, serhii.chainikov@nure.ua

[2] К.т.н., доцент, каф. медицинской и биологической физики и медицинской информатики,
Харьковский национальный медицинский университет, г. Харьков, Украина, andrew.sldv@gmail.com

МЕТОДЫ СТРУКТУРНОГО СИНТЕЗА И АВТОМАТИЗИРОВАННОГО КОНФИГУРИРОВАНИЯ ПРОГРАММНОЙ АРХИТЕКТУРЫ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Проведен анализ методов структурного синтеза и кастомизации программного обеспечения информационной системы. Указывается, что современные методы не удовлетворяют требованиям к эффективной адаптации программного обеспечения под изменяющиеся во времени требования конечного пользователя. Показано, что при разработке формальных графовых моделей программной архитектуры с использованием существующих методов эволюционные изменения требований конечного пользователя обычно не рассматриваются, что приводит к трудностям при решении задачи кастомизации. Несоответствие между возможностями существующих технологий проектирования и практической необходимостью адаптации программного обеспечения приводит к возникновению проблемы разработки эффективных формальных подходов к кастомизации. Анализ методов подтверждает актуальность решения задачи разработки эффективных формальных подходов к кастомизации программного обеспечения с учетом специфики конкретного предприятия и конкретного рабочего места путем использования ярусно-параллельных графовых моделей программной архитектуры.

ГРАФОВАЯ МОДЕЛЬ, АРХИТЕКТУРА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ, СТРУКТУРНЫЙ СИНТЕЗ, ЯРУСНО-ПАРАЛЛЕЛЬНАЯ ФОРМА, BACKTRACKING, КОНТРОЛЬНАЯ ТОЧКА, AGILE, GRID, TDD, СЕРВИС-ОРИЕНТИРОВАННЫЙ ПОДХОД

Чайніков С.І., Солодовніков А.С. Методи структурного синтеза й автоматизованого конфігурування програмної архітектури інформаційної системи. Проведено аналіз методів структурного синтезу і кастомізації програмного забезпечення інформаційної системи. Вказується, що сучасні методи не задовольняють вимогам до ефективної адаптації програмного забезпечення під змінні в часі вимоги кінцевого користувача. Показано, що при розробці формальних графових моделей програмної архітектури з використанням існуючих методів еволюційні зміни вимог кінцевого користувача зазвичай не розглядаються, що призводить до труднощів при вирішенні задач кастомізації. Невідповідність між можливостями існуючих технологій проектування і практичною необхідністю адаптації програмного забезпечення призводить до виникнення проблеми розробки ефективних формальних підходів до кастомізації. Аналіз методів підтверджує актуальність вирішення задачі розробки ефективних формальних підходів до кастомізації програмного забезпечення з урахуванням специфіки конкретного підприємства і конкретного робочого місця шляхом використання ярусно-паралельних графових моделей програмної архітектури.

ГРАФОВА МОДЕЛЬ, АРХІТЕКТУРА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ, СТРУКТУРНИЙ СИНТЕЗ, ЯРУСНО-ПАРАЛЕЛЬНА ФОРМА, BACKTRACKING, КОНТРОЛЬНА ТОЧКА, AGILE, GRID, TDD, СЕРВІС-ОРІЄНТОВАНИЙ ПІДХІД

S.I. Chainikov, A.S. Solodovnikov. Methods of structural synthesis and automated configuration of the program architecture of information system. Authors represent analysis of the methods of structural synthesis and customization of the information system software. It is indicated that modern methods do not satisfy the requirements for effective adaptation of software for time-varying end-user requirements. It is shown that when developing formal graph models of software architecture using existing methods, evolutionary changes in end-user requirements are usually not considered, which leads to difficulties in solving the problem of customization. The mismatch between the capabilities of existing design technologies and the practical needs to adapt software leads to the problem of developing effective formal approaches to customization. The analysis of the methods confirms the relevance of solving the problem of developing effective formal approaches to customization of software, taking into account the specifics of a particular enterprise and a particular workplace by using tier-parallel graph models of software architecture.

GRAPH MODEL, SOFTWARE ARCHITECTURE, STRUCTURE SYNTHESIS, MULTILEVEL STRUCTURE, BACKTRACKING, CONTROL POINT, AGILE, GRID, TDD, SERVICE BASED APPROACH

Введение

Многие области человеческой деятельности в связи с тенденцией к усложнению за последнее время требуют поддержки информационных технологий (ИТ) с целью оптимизации и автоматизации труда. В рамках конкурентной структуры рынка программного обеспечения (ПО) немаловажную роль в процессах проектирования и разработки играет не только качество, надежность, информационная безопасность, но и скорость формирования готового программного продукта. Так же заказчики предъявляют высокие

требования к скорости выполнения функций самим ПО. Особенно это характерно для информационных систем (ИС), характеризующихся структурной, функциональной, информационной сложностью, сложной динамикой поведения. Проектирование и разработка подобного рода ИС требует значительных трудовых, временных затрат. В связи с увеличением сложности ИС, согласно стандарту ISO/IEC 12207-2008, усложняются процессы проектирования программной архитектуры, менеджмента конфигурации, менеджмента повторного применения программ и сопровождения ПО ИС. Поэтому на современном этапе активно развиваются автоматизированные методы синтеза программмной архитектуры, её компоновки и конфигурирования. Возникает задача кастомизации ПО, выражающаяся в адаптации программной архитектуры и функционала программного продукта к требованиям конечного пользователя.

1. Методы структурного синтеза и автоматизированного конфигурирования программной архитектуры

Один из распространенных подходов к формированию архитектуры ПО – применение сборочного подхода или использование технологии композитных приложений, которая показывает свою эффективность при использовании готовых программных компонентов сторонних разработчиков [1]. Тем не менее, в некоторых случаях существуют проблемы, связанные с принципами стыковки программных компонентов, обеспечением совместимости их функций. Решение этих проблем происходит: а) в рамках использования ограниченного числа компонентов, известных пользователю; б) в случае принадлежности компонентов одному разработчику, когда их технологическая и методологическая совместимость изначально обеспечена [2]. В этих случаях логично использование проблемной ориентации целевого ПО [3], применяемого для задач моделирования, контроля, анализа, автоматизированного управления, которое ограничено узкой предметной областью (ПрО). Синтез таких программных систем или комплексов программных средств (ПС) возможен на базе существующих программных компонент, сервисов и программных модулей.

Сборка ПО осуществляется в ручном, автоматическом или полуавтоматическом (автоматизированном) режимах [4].

Автоматический режим, хотя и позволяет снизить время разработки программы, все же обладает рядом недостатков в сравнении с автоматизированными методами, а именно – сложность генерации ПО для распределенных или параллельных вычислительных систем (ВС) и при наличии недетерминированных, трудно формализуемых ПрО.

В общем смысле выделяют два подхода к синтезу программной архитектуры ИС, основываясь на степени формализации исходной модели ПрО: 1) логический и 2) структурный синтез [4]. Логический синтез программной архитектуры базируется на математическом исчислении, представляющем закономерности функционирования объектов и их взаимосвязи, и трудно применим для нетривиальных ПрО. Структурный синтез обладает большими преимуществами в таких случаях и позволяет использовать наглядное представление структуры.

Логичным совмещением формальных и неформальных средств описания архитектуры ИС является архитектурный фреймворк [5], вмещающий в себя конвенции, принципы и методы описания архитектуры.

Для выявления особенностей и подходов к автоматизации процесса синтеза программ рассмотрим существующие методы.

Для процесса проектирования программной архитектуры ИС за основу может быть взята одна из существующих технологий проектирования (SADT, IDEF, SSADM, Meris) [6], основываясь на критерии схожести интерпретации этапов жизненного цикла (ЖЦ) ИС. Однако для обеспечения автоматизации процесса синтеза архитектуры в качестве исходной информации используется формализованное описание ПрО, формализованное представление программной архитектуры, а также требования конечного пользователя к ПО.

Такой формальный подход реализован на базе совокупности формальных документов, адекватно отражающих ПрО, в виде инструментария – генератора проектов, который позволяет на конечных этапах генерировать программный код системы и выполнять технологическую сборку [7].

Разработки в данном направлении велись с 80-х годов [20, 9]. Однако, в случае использования метода генерации проектов имееют место недостатки, связанные с употреблением генерируемых скриптов вместе с текстами программного кода для сборки программ проекта по исходным текстам в соответствии с выбранной платформой. Это является причиной разработки и поддержки дополнительного ПО, которое анализирует полученные скрипты.

В 80-х годах в качестве инструментария для генерации программной архитектуры была предложена диалоговая система, позволяющая осуществлять синтез ПО для промышленных объектов [10]. Ключевыми особенностями данной разработки является оптимизация процесса

синтеза программной архитектуры путем: 1) запоминания состояния задачи на любом этапе с последующим восстановлением; 2) проверки исходных данных задачи до ее решения, в процессе решения и после него; 3) оперативного ввода исправлений в исходные данные; 4) предоставления пользователю возможности многосеансной работы. Такой функционал позволяет прорабатывать различные стратегии решения задачи, минимизировать время ожидания решения задачи за счет снижения количества ошибок, приводящих к сбою ВП и разбивать последовательность действий пользователя на этапы (сеансы) с длительными временными перерывами между ними. Для осуществления синтеза программной архитектуры делается упор на развитие проблемно-ориентированного языка.

В 90-х годах было предложено использование диалоговых систем для автоматизированного формирования ВП на основе маршрутов, выделяемых на графовой модели вычислений [11]. Однако развитие диалоговых систем пошло в сторону проектирования диалога на базе естественного языка и в дальнейшем получило развитие в применении искусственного интеллекта и баз знаний [12].

В настоящее время также известны некоторые системы автоматизированного технологического проектирования, работающие в диалоговом режиме, например, «ТехноПро» [13]. Все же на современном этапе уделяется мало внимания вопросам оптимизации диалогового интерфейса пользователя в смысле разграничения функциональной нагрузки между пользователем и ПО ИС. В целях повышения эффективности работы пользователя применяются попытки оценивания физиологических показателей [28].

Выделяют также технологию автоматического синтеза программной архитектуры с использованием онтологии прецедентов [15]. Онтологическое описание позволяет накапливать опыт разработки, выполнять автоматическую классификацию программ на основе их спецификаций и выполнять построение программ путем адаптации известных решений [16]. Онтологическое моделирование ПрО получило развитие в области GRID-технологии, облачных вычислений, технологий e-Science [17].

Другое направление в автоматическом синтезе архитектупы ПО связано с развитием технологии генетических алгоритмов, позволяющих определять структуру управляющего автомата, который в свою очередь является системой вложенных и взаимовызываемых автоматов [18]. Такой подход реализует технологию автоматного программирования и обладает такими преимуществами как автоматизация процесса верификации, документируемости, упрощение процедуры внесения изменений. Автоматный подход находит широкое применение не только в синтезе программного кода [19], но и в управлении поведением самой системы, запущенной на выполнение.

Заслуживают внимания методы и средства автоматического построения параллельных программ с использованием технологии CUDA по непроцедурным спецификациям [20]. Такая технология базируется на методах декларативного программирования, что позволяет получать программу с высоким уровнем абстракции с отражением самого метода решения, а не его реализацию при конкретных условиях. Недостаток такого подхода – ограниченная применяемость в силу недостаточной универсальности методов, неприменимых для другого аппаратного обеспечения и ПрО.

Одним из эффективных подходов на современном этапе является сервис-ориентированный подход. Он требует применения архитектурных фреймворков и банков видов моделей, методов, знаний, правил и алгоритмов для конструирования ИС в рамках выбранной методологии. Для описания программной архитектуры ИС широко используются ADL-языки (Architectural description languages) [21]. ADL-языки используются в качестве средств описания архитектурных спецификаций, их интеграции с целевыми моделями ИС, описанных аспект-ориентированными графами [22]. В случае параметрического синтеза ПО ИС применяются модели многослойного графа [23]. Кроме специализированых ADL-языков также применяются UML нотации для описания программной архитектуры.

Применение архитектурных шаблонов и проблемно-ориентированных языков описания архитектуры находит применение в развитии технологии построения композитных приложений [24].

С усложнением ПрО увеличивается сложность аппаратного и программного обеспечения. Повышаются требования к эффективности и производительности ПО (стандарт ISO/IEC 25041:2012). Это требует применения альтернативных технологий увеличения вычислительной мощности [25]. Решение проблем оптимизации ВП сводится к технологиям организации распределенных и параллельных вычислений. Однако в этой связи растет сложность проектирования и разработки качественного ПО. В случае невозможности непосредственного использования конечным пользователем ИС с проблемной ориентацией на базе многопроцессорной или распределенной архитектуры прользователю предлагается использование GRID-технологии и предоставление вычислительных мощностей компьютерного кластера в качестве сервиса [26].

Сервисно-ориентированные технологии организации кластерных вычислений являются на данный момент наиболее перпективными. Они порождают новое ответвление – облачные технологии. Однако в области сервисно-ориентированных технологий присутствует существенная проблема – обеспечение безопасности данных, находящихся во владении сторонних организаций. Среди известных типов угроз (сетевые атаки, вредоносное ПО, уязвимости в приложениях и ОС) при использовании облачных технологий добавляются сложности, связанные с контролем среды (гипервизора), трафика между гостевыми машинами и разграничением прав доступа [27].

В случае использования итерационных моделей ЖЦ ИС и динамически формируемых требований используются Agile-технологии, которые показывают свою эффективность для небольших компаний-разработчиков ПО. В случаях средних и крупных компаний, которые ведут разработку сложных программных систем с заданными высокими требованиями к надежности, точности и эффективности, применяются технологии, учитывающие функциональные и нефункциональные требования конечного пользователя. Программные спецификации, составляемые на основе требований, используются в качестве основы для разработки ПО через тестирование TTD [28] и с учетом поведенческих свойств ПО – BDD [29, 30]. Однако постепенный рост дополнительной функциональности готовых программных продуктов в процессе его технической поддержки приводит к неконтролируемому разрастанию программной архитектуры, увеличению сложности программ, что влечет за собой увеличение стоимости или прекращение сопровождения ПО. Это становится проблемой для технологий гибкой и интенсивной технологий разработки в современных условиях рынка ПО [31]. В качестве решения проблемы неконтролируемого разрастания функциональности программ предлагается использовать мониторинг актуального состояния ПО на предмет идентификации устаревшей функциональности и удалении ненужных функций из программной архитектуры. Автоматический процес сокращения функциональной сложности базируется на отслеживании изменения значимости функций приложения во времени от версии к версии в компании разработчика путем фиксации частоты использования функций ПО конечным пользователем.

Рост функциональной сложности, являющийся одной из причин повышения стоимости сопровождения программных продуктов, обусловливает необходимость применения также и другого подхода в решении этой проблемы. Кстомизация ПО для бизнес-процессов и ИТ-инфраструктуры является выходом из сложившейся ситуации, например, для больших программных комплексов такого класса как ERP-системы [32]. Поскольку такие системы требуют нескольких месяцев или лет для развертывания, внедрения и начала успешной эксплуатации [33]. Кастомизация подразумевает адаптацию ПО к организационной структуре ИС, основываясь на использовании сервис-ориентированной архитектуры и сервис-доминантной логики (Service Dominant Logic) [32]. Для автоматизации процессов кастомизации и снижения функциональной сложности ПО требуется применение формального аппарата –графовых модели программной архитектуры разрабатываемых ИС.

На современном этапе для решения проблем кастомизации и динамического конфигурирования компонентов ПО применяется развитая технология динамических линеек программных продуктов DSPL [34]. Такие системы адаптируются не только к изменяемым требованиям пользователя, но также и к среде функционирования, позволяя изменять свою функциональность без перезагрузки системы. DSPL технология применяется для разработки саморегулирующихся систем. Технология DSPL может базироваться на использовании сервис-ориентированной архитектуры программного продукта [35]. Недостаток этих технологий – недостаточное обеспечение безопасности данных.

По данным отчетов за 2015 и 2016 годы компании Panorama Consulting Solutions о применяемых технологиях в проектировании и разработке программных систем ERP класса наблюдаются тенденции значительного снижения процента использования сервис-ориентированой технологии в рамках модели «приложение как сервис» SaaS (с 33% до 17%); увеличение процента использования ERP-платформы на базе технологии ERP-облако (с 11% до 27%). При этом остается неизменным процент использования локального развертывания и использования ПО (56%) [36, 37]. Согласно данным этого же отчета предприятия, преобретающие программный продукт, отказываются от применения модели SaaS и технологии облачных вычислений по причине недостаточного уровня защиты данных: процент неудовлетворенности возрос с 20% до 29% на фоне снижения других причин (рисунок 1, а).

Для предприятий, которые внедряют ERP-системы, важным показателем является процент кастомизации приобретенного ПО, то есть процент доработок, осуществляемых при адаптации программного функционала к условиям бизнес-процессов предприятий и требованиям пользователей. В соответствии с отчетами 2015-2016 гг.

а *б*

Рис. 1. Данные отчетов: *а* — причины отказа от использования облачных технологий; *б* – требуемая кастомизация приобретенного ПО

наибольшее количество предприятий (41%) требуют 11-25% доработок [36, 37]. Однако на 2016 год их количество снизилось до 31% на фоне медленного увеличения количества предприятий (с 22% до 23%), которые требуют 26-50% кастомизации своего программного продукта (рисунок 1, б). Одновременно с этим, благодаря высокой стандартизации и унификации бизнес-процессов произошло значительное увеличение числа предприятий, требующих небольшие изменения в ПО после преобретения (уровень кастомизации 1-10%). Тем не менее, приведенные данные говорят о сохраняющейся актуальности использования своих технологических решений при проектировании и разработке ПО, которое развертывается локально на предприятиях, но, подобно DSPL или сервис-ориентированным технологиям, позволяет обеспечить высокую гибкость, динамическую конфигурируемость, адаптируемость программной архитектуры и функционала, предлагаемого на рынке ПО.

На основе проведенного анализа, можно выделить следующие методы синтеза архитектуры ПО ИС: 1) ручной; 2) автоматизированный (в т. ч. с использованием метода диалогового конструирования); 3) автоматический.

В рамках автоматизированного и автоматического методов наиболее широко используются следующие средства формализации: 1) логико-алгебраические спецификации; 2) автоматные модели; 3) графовые модели; 3) ADL-языки; 4) средства онтологического инжиниринга.

Перечисленные средства формализации могут применяться в рамках *синтезирующего* (на базе модели вычислений, отображающей понятия и отношения ПрО и программной спецификации), *композиционного* (на базе функций и операций композиции в логико-математической системе) и *сборочного программирования* (на базе модели сборки в виде ориентированного нагруженного графа) [24].

Что касается методов проектирования и разработки ПО ИС с использованием наиболее распространенного сборочного подхода, на данный момент существует следующая обобщающая классификация этих методов [24]: 1) модульно-ориентированный; 2) объектно-ориентированный; 3) компонентно-ориентированный; 4) метод генерации; 5) сервисно-ориентированный.

Каждый последующий в этом списке метод является развитием предыдущего. Для реализации данных методов необходимо использовать хранилища готовых решений, компонент повторного использования. При этом указывается на наличие существенных проблем – обеспечение межмодульного интерфейса при сборке ПО ИС [24] и присутствие конфликта между нефункциональными требованиями к компонентам.

Актуальность использования методов ситеза программной архитектуры ИС обусловливается сложными ПрО, которые характеризуются [38]: 1) структурной сложностью и территориальной распределенностью; 2) функциональной сложностью; 3) информационной сложностью; 4) сложностью динамики поведения при высокой изменчивости внешней среды.

Для проектирования современных технически сложных систем широко применяются системы автоматизированного проектирования (САПР). Основными функциями этих систем являются: автоматизация выполнения различных проектных процедур с целью нахождения оптимальных вариантов проектируемого объекта, автоматизация выбора схемы или конструкции, автоматизация составления проектной и технической документации. САПР, ориентированные на конкретную ПрО, используют специальные методы, алгоритмы и программы, оригинальные математические модели, учитывающие специфические качества объектов проектирования. В целях решения проблем повышения эффективности работы ПО ИС

и возможности координации действий программ указывается на необходимость разделения программной системы на управляющий объект (УО) и объект управления (ОУ). УО может служить некая исполнительная система, определяемая как компоновщик, который формирует общий программный код на основе атомарных фрагментов (существующих программных модулей) и алгоритма сборки. Компоновщик дополняет программную архитектуру диспетчером, который контролирует исполнение функций ПО ИС (рисунок 2) [39].

При указанной архитектуре компоновщик может быть сориентирован на повторное использование программных модулей, компонент, инструментального и прикладного ПО. Он может учитывать версионность сборок, а архитектура ИС при этом может быть расширена до включения в её состав нескольких компаний-разработчиков ПО (рисунок 3), что влечет за собой необходимость разработки процессов повторного применения программ [40].

Компания-разработчик может использовать версии ПО, изменять исходный код и, при необходимости, возвращать ПО данной версии с изменениями, указывая, что было модифицировано. В репозитории эти изменения интегрируются с базовой версией ПО. Интегратором является организация, разрабатывающая новые версии ПО. Подобная архитектура хорошо подходит для спиральных и итеративных моделей ЖЦ ПО. В случае спиральной модели осуществимо накопление и повторное использование программных компонентов, моделей и прототипов. Также возможна ориентация на развитие и модификацию ПО в процессе его проектирования.

Основные стадии ЖЦ ПО могут варьироваться в зависимости от выбранной модели ЖЦ (спиральная модель, RUP, MSF), однако процессы ЖЦ

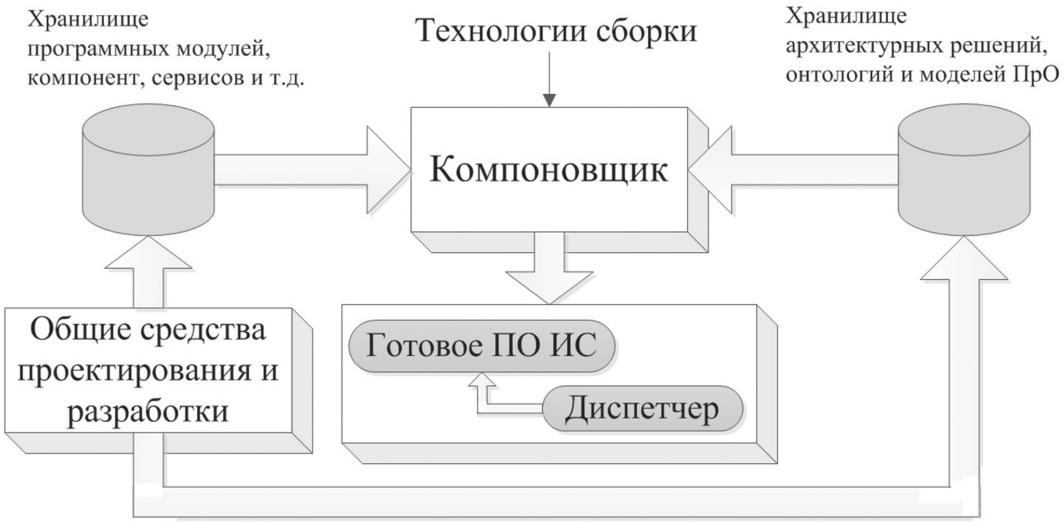

Рис. 2. Концептуальная структура инструментальных средств, реализующих сборку ПО ИС

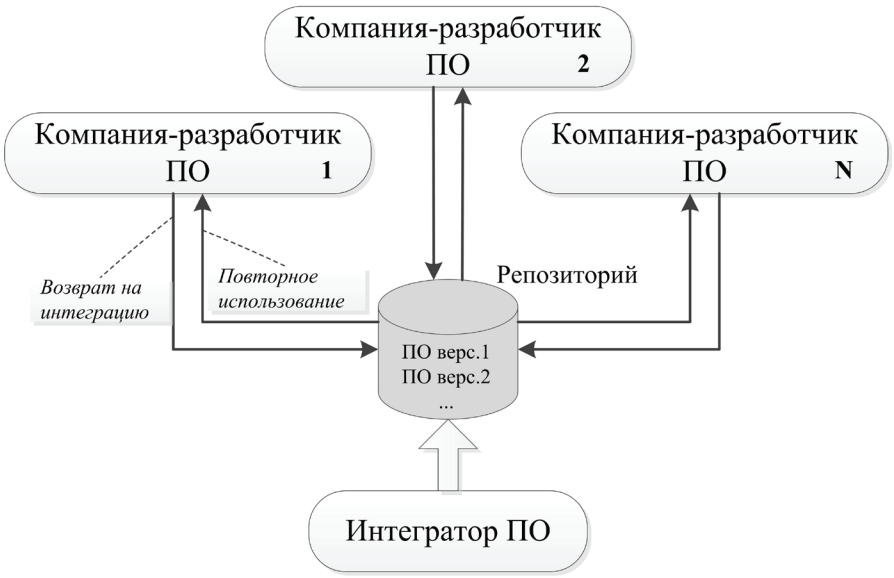

Рис. 3. Менеджмент повторного применения ПО

программных средств регламентируются соответствующими стандартами.

К итеративным моделям относятся наиболее распространенные MSF (Microsoft Solutions Framework) и RUP (Rational Unified Process), которые используют стандарты ISO/IEC [41].

К подвиду итеративных моделей относят модели, применяемые в рамках гибкой методологии разработки ПО (Agile software development): XP (eXtreme Programming), Crystal, FDD (Feature-Driven Development), Scrum, которые могут быть сориентированы как на стандарты CMMI v.1.2 – 1.3 так и на стандарт SPICE (ISO/IEC 15504) [28]. Последние модели (относящиеся к Agile методологии) ориентированы на небольшие компании и штат разработчиков.

Учитывая вариативность представлений стадий ЖЦ в зависимости от моделей, следует ориентироваться на существующие стандарты в этой области для выявления этапов и процессов ЖЦ, которые затрагиваются при применении подходов к оптимизацици ПС.

Согласно выполненному анализу методов синтеза в соответствии со стандартом ISO/IEC 12207:2008 затрагиваются следующие процессы ЖЦ: 1) процесс анализа требований к программным средствам; 2) процесс проектирования архитектуры программных средств; 3) процесс детального проектирования программных средств; 4) процесс конструирования программных средств; 5) процесс комплексирования программных средств. Теория графов нашла широкое применение в рамках указанных процессов. Графовые модели, обладая математической простотой, позволяют описать программную архитектуру и формализуемые задачи ИС. Такие модели являются наглядными и хорошо согласовываются с парадигмами объектно-ориентированного, функционально-ориентированного и компонентно-ориентированного программирования. В связи с тем, что способы представления информации о программной архитектуре ИС, необходимой для генерации программного обеспечения, играют важную роль, подробно рассмотрены графовые модели программной архитектуры.

2. Анализ формальных графовых моделей программной архитектуры информационной системы

Теория графов получила широкое распространение и применяется во многих областях, в частности – для описания программной архитектуры ИС. Графовые модели используются для различных целей в рамках процессов проектирования и разработки ПО: отображения информационных зависимостей между программными компонентами, последовательности выполнения функциональных задач системы, описания версии конфигурации ПО, схемы связей по управлению между элементами программной системы.

Для получения графовой модели программной архитектуры ИС обычно используется информация о заданной ПрО в виде таких моделей ПрО как [42]: 1) информационные модели; 2) модели потоков данных; 3) функциональные модели. Данные модели ПрО берутся за основу при проектировании ПО в рамках существующих технологий проектирования ARIS, IDEF, DFD или UML и являются исходной информацией для формального представления программной архитектуры. Преимуществом графовых моделей архитектуры ПО, базирующихся на формализме теории графов, является способность к визуализации и возможность к автоматизации как процессов проектирования так и разработки программных систем. Теория графов используется для описания архитектуры ПО в граф-ориентированных программных моделях для ВС с параллельной и распределенной архитектурой, а также в целях конфигурирования программных систем [38]. При этом в структуру программмной платформы, реализующей подход к разработке приложений с динамически конфигурируемой параллельной и распределённой архитектурой, включен программный модуль менеджера (диспетчера). Данный модуль осуществляет изменение конфигурации системы во время работы и система не нуждается в перезагрузке. На графовую модель такой программной системы ложится задача описания программных примитивов (программных модулей), реализующих функциональные задачи. В процессе динамического конфигурирования формируется последовательность программных модулей, которые запускаются на выполнение. Эта последовательность исполнения называется конфигурационным планом. Графовая модель представляет собой ориентированный граф, для которого конечному множеству вершин сопоставляются программные модули, а направленным дугам – показатели стоимости и временной задержки передачи данных от одной к другой вершине. Подход к разработке ПО, в основу которого положены граф-ориентированные программные модели, применяется для кластерных вычислений, web-сервисов, компонентно-ориентированных вычислений.

На современном этапе динамическое конфигурирование осуществляется на основе технологии функционального программирования, с использованием функционально-ориентированных языков программирования, таких как Scala [43].

Недостатком указанного граф-ориентированного подхода является то, что при его использовании не учитывается возможность объединения вершин графовой модели для получения супервершин с целью уменьшения связности и сцепления программных модулей, а также упрощения графового представления программной архитектуры ИС. Подобные графовые модели не содержат дополнительной информации, требуемой для их обработки, используются для описания словарей программных классов и представляются кортежем:

$$G = <VC, VA, VR, A, EC, ECO, EA, ER, ERO>,$$

где VC – конструктивные вершины; VA – заменяемые или изменяемые вершины; VR – вершины повторного использования; A – метки; EC – конструктивные дуги; ECO – необязательные конструктивные дуги; EA – изменяемые дуги; ER – дуги повторного использования; ERO – необязательные дуги повторного использования.

Каждой вершине данной графовой модели сопоставляются классы программных модулей. Выделеные на множестве вершин модели подмножество классов повторного использования и подмножество изменяемых классов позволяют описать динамическую часть программной архитектуры. Подобные направленные ациклические графовые модели применяются также в функциональном программировании для автоматизации типизации, поиска соответствующих функций по их заданным аргументам.

Кроме того, на теоретико-множественном уровне представления ИС в качестве основы для процесса синтеза также выступают графовые модели. Примером служит подход, основанный на описании ВП в форме потоков заданий (workflow, WF). В этом случае WF – это ориентированный граф, вершинами которого являются запускаемые задачи, а ребрами – зависимости между задачами по данным и по управлению [44]. Такой подход находит применение для ИС, ориентированных на распределенную вычислительную среду.

Кроме того, среди графовых моделей внимания заслуживают вероятностные модели систем зависимостей. Данные модели на основе графов возникли на стыке многомерного статистического анализа, теории вероятностей, теории графов, теории информации и искусственного интеллекта. Данный класс моделей играет роль строгого языка представления знаний в условиях неопределенности (в частности, в экспертных системах нового поколения) и эффективного аппарата решения разнообразных аналитических задач.

Наиболее привлекательны модели на базе ациклических ориентированных графов (АОГ-модели). Выделяют такие достоинства АОГ-моделей [45]: наглядность, способность отображать причинно-следственные связи и прогнозировать последствия действий (решений), компактное представление систем зависимостей, вычислительная эффективность вероятностного вывода.

Эти свойства обеспечивают эффективное применение таких моделей в медицинской и технической диагностике, социометрическом, эконометрическом и эпидемиологическом анализе, моделировании генетических механизмов, распознавании речи в виде комплекса дисциплин e-Science.

Для задач детального описания архитектуры ПО ИС могут применяться четыре основные графовые модели: граф управления, информационный граф, операционно-логическая история и история реализации [45]. Первые две модели не зависят от входных данных и строятся непосредственно по тексту программы. Две последние модели для своего построения формально требуют слежения за выполнением всех операндов. Сложность построения модели возрастает в порядке указанного перечисления. Все указанные модели существуют для всех программ.

Существует множество сложных научных, инженерных и экономических задач, для решения которых на современных ВС требуется длительное время. При этом количество таких задач постоянно растёт. В крупномасштабных ВС, вероятность потери результатов вычислений очень высока [46]. Исходя из этого, существует серьёзная необходимость обеспечения отказоустойчивого выполнения программ на ВС и в тоже время оптимизации времени работы ВС.

В литературе известен подход к решению этой задачи под названием «backtracking» [47]. Смысл этого метода заключается в использовании «истории» взаимодействий для анализа программы. Вводятся вспомогательные переменные, которые хранят истории взаимодействия по каждому каналу программы. Для хранения историй вводится специальная *историческая переменная* – массив значений, последовательно переданных по соответствующему каналу. А далее при необходимости восстановления этой истории программа обращается к массивам. Кроме принципа backtracking в литературе известен и другой подход, основанный на создании контрольных точек (КТ) вычислений (checkpoint). Подходы к созданию КТ разделяются на реактивный (reactive) и проактивный (proactive) [48]. Наибольшее распространение получил реактивный подход, также называемый *Checkpoint/Restart* или *Rollback/Recovery* [49]. Он предусматривает периодическое создание КТ восстановления, хранящих состояние выполняющейся программы. В случае отказа одного или нескольких

вычислительных узлов (ВУ) любая КТ может быть использована для повторного запуска программы на исправной подсистеме. При этом работа продолжится с момента времени, соответствующего созданию этой КТ. Реактивный подход используется также для балансировки нагрузки ВС и в воспроизводящих отладчиках (Playback Debuggers) [50].

Применение механизма КТ связано с накладными расходами при выполнении параллельных программ (ПП). Данная операция характерна интенсивным использованием узлов ввода-вывода (УВВ), поэтому среднее время создания КТ может быть весьма значительным.

Существует достаточно много средств создания контрольных точек (ССКТ) [51], каждое из которых имеет свои преимущества и недостатки.

Различают две основные схемы взаимодействия ССКТ с защищаемой программой: явная и прозрачная (неявная). Средства создания КТ, построенные на основе явной схемы, позволяют задать ограниченный набор информации, которую необходимо сохранить в КТ. Недостатком явной схемы является необходимость модификации исходного кода, что не позволяет применять её к программам, доступным только в бинарном виде.

Средства создания КТ, построенные на основе прозрачной схемы, осуществляют сохранение КТ незаметно для программы. Недостатком данного подхода является больший объём дискового ввода-вывода информации, так как сохраняется всё пространство памяти.

По классам поддерживаемых программ ССКТ подразделяют на сосредоточенные и распределённые. Сосредоточенные ССКТ обеспечивают отказоустойчивость выполнения одного или нескольких процессов в рамках вычислительного узла ВС. Распределённые ССКТ обычно строятся на базе сосредоточенных и применимы для распределённых и параллельных программ, что делает их важным инструментом организации функционирования ВС.

Для распределённых ССКТ различают координированный и некоординированный подходы. При создании РКТ каждый процесс РП сохраняет свое состояние в КТ. Целостной РКТ называется набор из N локальных КТ, формирующих допустимое состояние программы [49]. Такая РКТ может быть использована для восстановления программы после сбоя. При координированном подходе создание КТ происходит синхронно, что гарантирует целостность РКТ. При некоординированном подходе каждый процесс создает КТ независимо от других. Следовательно, при восстановлении необходимо выполнять поиск целостного состояния программы на основе набора независимых КТ. Последнее вносит дополнительные накладные расходы.

Распределённые ССКТ также можно разделить на универсальные и MPI- ориентированные. Первые ССКТ позволяют создавать РКТ для любых распределённых и параллельных программ, в том числе для различных реализаций модели передачи сообщений (PVM, MPI). Что касается вторых, то существует несколько ССКТ, построенных на базе конкретных реализаций MPI (например, OpenMPI, MVAPICH2). Все они используют пакет BLCR для создания сосредоточенных КТ и реализуют собственные механизмы сохранения графа связей и транзитных сообщений.

Технология автоматного программирования применяется при проектировании такого ПО как системы автоматизации ответственных объектов управления. Стандарт IEC 61499, унифицирующий правила создания распределенных управляющих систем, рекомендует описывать базовые функциональные блоки с помощью конечных автоматов (КА). Выбор в пользу автоматных моделей обусловливается требованиями к организации бесперебойного сохранения и восстановления данных ВП, устранения блокировок и минимизации ошибок [52, 53], что позволяет обеспечивать высокую надежность работы ПО.

Выводы

Выполнен анализ существующих методов структурного синтеза программной архитектуры ИС, в том числе на основе графовых моделей. На основании проведенного анализа установлено, что решение задачи разработки методов и информационных технологий структурного синтеза программной архитектуры, адаптации ПО под изменяющиеся во времени требования конечного пользователя остается актуальным, поскольку на современном этапе достаточно большое число производственных предприятий и компаний (31% – 41%) не удовлетворены уровнем защиты данных для применяемых технологий, а также требуют большой объем доработок исходного ПО, приобретаемого заказчиком. В последнем случае процент кастомизации составляет примерно 11%-25%.

Литература

[1] A Modular Reference Structure for Component-based Architecture Description Languages/Misha Strittmatter, Kiana Rostami, Robert Heinrich [et al.] // ACM/IEEE 18th International Conference on Model Driven Engineering Languages and Systems, September 28, 2015. – Ottawa, Canada, 2015. – P. 36–41.

[2] Ковальчук С. В. Интеллектуальная поддержка процесса конструирования композитных приложений

в распределенных проблемно-ориентированных средах / С. В. Ковальчук, В. Г. Маслов // Изв. вузов. Приборостроение. – 2011. –Т. 54, –№ 10 – С.29–36.

[3] An evolutionary multiobjective optimization approach to component-based software architecture design / R. Li, R. Etemaadi, M.T.M. Emmerich, [at al.] // Congress on Evolutionary Computation (CEC), 5-8 June 2011. – New Orleans, LA.: IEEE, 2011. – P.432–439.

[4] Малышкин В.Э. Параллельное программирование мультикомпьютеров / В. Э. Малышкин, В. Д. Корнеев – Новосибирск: Новосибирский государственный технический университет, 2006. – 452 с.

[5] Левыкин В. М. Модель архитектурного фреймворка ускоренной разработки информационной системы / В. М. Левыкин, М. В. Евланов // Нові технології. – 2013. – № 1-2 (39-40). – С.51–57.

[6] Генератор проектов – средство автоматизации проектирования прикладных информационно-вычислительных систем. / Флёров Ю. А., Вышинский Л. Л., Гринёв И. Л. [и др.] // Автома¬ти¬зация проектирования инженерных и финансовых информацион¬ных систем средствами «Генератора проектов». – М.: ВЦ РАН. – 2010. – С. 3–15.

[7] Инструментальная система ФАКИР / Л. Л. Вышинский, Ю. Д. Прибытков, В. И. Шиленко [и др.] // Известия АН СССР, техническая кибернетика. – 1986. – №3. – С. 6.

[8] [8] Инструментальные средства САПР / Вышинский Л. Л., Гринёв И. Л., Шиленко В. И. [и др.] // Задачи и методы автоматизированного проектирования в авиастроении. – 1991. – С.52–70.

[9] Диалоговая система синтеза многосвязных структур промышленных объектов / [Зайцев И.Д., Кисиль И.М., Вайнер В.Г., Губницкий С.Б.] – Киев: Институт кибернетики. – 1981. – 51 с.

[10] Перевозчикова О. Л. Диалоговые системы / О. Л. Перевозчикова, Е. Л Ющенко; [ин-т кибернетики им. В. М. Глушкова]. – Киев: Наук. Думка, 1990. – 184 с.

[11] Masahiro Sh. Dialog System for Open-Ended Conversation Using Web Documents / Masahiro Shibata, Tomomi Nishiguchi, Yoichi Tomiura // Informatica. – 2009. – № 33. – P.277–284.

[12] Суровцева О. А. Использование потенциала САПР ТП «ТехноПро» для формирования интегрированных комплексов на основе CALS технологий // Состояние и перспективы развития сельскохозяйственного машиностроения: 9-ая междунар. научн.-практ. конф. в рамках 19-й междунар. агропромышленной выставки «Интерагромаш-2016», 2016. – Т. 9. – С. 330–332.

[13] Dan T. An Effort-Based Framework for Evaluating Software Usability Design / Dan Tamir, Carl J. Mueller, Oleg V. Komogortsev // ARPN Journal of Systems and Software. – 2013. –Vol. 3 – № 4. – P.65–77

[14] Корухова Ю. С. Автоматический синтез программ с использованием онтологии прецедентов / Ю. С. Корухова, Н. Н. Фастовец // Программные системы и инструменты: тематический сборник. – Т. 12. – 2011. – С.203–215.

[15] Палагин А. В. Методика проектирования онтологии предметной области / А. В. Палагин, Н. Г. Петренко, К. С. Малахов // Комп'ютерні засоби, мережі та системи. –2011. –№10. – С.5–12.

[16] Зінькович В. М. Онтологічне моделювання предметної області з проблематикою e-Science / В. М. Зінькович // Проблеми програмування. – 2011. – № 3. – С. 30–37

[17] Автоматический синтез системы управления мобильным роботом для решения задачи «Кегельринг» / С. А. Алексеев, А. И. Калиниченко, В. О. Клебан [и др.] // Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. – 2011. – № 2 (72). – С.26–31.

[18] Канжелев С. Ю. Автоматическая генерация кода программ с явным выделением состояний / С. Ю. Канжелев // Software Engineering Conference (Russia) – 2006 (SEC (R): матер. конф. 2006. – 2006. – С. 60–63.

[19] Андрианов А. Н. Автоматическая генерация программ для графических процессоров по непроцедурным спецификациям / А. Н. Андрианов, А. Б. Бугеря, Е. Н. Гладкова // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. – 2014. – № 1(3). – С.5–16.

[20] Левыкин В. М. Модель архитектурного фреймворка ускоренной разработки информационной системы / В. М. Левыкин, М. В. Евланов // Нові технології. – 2013. – № 1-2 (39-40). – С.51–57.

[21] Woods E. Using an Architecture Description Language to Model a Large-Scale Information System – An Industrial Experience Report / E. Woods, R. Bashroush // Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 20-24 Aug. 2012 – Helsinki.: IEEE, 2012. – P.239–243.

[22] Coelho K. From Requirements to Architecture for Software Product Lines / K. Coelho, T. Batista // 9th Working IEEE/IFIP Conference on Software Architecture (WICSA), 20-24 June 2011. – Boulder, CO: IEEE, 2011. – P. 282–289.

[23] Агеев Д. В. Параметрический синтез инфокоммуникационных систем с использованием модели многослойного графа / Д. В. Агеев, Фуад Вехбе // СВЧ-техника и телекоммуникационные технологии (КрыМиКо 2013): 23-я междунар. Крымская конф., 8-13 сентября, 2013 г.: тезисы докл. в 2 т. – Севастополь, 2013. – С. 507–508.

[24] Лаврищева Е. М. Software Engineering компьютерных систем. Парадигмы, технологии и CASE-средства программирования / Лаврищева Е. М.– К.: Наук. думка. – 2013. – 283 с.

[25] Фельдман Л. П. Эффективность реализации параллельных вычислений для кластерных систем на базе интерфейса MPI / Л. П. Фельдман, И. А. Назарова // Наукові праці Донецького національного технічного університету. Серія : Інформатика, кібернетика та обчислювальна техніка. – 2016. – № 1. – С. 136–141.

[26] Mohammadkhanli L. Ranking Approaches for Cloud Computing Services Based on Quality of Service: A Review / Leyli Mohammadkhanli, Arezoo Jahani //ARPN Journal of Systems and Software. – 2014. – Vol. 4. – № 2. – P.55–62.

[27] Сергеев Ю. Управление доступом к виртуальной инфраструктуре с помощью продукта HyTrust / Ю. Сергеев // Jet Info Информационный бюллетень. – 2012. – №3 (224). – 44 с.

[28] Веденеев В. С. Применение экстремального программирования при разработке научных приложений / В. С. Веденеев, И. В. Бычков // Математические структуры и моделирование. – 2014. – №. 4 (32). – С.180–184.

[29] Amodeo E. Learning Behavior-driven Development with javascript + Code / Enrique Amodeo. – Birmingham: Packt Publishing, 2015. – 392 p.

[30] Smart J. F. BDD in Action: Behavior-driven development for the whole software lifecycle / John Ferguson Smart / Publisher: Manning Publications, Shelter Island, NY, 2015. – 384 p.

[31] Marciuska S. Automated Feature Identification in Web Applications / S. Marciuska, C. Gencel, P. Abrahamsson // International Conference on Software Quality. – Springer International Publishing, 2014. – P. 100–114.

[32] Customer-Induced Interactions And Innovation In Professional Services: The Case Of Software Customisation / M. Schaarschmidt, W. Gianfranco, B. Matthias, V. K. Harald // International Journal Of Innovation Management. – 2015. – P.1–38. – Режим доступа: https://www.academia.edu/21938879/Customer-Induced_Interactions_ and_Innovation_in_Professional_Services_The_Case_of_Software_Customization/

[33] Nwankpa J. K. Real Options and Subsequent Technology Adoption: An ERP System Perspective / J. K. Nwankpa, Y. Roumani // System Sciences (HICSS): 48th Hawaii International Conference. – IEEE, 2015. – P. 5020–5027.

[34] Learning and Evolution in Dynamic Software Product Lines / Amir Molzam Sharifloo, Andreas Metzger, Clement Quinton, [at al.] – 2016. – 8 p. – Режим доступа: https://hal.archives-ouvertes.fr/hal-01280837

[35] Baresi L. Service-oriented dynamic software product lines / L. Baresi, S. Guinea, L. Pasquale // Computer. – 2012. – Т. 45. – №. 10. – P. 42–48.

[36] 2015 ERP report [Електронный ресурс] // Panorama Consulting Solutions, LLC. – Denver, Colorado, 2015. – Режим доступа: http://panorama-consulting.com/resource-center/2015-erp-report/

[37] 2016 Report on ERP systems and enterprise software [Електронный ресурс]// Panorama Consulting Solutions, LLC. – Denver, Colorado, 2016. – 32 p. – Режим доступа: http://panorama-consulting.com/resource-center/2016-erp-report/

[38] Вендров А. М. Современные технологии создания программного обеспечения. Обзор / А. М. Вендров // Jet Info. –2004. –№4 (131). – С.3–32.

[39] Lifecycle Management of Open-Source Software in the Public Sector. A Model for Community-Based Application Evolution / Ju. Kääriäinen, P. Pussinen, T. Matinmikko, T. Oikarinen // ARPN Journal of Systems and Software. – 2012. – Vol. 2. – № 11. – P.279-288.

[40] Брагина Т. И. Сравнительный анализ итеративных моделей разработки программного обеспечения / Т. И. Брагина, Г. В. Табунщик // Радіоелектроніка, інформатика, управління. – 2010. – Вып. № 2 (23). – С.130–139.

[41] Чайников С. И. Методы и алгоритмы априорной оценки параметров вычислительных процессов: автореф. дис. на соиск. уч. степени канд. техн. наук: спец. 05.13.01 «Техническая кибернетика и теория информации» / С. И. Чайников.– Харьков: ХИРЭ – 1983.–16 с.

[42] Casadei R. Towards Aggregate Programming in Scala / R. Casadei, M.Viroli // First Workshop on Programming Models and Languages for Distributed Computing. – ACM, 2016. – P. 5.

[43] Князьков К. В. Предметно-ориентированные технологии разработки приложений в распределенных средах /К.В.Князьков, А.В.Ларченко // Изв. вузов. «Приборостроение». – 2011. – Т. 54 – № 10 – С.36–43.

[44] The TETRAD Project: Constraint Based Aids to Causal Model Specification / Richard Scheines, Peter Spirtes, Clark Glymour [et al.] // Multivariate Behavioral Research. – 1998. – Vol. 33 –№ 1. – P.65–118.

[45] Воеводин В. В. Параллельные вычисления / В. В. Воеводин, Вл. В. Воеводин. – Санкт-Петербург.: БВХ-Петербург, 2002. – 608 с.

[46] Поляков А. Ю. Оптимизация времени создания и объема контрольных точек восстановления параллельных программ / А. Ю. Поляков, А. А. Данекина //Вестник СибГУТИ. –2010. – №2 –С.87–100.

[47] Антонов В. В. Построение формальной модели предметной области с применением нечеткой кластеризации / В. В. Антонов, Г. Г. Куликов, Д. В. Антонов // Уфа: УГАТУ. – 2011 –Т. 15 –№ 5 (45). – С. 3–11.

[48] Proactive fault tolerance for HPC with Xen virtualization / A.B. Nagarajan, F. Mueller, C. Engelmann, S. L. Scott // ICS 2007: proc. of the 21st Annual International Conference on Supercomputing. – ACM, New York, 2007. – P. 23–32.

[49] A survey of rollback-recovery protocols in message-passing systems / Elnozahy E. N., Alvisi L., Wang Y. M. [at al.] // ACM Computing Surveys. – 2002. – Vol. 34 – №3. – P. 375–408.

[50] Proactive process-level live migration in HPC environments / C. Wang, F. Mueller, C. Engelmann, S. L. Scott // In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC). – 2008. – P.1–12.

[51] The design and implementation of checkpoint/restart process fault tolerance for Open MPI / Hursey J., Squyres J.M., Mattox T.I. [at al.] // In Proceedings of the 21st IEEE International Parallel and stributed Processing Symposium (IPDPS). – IEEE Computer Society. – 2007. – Vol. 3– №26 – P.1–8.

[52] Построение автоматных программ по спецификации с помощью муравьиного алгоритма на основе графа мутаций / Чивилихин Д. С., Ульянцев В. И., Вяткин В. В. [и др.] // Научно-технический вестник информационных технологий, механики и оптики. – 2014. – № 6 (94). – С. 98–105.

[53] Шелехов В. И. Язык и технология автоматного программирования // Программная инженерия. – №4. – 2014. – С. 3–15.

Поступила в редколлегию 14.05.2019

UDK 519.62

Maksim.V. Shopynskyi[1], Nataliia.V. Golian[2], Iryna.V. Afanasieva[3]

[1]Department of Software Engineering, NURE, Ukraine, maksym.shopynskyi@nure.ua
[2]Department of Software Engineering, NURE, Ukraine, nataliia.golian@nure.ua
[3]Department of Software Engineering, NURE, Ukraine, iryna.afanasieva@nure.ua

PRINCIPLES OF SEARCHING AND SORTING OPTIMIZATION IN SOCIAL NETWORKS USING A MULTI-FACTOR ASSESSMENT SYSTEM

The analysis of social networks, which focuses on the relationship between social entities today is an area of active research. It is a set of tools for research, in particular, in combination with artificial intelligence methods such as machine learning, deep learning. The paper examined the current quality of the assessment of information in social networks, analyzed the methods of searching and sorting information in various social networks, as well as the process of providing recommendations to users. Social media data is an inexhaustible source of research and business opportunities. In general, social media data is information gathered from social networks that shows how users interact with content. Methods of improving search results for personalizing recommendations in social networks are given. These indicators and statistics provide an effective understanding of the strategy of behavior in social networks. The advantages and disadvantages of a multifactor assessment system are considered. The possible ways of integrating the combined system of evaluating information elements by the user to optimize search queries and filtering big data are identified.

SOCIAL NETWORK, SEARCH, FILTERING, MULTI-FACTOR ASSESSMENT SYSTEM, RATING SCALE, BIG DATA

Introduction

Social networks are the main direction of dissemination of information on the Internet in recent years. Most of them use a news feed for distribution of records among users.

Social media data is the source that comes from developing or analyzing social networks. After extracting data, analytics is used to sort out raw information. The more data can be collected, the more informed decisions can be made, which will lead to a better result.

Due to the need of providing relevant information, the actual problem is to optimize the searching and filtering the data provided to the user while viewing the feed.

The purpose of this research – analyze the basic algorithms and methods of searching, filtering and sorting information in social networks, possible ways of their optimization and implementation of these methods to the existing and future networks.

To achieve a goal, the following tasks were set:

- to analyze the basic methods of information search in social networks;
- to formulate possible ways of optimization the sorting and filtration of records in the news feed;
- to explore possible options for implementing the obtained methods to the existing social networks.

1. Analysis of the research problem

Before determining the social profitability of investments, it is necessary to first identify key performance indicators (KPI). These are the various business indicators used to measure and measure success. KPI of social networks are indicators that will help suggest what gives results and what does not. In other words, it is necessary to have data, which key performance indicators of social networks should be monitored and analyzed. Only then can one understand by metrics whether a social strategy is being implemented.

For more than 20 years of history of the social networks' existence, the principles for creating a news feed were largely based on user activity [1]: the posts of friends/subscriptions, groups, and other news sources. In the search process, to sort the found information, initially there was used only the number of preferences (likes) received by a post. During the process of development of neural networks and artificial intelligence, their achievements began to be used to form a "smart" news feed and sort provided information, both in general terms and for each user separately. The relevance of each record is now calculated not only by the total number of preferences, but also by the number of views, related to the subject of the news feed [2].

One of the most important moments when collecting data in social networks is the availability of sufficient information for making an reasoned decision. Understanding user behavior and preferences is important for any business purpose: from determining what content people want to see, to controlling the mood of the community as a whole. The data exists regardless of the network in which the need for research arises: the question is in carrying out the necessary analysis.

Data-driven marketing is the process of obtaining information based on the analysis of indicators extracted from large data on consumer interaction in order to make predictions about future behavior. By collecting data in social networks using targeted platforms, it is possible to make the process of making business decisions more balanced.

Despite the progress in the development of social networks, they have one serious disadvantage. Since

most networks use likes (or their counterparts) to get a response from users [3], for the analysis of user preferences, the algorithms for providing information can only be guided by binary factors (like/dislike). Such a model cannot fully provide a clear assessment and feedback of the extent and the context derived from the user who liked a post.

Take for example one of the most popular social networks – Twitter. Twitter has two basic parameters for evaluating each post (tweet):

- the number of likes (bookmarks). It has two usage options: "liked" (the user rated a tweet) and "not defined" (the user did not rate a tweet);
- the number of retweets (allows to share tweets on your page). Also, it has only two states: the user either shared a tweet or not;
- the owner of the tweet. It allows to focus on the number of author's subscriptions to be displayed in the search top.

Twitter has two methods for displaying found tweets: top results for the latest time and tweets sorted by date of publication. The formation of the top is most influenced by the status of the tweet's author (number of subscribers, total activity). The number of retweets and likes [4] affect tweet's rating less significantly.

The first problem of forming the top of the tweets when searching on Twitter is the universality of the results. The search considers only the user's location (country and city) to create the feed of tweets. There is no more personalized search, since the feedback to users is quite low.

The second problem follows from this: Twitter does not have a unique method for analyzing user actions, since the likes are acting as favorites and cannot be used for real evaluation, and the number of retweets is not informative option for content creation algorithms.

So, the general problem of social networks is paying little attention to active user actions [5]. To create a news feed, they involve highly effective algorithms for machine learning, which should predict what might be liked by a user without having enough information. Due to this, search results in social networks are rather approximate, this may lead to an information collapse against the backdrop of exponential growth in the amount of data in the Internet, depriving many users of the benefit of one of the most valuable information sources.

3. Institution of segmented grading scale

For more detailed feedback from the reaction of users to an information post, it is necessary to expand the simplified scheme of a post evaluation in a social network.

Instead of using likes, the notion of rating ranking is introduced. This rating should have a differentiated scale, depending on the required degree of detail. Major scales may range from 1 to 5 (like a five-point rating system used by some educational institutions) and a scale from 1 to 10 (due to the use of the decimal number system). In addition, since the emoji have become widely popular recently, they can also be used as an emotional assessment (from annoying emotion to satisfaction one, from a sad one to a fun one, etc.).

As the system of records evaluation, in the form of likes, has an advantage in the form of greater ease of use [6], to implement a segmented scale, it will be necessary to re-engineer the user's post evaluation interface to maintain satisfactory ease of use.

The segmented scale requires one action more than the likes system (in the worst situations, where the scale is highly differentiated, there may be needed more actions, but such options are not considered here). In addition, for web-based systems and high-resolution smart devices, additional actions may not be even needed – the whole scale can be compactly positioned on the user's screen. Considering that such a system will not greatly influence the usability of the evaluation, it can be assumed that such a method would not cause dissatisfaction with users.

The segmented scoring scale is now used in many online services: movie evaluation systems, mobile app distributors (Google Play), online stores (AliExpress, Rozetka), and more. It should be noted that the rating scale has already been applied to social networks. An example is a network, where the scale from 1 to 5 is used to evaluate photos. But due to incorrect design and lack of a logical context on the rating scale, this rating system was not able to get the right value.

A differentiated scale affects several aspects of social networking optimization. Thus, the user is given a wider choice to evaluate an item: instead of one option, he can choose from several (depending on the segmentation of the scale). In addition, users can express a negative reaction using low scores on the scale (like the score on Google Play, where the dissatisfied users give an app 1-2-star, etc.).

From the point of view of the internal structure of the social network, the assessment scale will help to optimize the statistics for an information element [7]. Instead of prioritizing the quantitative factor of evaluation (total number of likes), the qualitative factor (number of positive assessments and average score) is now on the foreground.

Also, in addition to the information item ratings, the segmented scale can optimize the recommended entries that are provided to the user. Now, recommendations can consider not only the records that were reviewed and approved by the user, but also their overall rating and preferences of the user (evaluation of related records). This will create an additional impact factor on

search results and recommendations that will make them more differentiated and independent for each network user.

The algorithm for ranking of the record displayed in the search results should change according to the entered scale. For a social network with estimates in the form of likes, in the simplest form it may look like this:

$$R = K * likes \quad (1)$$

R – overall post's rating; K – the user influence factor (calculated for each network separately, depends on the user's environment and their activity); *likes* – the number of likes received by a post.

The easiest version of a ranking with a segmented scoring scale is as follows:

$$R = K * \frac{sum_{rate}}{totalRate} \quad (2)$$

sum_{rate} – the total sum of received rating; *totalRate* – the maximum amount of ratings for a post (the number of ratings multiplied by the maximum rating value in numerical form).

Given that most modern systems use a special policy for minimum and maximum ratings to protect against deliberate increase of the rating, the formula (2) allows you to enter the coefficients of the significance of each estimate from a segmented scale. Then the formula will look like this:

$$R = K * \frac{\sum(rate_i * rateK_i)}{totalRate} \quad (3)$$

$rate_i$ – the i-th element of the assessment scale in numerical form; $rateK_i$ – the coefficient of significance of the i-th element of the scale.

Therefore, a segmented scale for rating of information elements leads to more feedback from the user, which can help to customize the data filtering for each user more independently.

3. Assessment of the information post by factors

One of the major drawbacks in the evaluation of information elements with the help of the likes is the lack of a concrete reason for the assessment or its absence [8]. This system partially answers the question "Did the user like this item?", Which has options for "Yes" and "Not at all". Such an assessment is not very informative for the search algorithms of the social network, therefore for its additional analysis, the subject of the element and its affinity with other elements are considered.

The answer to the question "What exactly the user liked the item?" should expand the range of possible choices for the user. One of the possible solutions to this problem is the introduction of factors for the evaluation of information elements.

Evaluation factor is the unit of evaluation of the information element in the social network. It is an analog of the like, but it has a certain context or category that is used to specify the user's motives.

Compared to the segmented scale, which is responsible for the quantitative assessment of the elements, the task of the factor system is a qualitative analysis of data and obtaining reasons for the preference of an element by the user.

The factors can be categorized according to their versatility, way of display, and the purpose.

In terms of universality, the factors are divided into:

- general – can be applied to any information elements, there are universal evaluation options (for example, "Utility", "Relevance", "Truth", etc.);
- specific – inherent to elements of a certain type or certain subjects (for example, factors such as "Perspective", "Processing", "Exposition", etc.) can be used for evaluation of the photographer works;
- custom – factors created by the user for an individual assessment of the information element.

By the way of displaying, factors can be divided into:

- linguistic (verbal) – transmit values using words and phrases;
- figurative (graphical) – transmit values using images (pictures, emoji, etc.).

By the purpose, the factors are:

- statistical – used to consider the popularity and calculation of the overall rating of the information element;
- logical – used as a voting system, have several alternative choices for shaping further actions and making decisions based on user ratings;
- emotional – used as a psychological element evaluation; have a similarity to a segmented scale of evaluation (when using a scale of images), but they focus not on the quantitative but on the qualitative characteristics of the element.

The introduction of the factor system of evaluation takes place by increasing the number of factors of evaluation from one (likes) to several. At the same time, the main disadvantage of such a system is an increase in the complexity of the element's evaluation by the user. In terms of the latter, the complexity increases in proportion to the number of factors. By Miller's law [9] (about the number of objects an average person can keep in working memory), the maximum number of factors for evaluating one element should be 7 ($\pm$ 2) variants.

Despite the mathematically increasing complexity of the factor system evaluating, many users already have experience working with such a system. Facebook, Github, and Stack provide the evaluation feature with emoji, which is one of the options for factor estimation, so in practice, the introduction of the multifactorial system will not cause inconvenience when evaluating the information elements of the network.

From the point of view of the internal structure of the social network, the introduction of the multi-factor evaluation system can greatly affect the algorithms for data search and filtration [10]. Like filtration systems in online stores, factors can be considered when getting search results, while reducing the number of possible variations and increasing their quality.

If we consider as the basic formula of the rating of an element for a one-factor evaluation system formula (2), then when introducing a multi-factor system, the formula of the element rating in filtering by a factor takes the following form:

$$R_f = K * \frac{factorsCount_f}{sum(factorsCount)} \quad (4)$$

R_f – post rating when filtering by selected factor; $factorsCount_f$ – the number of element ratings for the selected factor; $sum(factorsCount)$ – the number of element ratings by all factors.

To form a general assessment based on factors, coefficients of the influence of a factor on the overall assessment of the element can be introduced. In this case, the formula for the overall rating of the item becomes the following:

$$R = K * \frac{\sum(K_i^f * factorsCount_i)}{totalRatesCount} \quad (5)$$

K_i^f – the coefficient of influence of the i-th factor on the overall rating; $factorsCount_i$ – the number of element ratings on the i-th factor; $totalRatesCount$ – the total number of element ratings by all factors.

Thus, the multi-factor system allows users to evaluate information elements in a qualitative way. In addition, filtering can be applied to certain data by searching for data, which can greatly optimize the search results, including for each user independently (based on only his factor ratings).

4. Ways of factor system integration

To optimize the search and filtering of information in social networks and increase the role of users in the evaluation of elements, the combination of a segmented scale and a multi-factor evaluation system is optimal [11].

In this case, each factor will have a universal scale (numeric or figurative) for element evaluation. Thus, the number of options for the user's assessment will increase $g * f$ times (g – the number of segmented scale options; f – the number of factors) in comparison with the one-factor evaluation system based on the likes.

Given the formulas (3), (5) and integration options the general formula of ranking information element in the social network with a combined system testing (segmented scale with the multifactor system) should be as follows:

$$R = K * \frac{K_i^f * factorsCount_i * avgRate_i}{totalRatesCount} \quad (6)$$

$$avgRate_i = K * \frac{\sum(rate_j * rateK_j)}{totalRate_i} \quad (7)$$

R – the overall post rating; K – the user influence coefficient; K_i^f – the coefficient of the influence of the i-th factor on the overall rating; $factorsCount_i$ – the number of element ratings on the i-th factor; $totalRatesCount$ – the total number of element ratings by all factors; $avgRate_i$ – the average element evaluation for the i-th factor; $rate_j$ – the j-th element of the estimation scale in numerical form; $rateK_j$ – the coefficient of the significance of the j-th element of the scale; $totalRate_i$ – the maximum sum of ratings for a certain element on the i-th factor.

The integration of such a combined assessment system is simpler to consider using the Facebook social network as an example. The network already has a simplified system of factors for evaluating records (in the form of emoji). You can use a 5-point or 10-point rating system with stars (or other images) to enter a segmented scale for each factor. In this case, the overall assessment composition will continue (use of graphic images). The disadvantage of such a system is the increase of the number of user actions to evaluate an item. Depending on the device, the number of actions displayed will increase by one for the users of the web interface of personal computers (when the scale for each factor in the form of a popup card is displayed) and two for mobile users (opening of the scale of the factor and the choice of an option).

The integration of such a system into the Twitter network, which was discussed earlier, needs to be done gradually, as this network, unlike Facebook, has neither a multivariate system nor a segmented scale of evaluation. The first is to integrate the system of factors based on graphic images (like those already known to the user evaluation options). With enough level of development by the users of the factor system, you can enter a rating scale for each factor (like Facebook, as described above). In this case, the role of likes becomes minor, they should be revisited for use only as a saving of the elements to the chosen bookmarks.

Integration of such a system involves possible risks during operation. The most likely of these is the increase in the complexity of the assessment system and the lack of instant evaluation with one action. On the one hand, the complexity of evaluating a multi-factor system using a segmented scale causes the user to spend more time to make a decision, but on the other hand, such a system provides the user with much greater freedom of action and options, which, of course, brings the social network to a new level of progress in terms of searching, filtering and sorting information.

Depending on the number of factors and the differentiation of the scale, the number of options for evaluating one element for the user increases by 10-50 times, which will take into account the preferences of the users when giving recommendations (including advertisements) and provide the user with an expression of their opinion when evaluating the item much more concretely and clearer than usual "like it".

Conclusion

The main problem of popular social networks in data retrieval and filtration is the lack of an extensive system of evaluation of information elements that allows the user to evaluate the records more qualitatively and specifically.

The paper considers the main evaluation options used in popular social networks. Using Twitter as the example of the social network, the features were analyzed of the use of the evaluation system in the form of likes and possible options for its improvement. The problems connected with the sorting of data during search and giving personal recommendations to the user were analyzed too.

It has been determined that the introduction of a segmented scale will help users to evaluate information more variably and critically (like the evaluation of movies in online cinemas and goods in online stores). On the other hand, it was found that the multi-factor system greatly improves the qualitative assessment of the element, allowing users not only to express their preferences for some information element, but also to indicate the reason for such an assessment.

Determined two development options of the evaluation system, it has been determined that a combination of a segmented scale with a multi-factor system increases the number of possible evaluation options for the user by tens of times. This allows to optimize search results using variables scale factors and increase the number of possible options for filtering data using factors and their impact coefficients on the overall rating of the information elements.

Using Facebook and Twitter as examples of the social networks there were considered the possible ways of integrating the combined assessment system, possible risks and ways to eliminate them. Thus, the emphasis is placed on the fact that a social network requires a more sophisticated evaluation system, which will help to sort information qualitatively with the increase in the amount of data in the network.

Subsequent studies include a detailed analysis of the integration of such a system of evaluation and creating a new social network based on it. In addition, one of the key tasks is to get feedback from users on the complexity of the combined rating system to determine the main areas of information evaluation development to optimize the sorting and filtering of big data [12], whose number is continuously increasing.

Conflict of interest

The authors declare no Conflict of interest.

References

[1] *Wasserman S.* Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences) // Cambridge University Press. – 2012. – 857p.

[2] *Kadushin C.* Understanding Social Networks: Theories, Concepts, and Findings // Oxford University Press. – 2011. – 264p.

[3] *Easley D, Kleinberg J.* Networks, Crowds, and Markets // Cambridge University Press. – 2010. – 744p.

[4] *Cha M., Haddadi H., Benevenuto F., & Gummadi P. K.*. Measuring User Influence in Twitter: The Million Follower Fallacy // Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. – 2010. – P. 10-17.

[5] *McCulloh I., Armstrong H., Johnson A.* Social Network Analysis with Applications // Wiley. – 2013. – 320p.

[6] *Tsvetovat M., Kouznetsov A.* Social Network Analysis for Startups: Finding Connections on the Social Web // O'Reilly Media. –2011. – 192p.

[7] *Golbeck J.* Analyzing the Social Web // Morgan Kaufmann. – 2013. – 290p.

[8] *Stieglitz S., Mirbabaie M., Ross B., Neuberger C.* Social media analytics – Challenges in topic discovery, data collection, and data preparation // International Journal of Information Management. – 2017. – P. 156-168.

[9] *C. Amit, J. Van Hillegersberg.* Exploring the Impact of Socio-Technical Core-Periphery Structures in Open Source Software Development, // Journal of Information Technology. – 2010. – P. 216-229.

[10] *T. Kesava, G. Mohan Ram.* A Novel Sorting Algorithm for Data Analysis // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. – 2018. – P. 1454-1456.

[11] *S. Stieglitz, L. Dang-Xuan, A. Bruns, C. Neuberger.* Social Media Analytics – An Interdisciplinary Approach and Its Implications for Information Systems // Business & Information Systems Engineering. – 2014. P. – 89-96.

[12] *Yang M., Kiang M., Shang W.* Filtering big data from social media – Building an early warning system for adverse drug reaction // Journal of Biomedical Informatics. – 2015. – P. 230-240.

The article was delivered to your editory stuff on the 22.05.2019

UDK 004.89

Bilous N.[1], Tereshchenko G.[2], Kyrychenko I.[3]

[1] Candidate of Technical Sciences, Professor of Software Engineering Department, Kharkov National University of Radio Electronics, nataliya.bilous@nure.ua, ORCID iD: 0000-0002-8850-9316

[2] Graduate students of the Department of Software Engineering, Kharkov National University of Radio Electronics, hlib.tereshchenko@nure.ua, ORCID iD: 0000-0001-8731-2135

[3] Candidate of Technical Sciences, Assistant of the Department of Software Engineering, Kharkov National University of Radio Electronics, iryna.kyrychenko@nure.ua, ORCID iD: 0000-0002-7686-6439

COPYRIGHT PROTECTION USING BLOCKCHAIN

This article examines the potential and limitations of blockchain technology and blockchain-based smart contracts in relation to copyright. Copyright has long been enforced through technological means, specifically Digital Rights Management. With the emergence of blockchains, many are now predicting a new era regarding the administration and enforcement of copyright through computer code. The article introduces the technology and related potential and limitations while stressing its capacity to act as a form of normative ordering that can express public or private objectives.

DIGITAL RIGHTS MANAGEMENT, COPYRIGHT, BLOCKCHAIN, SMART CONTRACTS, PRIVATE ORDERING, PERMITTED USES

Introduction

Current issue for today is who will regulate the author's and property issues in the 21st century. The matter concerns not only works of art and science, it concerns even your things, houses, motorcycles. Who or what will control the property and legal documents in the 21st century?

It is clear that this will not deal with the old market with offices and corporations dictating the conditions. Publishers, producers, promoters and PR people will be forced to look for new ways to earn on air and adapt to the growth of individualism, which is so often labeled "Western" and advised to get rid of many religions of the world. A system where the majority has to put up with the rules for the dissemination of information and goods, imposed by several large uncles, is a thing of the past.

And perhaps you ask: why so many people are deprive of work? After all, many will have a hard time, when it is impossible to earn on the fruits of someone else's work and creativity. Well, firstly, while the author receives "interest from sales" the media giant has superprofits, which it is sometimes so convenient to hide. This is not fair to the author. Secondly, if the authors themselves are not against the free circulation of everything being created, and some of them are even actively advocating, we must clearly understand and admit that we have the right to reform [1].

In addition, some products of creativity can impose cryptographic methods of protection. In the New World there will be no bloody revolution, the technology itself will make some ideas obsolete and others will be relevant, and since technology is inextricably linked with progress, if it is not weapon technology, it will only move it forward by overcoming artificial restrictions: superstitions, religion, ideology, dogma, titles, titles, legislative prohibitions, important pathos, and so on.

1. Blockchain

Blockchain — is a decentralized database, the storage devices of which are not connected to a shared server. Any member of the system can make records in it, which after verification are automatically displayed on all computers on the network. Thus, the main advantage of the blockchain is the safety of the information stored in it [2]. No one can forge or replace it, as unilateral changes on each of the network devices will require huge computing capabilities that are not available to the ordinary participant of the system. In the field of intellectual property (IP) protection, the technology will provide safe storage and prompt updating of information about any IP objects: from a patent for high-tech development to a musical work.

2. Copyright protection

The system for the protection of trademarks and patents is currently based on registers of rights to IP objects maintained by authorized state bodies. It is also important here that when entering the above-mentioned registers, non-trivial manual work should be performed to verify the IP object for protection. The system for the protection of trademarks and patents is currently based on registers of rights to IP objects maintained by authorized state bodies. It is also important here that when entering the above-mentioned registers, non-trivial manual work should be performed to verify the IP object for protection. Strange as it may seem, new technologies, some of which are called cryptoanarchic, came to the right holder to help. This is a

decentralized system for exchanging, storing and processing data. Or about the blockchain. The well-known advantages of the new technology are that it is possible to deposit authorship without the participation of a third party and without binding to geography. In the decentralized registry, you can store information about the output parameters of the author's object, as well as the object itself (or its digital imprint, if you need to save on the volume of the file blockchain). The authenticity of the object is confirmed by a cryptographic guarantee - a kind of digital seal. There are several startups that implement the certification of documents uploaded to a distributed database. Potentially, they will be able to solve, for example, the question of the authenticity of the authorship of photographs, which are purchased in stock shops. Such mechanisms can be used to write to the block and the right to own a licensed software (and check the license by the manufacturer or automatically when connected to the blockchain).

A revolution, or at least tectonic shifts, is called blockchain in the music industry. The fact is that everything related to royalties paid to the author of a musical work is a very complex and often opaque process. Blockchain and smart contracts solve this problem, as they eliminate the functions of organizations managing copyright and related rights - no more mediators, and hence distortions and additional costs. This is especially true for unknown performers who are just starting their career - these musicians simply do not have the money to enter into contracts with major labels.

However, entering into a certain state or non-state register can be useful for a number of purposes, including facilitating the receipt by users of information about the current rightholder. There may well be a prospect for using a detachment [3].

For example, downloading a movie (book, program) from the Internet, you along with it can get information about the current copyright holder and the terms of his public license (that is, how much you have to pay for downloading). If proper software is available, the download itself may be carried out upon payment in accordance with these conditions. Or, based on this technology, you can organize the exchange of electronic copies of movies (books, etc.), if, of course, this is the permission of the original copyright holder. Then users will be able to transfer files to each other (as the bitcoins are being transmitted right now), similar to how a paper book or a movie disc is transmitted. Including for money (real or virtual).

Each content unit has its own price. Actions in many content platforms on the blockchain, for example "Voice", are reduced to money. That is, each unit of text, each photo, a comment in the system is automatically assigned a certain amount. This amount can grow depending on how other users interact with this content: read it or create something new based on it. That is, each unit of text, each photo, a comment in the system is automatically assigned a certain amount

The author can choose the appropriate privacy policy for each content unit. When a user simply enters the block with the author's text, the author receives a reward. When the reader wants to use a fragment of the text for his own purposes, he chooses the appropriate item (if it is provided for by the privacy policy) and the author gets even more reward.

In addition, information about the source (author) of the text fragment will be in the chain with the new content created on its basis. But here is another thing.

Content-platforms on the blockchain themselves monitor the observance of your copyright. After the content unit is created and published in the system, it is assigned a code that is automatically "checked for uniqueness" for all units of the block system. You correctly understood: after adding each new element the whole system is updated.

It's as if Google were updating its search engine after the next page appeared. Only much, much faster. First of all, due to the fact that all data is not on any servers, like Google, but distributed among all participants of the block system. That is everywhere and nowhere at the same time.

Since your most evil critic is yourself in a year, many blockchain based content platforms open up opportunities for editing or deleting blocks with content without destroying the whole chain. Such developments are, for example, in Bandcamp and Accenture. That is where, the block will disappear, but if the deleted text fragment has already been used somewhere, information about you as an author will be preserved. This is what Americans call "legacy", a legacy. The bad news: no one will forget what you wrote that night. Good: you can manage content as a valuable asset, bequeathing it to heirs.

The principle of blockchain solves a very important issue in the world of information domination over the individual: how to preserve mercantilist benefits and leave a trace in history without investing in a PR brand, simply doing its own thing. In order to earn more, the author can assign any meta tags to the content units, improving their visibility in the system, but in the top of the tape there will still be texts with the greatest number of interactions from real users-backers. No cheat from the bots.

Blockchain can offer revolutionary changes for marketing, but this has never come into fashion among marketers and has not become a trend, like Snapchat and online video.

3. Future of blockchain

The chain of blocks or blockchain of bitcoin is well known due to its use as a kind of ledger for dealing with digital currencies. At the same time, this technology has the potential to be used to solve other, very radical, tasks. In particular, he gives us an idea of how bitcoin might one day affect the scope of intellectual property and intellectual rights.

And what about the disputes between the numerous "authors"? Constantly someone is suing someone for property or copyright, although the dispute can be resolved by a timely small entry in a block of several tens of kilobytes. Authors of works can prove authorship with the help of this system. It will no longer be necessary to resort to special complicated legal manipulations. It is enough to create in the locker, which is guaranteed never to be changed or erased, an encrypted entry with the first, seventeenth and last page of your book, for example. Leaving or receiving data, according to which you can be accurately acknowledged by the author of this record.

In the not too distant future, mankind will go on to write down not only pieces of its creativity at a certain time in the structure of the detachment, but also the management of the "logs" of life: incidents, data, discoveries, laws and rights will be written into the blockchain. It is not at all necessary to have an archive building or a state "office" servicing owners' catalogs. Governments in general pretend that they do not understand: maintaining secret documentation in an encrypted block system helps to avoid leakage.

Thus, it is possible to exclude the influence of centralized bodies - all-pervasive and senselessly harsh - on consumers and creators of content and products. The severity of corporations is caused by the desire for power, and the people always, at least subconsciously, strive for freedom. Imagine, by the way, how much money for the budget we will save! We do not need a building, furniture, light and water, computers and food in order to build archives and offices, hire staff there. Confirm the right to own creativity and inventions can be using a smartphone.

How do you model a step-by-step creation of a real hit, where do developers, actors or stunt people consult about all the details of the game, the movie or even the play with bitcoin investors? Already today, just adding a button "type" to our torrent distributions and receiving constructive comments, we partially step on this innovative path [4].

Your ideas can no longer be stolen! Everyone will be able to record unconditional priority and authorship of both artistic and scientific works. The world "Bit-net" will provide an opportunity to link real assets and assets to the blockchain. Truthfulness is assured, and a whole bunch of bureaucrats, lawyers and other air sellers are left without work. But the work will appear among millions of authors, inventors, artists around the world.

Some believe that their creativity is of no interest to anyone, but this is far from true! It is exclusive consumption that will become a characteristic feature of the first bitcom users. Authors of news, scientific, entertainment and other content, rare goods and services have already learned to advertise themselves through the Internet and attract public attention, at the same time "burying" television and physical advertising.

Now it's time to start taking real money for your efforts. Your audience is not limited to the territory of your native city or home state. Why should its finances be limited to this territory?

While bitcoins will flow into your pockets, copyright will undergo some "moral" changes, so that suddenly it does not turn out that bitcoins flow into the pockets of some other company that stole your idea. To get acquainted with works of art and science will be completely free. As a consumer, you will be able to search, read, watch, listen and copy to yourself anything you like without restrictions and legitimately. In some corners of the Internet, similar services are already available even in the CIS, payment? Bitcoins, of course! If you think that only drugs are sold for crypto currency, you live in 2009.

As the author of music, books or games, you can make a profit and "tip" from people from all over the planet, in volumes that are still difficult to predict. Not only will you receive material support from the fan community, you will also multiply it. Since the rate is very variable, people's interest will only grow, and your current "state" of $ 5 tomorrow may turn into 25 ... or 50.

The Internet bit of the future will allow ordinary people with small and medium income to get a chance to enjoy the results of creative and intellectual work of all ages, at the same time offering the whole world their services. Not only classics, but modern works will be available anywhere in the world. The importance of a total rethinking of copyright in the 21st century is great, because now there is a huge amount of priceless information, which should be caught literally "on the fly". The first thing that comes to mind is that the proof of existence can be used to confirm the authenticity of the certificate of property without revealing its contents.

You put a hash next to the link to download the file and check the hash yourself. Even if someone breaks into your server, it will not be able to change it. Using this method, you can unambiguously prove that the document or part of the code was checked at a particular time, and the global database of transaction accounting in the Bitcoin network is ideally suited as a means of its implementation. These are just some of the directions for applying this service.

Taking into account the described potential of services, which are possible for implementation on the basis of evidence of existence, this principle may prove even more valuable than the cost of bitcoin, on which most investors are obsessed today.

Digital property can sometimes be viewed as intellectual, and blockchain technology can prove ownership of such property. For example, if you write an article or you have an idea suitable for a patent, in some cases you have to prove that you owned this idea or document earlier than someone else. A check is an example of the potential of a blockchain beyond simple monetary innovation.

Now you need someone to be a third party in proof of identity - like Facebook, Twitter or Google. You could very well do the same, using the architecture of the blockbuster.

4. Blockchain in economy

The technical side of the invention of Satoshi Nakamoto will allow you to develop business in different directions, this will not be exactly for your competitors, and this need not be reported to the authorities. When you build a business, people will buy everything that you invent and offer, if it is useful and interesting. But you need to prove that you are the author. Blokchain provides such a proof, and Bitkoyn is a method of anonymous international payments (anonymous means very fast). In addition, taking the crypto currency for payment today, you create a powerful foundation for your reputation and wealth tomorrow. Do you think that the pioneers of the Internet are rich? What can you read about the first directors, scientific innovators, cosmonauts? And what about the conquerors of America? Sometimes for success it is necessary not to run faster than everyone, but to run out early, and to do it as uniquely as possible.

It is important, however, to understand that the technology of blockchain (the system of keeping the register of rightholders) does not by itself protect against piracy. Since books and films must eventually be converted into a human-readable form, it seems that you can always make an unprotected copy from this form [5].

So in the final account for protection still have to go to the courts, and records in distributed registries can then be used as evidence. Well, that, of course, if the judge knows the word “blockchain”.

5. Blockchain Technologies in the Copyright Domain

The potential of blockchain as a general-purpose technology is currently being experimented with in many domains, including copyright law. Over the past months and years variegated suggestions as to how the technology could be deployed for the management of copyrighted works and neighboring rights objects have been voiced by industry and in the academic literature. In this section, we provide a cursory overview of expected application of these technologies. We organize the following overview around three main drivers leveraging the main characteristics of blockchain technologies. The first driver revolves around the potential capacity of blockchain technology to precisely identify a digital asset and thereby counter the problem of digital “fluidity”. The second driver is related to the ability of blockchain technologies to foster transparent and disintermediated transactions. The third axis focuses on the potential of blockchains to be developed as a DRM system. Finally, in the second sub-section we introduce some structural limits of blockchain technologies such as the so-called “garbage-in garbage-out” problem [6].

6. Prospects for Application

Firstly, it has been argued that DLT could be used to create artificial scarcity in the digital market. Indeed in the copyright domain tokens may represent various elements including a copy of a protected work. This may solve a number of issues related to the fluidity of digital objects and create new business models. This may lead to the commodification of digital works and thereby allowing the creation of new markets. Some projects have already been implemented, in particular in the field of artworks leveraging the fact that blockchain technologies make digital artworks more attractive for collectors. It has also been speculated that these developments create the necessary preconditions for flourishing, technologically-enabled secondary markets for digital content.

DLT may also enable the precise tracking of certain digital assets (through tokens) that could be used as evidence of authorship and provenance. In relation to attribution, hashing can create a unique fingerprint of copyrighted material that allows verification of authorship and that the creative work existed at a given time without revealing the actual contents. The hash allows monitoring of provenance in through recording ownership and usage. DLT has been presented as a “revolution in how to keep track of rights”. Tokens can encode information including the terms of use of protected material (such information can be mentioned under the definition of RMI). For unregistered intellectual property (IP) rights such as copyright and neighboring rights, blockchain technologies offer the benefit of providing a time-stamped record of its conception, use and qualification requirements. For example, the hash may facilitate evidence in court cases concerning copyright authorship and violation of the terms of use [7].

Blockchains’ characteristics provide an opportunity to conceive of a global registry for copyright and neighboring rights. Indeed only the existence of a global registry holding RMI would allow for the development

of potential benefits into real benefits of blockchain and smart contracts as described in the following paragraph. In this regard it is worth mentioning the project currently implemented by PRS for Music, ASCAP and SACEM which aims at improving data accuracy for right holders. This point will be further developed in the next section.

Secondly, there is the prospect of transparency and cost savings related to smart contracts as once a user purchases the digital asset from a website, the smart contract can be triggered immediately so that all other actions – e.g. payment of royalties to right holders – are automated. Combined with digital currencies, this enables micropayments, which could change pricing models in relation to copyrighted materials. Micropayments – meaning the payment of a small sum, such as EUR 0002 – is currently not an economically viable solution as transaction fees exceed the price itself. The advantages of this method are manifold: "the smart contract facilitates microtransactions at little to no fee, and payment is divided nearly instantaneously – per the strict logic of the smart contract code – and is immediately disbursed to the musicians in amounts of less than $0.01". This innovation could also serve to enable an instantaneous, fairer and transparent remuneration of authors and artists. To illustrate, Ujo Music uses smart contracts to facilitate the sale of digital music files. The payment of a certain sum to download a song triggers the smart contract, which divides payment between the various contributors to the song. Notably, this transaction can theoretically occur without the need for a traditional intermediary such as a publisher, a music label or performance organization. Notwithstanding, these platforms still constitute a new form of for-profit intermediaries so that it is still to be determined what economic impact such solutions will eventually have. Thus blockchain promises allowing artists to independently determine prices and individually license their works in a "direct-to-fan" fashion. This appears to offer some remedies for the digital era's challenge of easy unauthorized access to and distribution of copyrighted works.

Some hope that smart contracts will generate disintermediation which would affect incumbents at different levels, including: (1) publishers and music labels, (2) collective management organizations (CMOs), and (3) online platforms. According to others, complete disintermediation is unlikely as blockchains may simply introduce new stakeholders. In the field of online music many blockchain projects promise disintermediation between artists and audience. Yet in reality such actors can be seen as new intermediaries. Indeed, while current discourse frequently envisages authors and artists themselves programming their smart contracts and thus directly defining terms of use, it appears that for numerous reasons of an economic, cultural and technological nature, this is an unrealistic prospect. Sometimes the role of intermediaries goes further than the mere management of legal tools being more related to marketing strategies. In any case, for the "direct-to-fan" model to take hold, solutions need to be devised that can provide a user-friendly form of smart contract management, which does not require the user to personally code the smart contract. It should be noted that some are already working on corresponding solutions [8].

Smart contracts may also play a role in standardizing licensing terms and conditions for copyright works across uses and jurisdictions. Standardized smart contracts, the terms of which can be described in comprehensible language, augment transparency and reduce barriers to using contracts for transactions. The technology could also be used to generate custom smart contracts with the terms of license payment and even its split between various beneficiaries.

Thirdly, some believe that DRM itself may be disrupted by blockchain technology. A number of projects are already underway in this domain. Sony recently applied for a patent for a DRM solution based on blockchain. Kodak launched a similar project, KodakOne, which is aimed specifically at photographers and agencies. Whereas over one trillion photos are uploaded to the web each year, most of them fall into the category of orphan works because it is burdensome for photographers to administer image licensing, infringement detection and reporting. KodakOne seeks to change this by creating an image rights management platform, combined with tokens (to manage instantaneous royalty payments) and smart contracts (to document licenses).

These early initiatives underline that blockchain can serve to create a hard-to-amend record of initial ownership with smart contracts being encoded to license the use of copyrighted works. Here, smart contracts are deployed to automate and standardize copyright-related transactions (such as use and exploitation of content as well as remuneration) in relation to blockchain-based tokenized elements. 144 Smart contracts would be modelled to hold, execute and monitor contractual code. The idea is that smart contracts would be used to establish and self-enforce copyright agreements such as licenses, and provide information about rights in copyrighted materials.

Highlighting that traditional DRM solutions rely on single points of failure, are expensive, can be overcome by a single hacker and interfere negatively with consumer expectations, blockchains' resilience-through-replication is appealing. It is important to note, however, that automated licensing through smart contracts is not to be confused with traditional DRM systems which always involve control of access and use of digital

subject matter. Blockchains may offer right holders greater security and stronger protections against possible attackers including copyright infringers that seek to access the digital asset. User rights would be encoded on a blockchain. Connected systems would then verify these rights and decrypt the related copyrighted content where appropriate. A smart contract would then be used to allocate access to the digital asset via tokens (such as bitcoin, ether, etc.) that reside on the chain, the role of which consists in facilitating remuneration and payments.

Blockchains do not hold the copyrighted digital asset itself in light of the technology's limited processing capabilities but rather facilitate a smart contract that contains information regarding related rights and permissions. However, when users use a work through their device, they trigger communication with the distributed ledger. The DRM system can scan the record for the necessary permission and give the user access to the acquired work. For example, if the user has purchased a limited-duration license, the system can consult a trusted timeserver and compare the time with the contract terms coded on the blockchain and take away access once the user's license has expired. Blockchain technology could therefore control use-rights and just as is the case with current DRM systems, smart contracts do not necessarily encode legally permitted copyright uses. We will return to this point in the following section.

After having summarized the main themes revolving around copyright management by means of blockchain uses, it is worth stressing some structural limitations of blockchain technologies [9].

7. Smart Contracts

In essence, a smart contract is self-executing computer code that automatically processes its inputs when triggered. It is essentially a small computer program that is deployed on a blockchain. Thus while smart contracts are currently being avidly discussed in relation to blockchain, similar mechanisms have been used for a long time, also by DRM systems. As explained, DRM technology essentially embedded copyright law into digital files by limiting the user's ability to view, copy, play, print, or otherwise alter the works. For example, digital audio files encrypted with DRM technology were not subject to the double-spending problem because they contained a basic smart contract, which referenced a centralized network (that is, for example, Apple's server programmed to enforce the iTunes Store Terms and Conditions). Beyond this rather basic definition there is little consensus as to what this terminology really refers. It is worth noting that depending on the adopted definition, smart contracts are not necessarily linked to blockchain technology. Given that they are discussed in relation to distributed ledgers in the literature we introduce below, we will also examine them as such here.

To some, a smart contract is simply a piece of computer code with specific characteristics. Vitalik Buterin portrays smart contracts as "cryptographic 'boxes' that contain value and only unlock it if certain conditions are met". In the blockchain context, a smart contract is "programmed logic that runs on a ledger-like distributed system in response to transaction submission". From a technical perspective, smart contracts are thus simply computer programs that can be consistently executed by a network of computers without the need for an intermediary. Because of their distributed nature and thus guaranteed execution, smart contracts are resilient to tampering, which makes them appealing in many scenarios including the transfer of value.

Yet, because these technical tools can be relied upon to automate transactions typically governed by contract law (such as value transfers) smart contacts can be useful tools in a contractual setting. Nick Szabo indeed coined the term "smart contract" in 1994, to denote "a set of promises, specified in digital form, including protocols within which the parties perform on these promises". Szabo envisaged the creation of computer software resembling contractual clauses to connect parties in a fashion that would make it difficult for one party to unilaterally terminate an agreement. Seen from this perspective, a smart contract can be a Ricardian contract, the objective of which is to create contracts that can be read by humans and machines alike. Indeed to some, a smart contract is "a computer program that both expresses the contents of a contractual agreement and operates the implementation of that content, on the basis of triggers provided by the users or extracted from the environment".

The main value proposition of smart contracts is that of their automated execution. As second-layer applications, smart contracts benefit from the tamper-proof nature of the underlying blockchain infrastructure that anchors their automated execution. Given that many blockchain nodes run smart contract code, it "is not controlled by – and cannot be halted by – any single party". Smart contracts execute automatically and cannot be halted unless this option is specifically built into the code. This enables transactions in situations devoid of human or institutional trust, lowers transaction costs and reduces counterparty risk and interpretative uncertainty. Once an agreement has been translated into code, the intervention of a party or intermediary (other than the respective oracle) triggering contractual execution is replaced by the software's automated execution.

Where smart contracts are used to automate the execution of contractual obligations, performance is thus hard-wired into the code. For example, the software can be used for the automatic transfer of collateral in

the event of default. Automated execution of course not only provides benefits but also disadvantages. Where software executes automatically, unwanted transactions can no longer be rolled back. This can be problematic, such as when a party lacks legal capacity or a party decides to default on its obligations. Modifications, such as those mandated by law or court decisions also cannot easily be accommodated. Through these characteristics, smart contracts promise to trigger efficiency gains particularly attractive in commercial settings, including in relation to copyrighted materials [10].

References

[1] *Gates M.* Blockchain: Ultimate guide to understanding blockchain, bitcoin, cryptocurrencies, smart contracts and the future of money. / Mark Gates., 2017. - 125 p.

[2] *Wright A.* Blockchain: Uncovering Blockchain Technology, Cryptocurrencies, Bitcoin and the Future of Money: Blockchain and Cryptocurrency Exposed (Blockchain and Cryptocurrency as the Future of Money / Alan Wright., 2017. - 130 p.

[3] *Buterin V.* The Business Blockchain: Promise, Practice, and Application of the Next Internet Technology / V. Buterin, W. Mougayar., 2016. - 208 p.

[4] *Vigna P.* The Truth Machine: The Blockchain and the Future of Everything. P. Vigna, M. Casey., 2018. - 302 p.

[5] *Swan M.* Blockchain: Blueprint for a New Economy / Melanie Swan., 2015. - 152 p.

[6] *Williams S.* Blockchain: The Next Everything / Stephen Williams., 2019. – 208 p.

[7] *Antonopoulos A.* Mastering Bitcoin: Programming the Open Blockchain / Andreas Antonopoulos., 2017. – 416 p. – (2nd Edition).

[8] *Tapscott A.* Blockchain Revolution: How the Technology Behind Bitcoin and Other Cryptocurrencies Is Changing the World / A. Tapscott, D. Tapscott., 2018. – 432 p. – (Reprint edition).

[9] *Werbach K.* The Blockchain and the New Architecture of Trust (Information Policy) / Kevin Werbach., 2018. – 344 p.

[10] *Wright A.* Blockchain and the Law: The Rule of Code / A. Wright, P. De Filippi., 2018. – 312 p.

The article was delivered to your editory stuff on the 17.04.2019

UDK 004.89

Nazarov A.[1], Kozel N.[2], Gruzdo I.[3], Kyrychenko I.[4]

[1] Candidate of Technical Sciences, Professor of Software Engineering Department, Kharkov National University of Radio Electronics, oleksii.nazarov1@nure.ua, ORCID iD: 0000-0001-8682-5000

[2] Senior Lecturer at the Department of Software Engineering, Kharkiv National University of Radio Electronics, natalia.kozel1@nure.ua, ORCID iD: 0000-0001-9276-9877

[3] Candidate of Technical Sciences, Associate Professor of Software Engineering, Kharkov National University of Radio Electronics, irina.gruzdo@nure.ua, ORCID iD: 0000-0002-4399-2367

[4] Candidate of Technical Sciences, Assistant of the Department of Software Engineering, Kharkov National University of Radio Electronics, iryna.kyrychenko@nure.ua, ORCID iD: 0000-0002-7686-6439

SECURITY IN DECENTRALIZED DATABASES

Blockchain is a distributed network that records digital transactions on a publicly accessible ledger. This paper explores whether blockchain technology is a suitable platform for the preservation of digital signatures and public/private key pairs. Conventional infrastructures use digital certificates, issued by certification authorities, to declare the authentication of key pairs and digital signatures. This paper suggests that the blockchain's hash functions offer a better strategy for signature preservation than digital certificates. Compared to digital certificates, hashing provides better privacy and security. It is a form of authentication that does not require trust in a third-party authority, and the distributed nature of the blockchain network removes the problem of a single point of failure.

DIGITAL SIGNATURES, BLOCKCHAIN, KEYS, ENCRYPTION, AUTHENTICITY, TRUST.

Introduction

Digital signature (ES) is a special requisite of the document, which allows you to establish the absence of distortion of information in an electronic document since the formation of the ES and confirm that the ES belongs to the owner. The value of the props is obtained as a result of cryptographic transformation of information.

Digital signature allows you to:

- Confirm the authorship of the message sender
- Ensure that no one can forge a message sent and confirmed via an ES

So, with a private key, we sign "letters of transfer of ownership" (transactions), and thus, for example, give our coins to someone else. With the public key (certificate) we verify the authenticity of the transactions of others.

Hashing — transformation of an input array of arbitrary length into a (output) bit string of fixed length. The function that implements the algorithm and performs the conversion is called the "hash function" or "convolution function". The source data is called the input array, "key" or "message". The result of the conversion (output) is called "hash", "hash code", "hash sum", "message summary".

A cryptographic hash function is any hash function that is crypto-resistant, that is, satisfies a number of requirements specific to cryptographic applications.

The fundamental part of Bitcoin are cryptographic algorithms. In particular, the ECDSA algorithm is an Elliptic Curve Digital Signature Algorithm that uses elliptic curves and finite fields to sign data so that a third party can confirm the authenticity of the signature by eliminating the possibility of falsification. ECDSA uses different procedures for signing and verification, consisting of several arithmetic operations[1].

1. Elliptic curves

One form of elliptic curves is Weierstrass curves.

$$y^2 = x^2 + ax + b.$$

For coefficients $a = 0$ and $b = 7$ (used in Bitcoin), the graph of the function takes the following form(Fig.1):

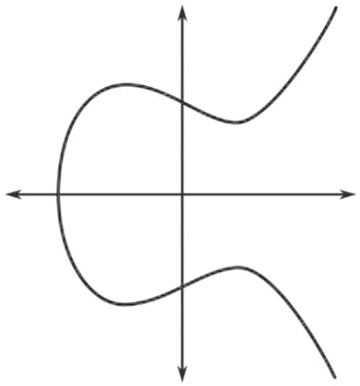

Fig. 1. Elliptic curve

Elliptic curves have several interesting properties, for example, a non-vertical line intersecting two non-tangent points on a curve will cross a third point on the curve. The sum of two points on the $P + Q$ curve is called the R point, which is a reflection of the $-R$ point (constructed by continuing the straight line (P; Q) to the intersection with the curve) relative to the X axis (Fig. 2) [2].

If we draw a straight line through two points having coordinates of the form $P(a, b)$ and $Q(a, -b)$, then it will be parallel to the ordinate axis. In this case there will be no third intersection point. To solve this problem,

a so-called point at infinity (point of infinity) is introduced, denoted as O. Therefore, if there is no intersection, the equation takes the following form $P + Q = O$.

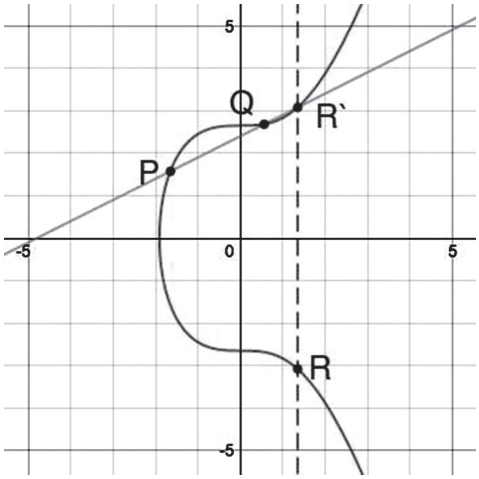

Fig. 2. The sum of two points on the curve

If we want to add the point to itself (double it), then in this case the tangent to the point Q is simply drawn. The resulting intersection point is reflected symmetrically with respect to the X axis (Fig. 3).

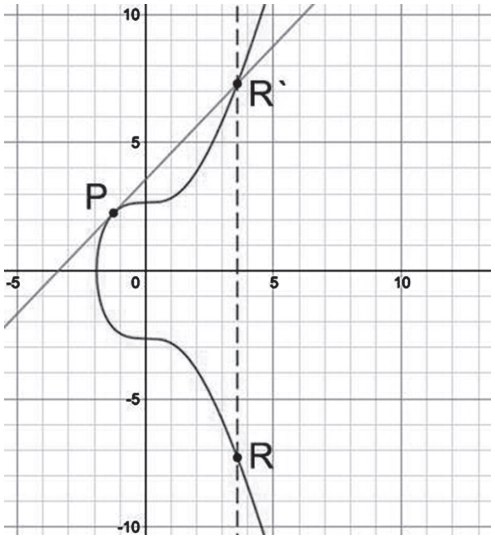

Fig. 3. Double point

These operations allow scalar multiplication of the point $R = k * P$, adding the point P with itself k times. However, note that faster methods are used to work with large numbers [3].

2. Elliptic curve over a finite field

In elliptical cryptography (ECC), the same curve is used, only considered over some finite field. The final field in the context of the ECC can be represented as a predefined set of positive numbers, which should be the result of each calculation.

$$y^2 = x^3 + ax + b \pmod{p}$$

For example, 9 mod 7 = 2. Here we have a finite field from 0 to 6, and all operations modulo 7, no matter how many times they are carried out, will give a result that falls in this range.

All the properties mentioned above (addition, multiplication, point at infinity) for such a function remain in force, although the graph of this curve will not resemble an elliptic curve. The bitcoin elliptic curve, $y^2 = x^3 + 7$, defined on the finite field modulo 67, looks like this (Fig. 4):

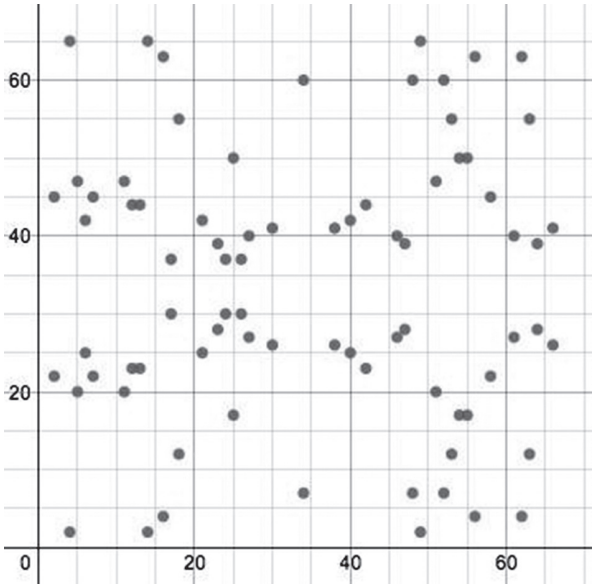

Fig. 4 . Bitcoin elliptic curve defined on a finite field module 67

This is a set of points where all values of x and y are integers between 0 and 66. Straight lines drawn on this graph will now "wrap" around the field as soon as they reach barrier 67 and continue from the other end of it. while maintaining the same slope, but with a shift. For example, the addition of points (2, 22) and (6, 25) in this particular case looks like this (Fig. 5) [4]:

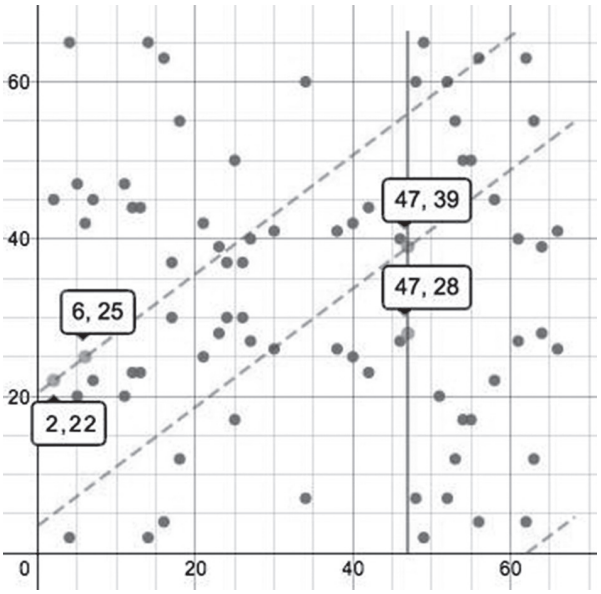

Fig. 5. Addition of points (2, 22) and (6, 25)

3. Bitcoin ECDSA

The Bitcoin protocol contains a set of parameters for an elliptic curve and its finite field, so that each user

uses a well-defined set of equations. Among the fixed parameters, the equation of the curve (equation), the value of the field modulus (prime modulo), the base point on the curve (base point) and the order of the base point (order) are distinguished. About calculating the order of the base point you can read here. This parameter is chosen specifically and is a very large prime number.

In the case of bitcoin, the following values are used:

The equation of an elliptic curve: $y^2 = x^3 + 7$

Simple module: 2256 – 232 – 29 – 28 – 27 – 26 – 24 – 1 = = FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFE FFFFFC2F

Base point:

04 79BE667E F9DCBBAC 55A06295 CE870B07 029BFCDB 2DCE28D9 59F2815B 16F81798 483ADA77 26A3C465 5DA4FBFC 0-1108A8 FD17B448 A6855419 9C47D08F FB10D4B8

The bold font is the x coordinate in hexadecimal notation. It is immediately followed by the Y coordinate.

Order: FFFFFFFF FFFFFFFF FFFFFFFF FFFFFFFE BAAEDCE6 AF48A03B BFD25E8C D0364141

This set of parameters for an elliptic curve is known as secp256k1 and is part of the SEC (Standards for Efficient Cryptography) family of standards proposed for use in cryptography. In Bitcoin, the secp256k1 curve is used in conjunction with the ECDSA (elliptic curve digital signature algorithm). In ECDSA, a secret key is a random number between one and an order value. The public key is generated based on the secret: the latter is multiplied by the value of the base point. The equation has the following form:

$$\text{Public key} = \text{private key} * G$$

This shows that the maximum number of secret keys (consequently, Bitcoin addresses) is, of course, equal to the order. However, the order is an incredibly large number, so accidentally or intentionally pick up the secret key of another user is unrealistic.

The public key is calculated using the same doubling and adding points operations. This is a trivial task that an ordinary personal computer or smartphone solves in milliseconds. But the inverse problem (obtaining a secret key publicly) is a problem of discrete logarithmization, which is considered computationally complicated (although there is no strict proof of this fact). The best known algorithms for its solution, like Pollard's ro, have exponential complexity. For secp256k1, in order to solve a problem, you need about 2128 operations, which will require a computation time on a regular computer, comparable to the lifetime of the Universe[5].

When a private / public key pair is obtained, it can be used to sign data. This data can be of any length. Usually, the first step is to hash the data in order to obtain a unique value with the number of bits equal to the bit order of the curve (256). After hashing, the z-signature algorithm is as follows. Here, G is the base point, n is the order, and d is the secret key.

- Some integer k is selected from 1 to n-1
- Calculate the point $(x, y) = k * G$ using scalar multiplication
- Is $r = x \bmod n$. If $r = 0$, then return to step 1
- There is $s = (z + r * d) / k \bmod n$. If $s = 0$, then return to step 1
- The resulting pair (r, s) is our signature.

After receiving the data and signing it, a third party, knowing the public key, can verify it. The steps to verify the signature are (Q — public key):

- Check that both r and s are in the range from 1 to n-1
- Calculated $w = s - 1 \bmod n$
- Calculated $u = z * w \bmod n$
- Calculated $v = r * w \bmod n$
- Calculate the point $(x, y) = uG + vQ$
- If $r = x \bmod n$, then the signature is true, otherwise it is invalid.

Indeed,

$$uG + vQ = u + vdG = (u + vd)G = (zs - 1 + rds - 1)G = (z + rd)\, s - 1\, G = kG$$

The last equality uses the definition of s at the stage of creating a signature.

ECDSA security is related to the complexity of the secret key search task described above. In addition, the security of the original scheme depends on the "randomness" of choosing k when creating a signature. If the same k value is used more than once, then the secret key can be extracted from the signatures, which is what happened with the Therefore, modern implementations of ECDSA, including those used in most bitcoin wallets, generate k determinedly based on the secret key and the message being signed[6].

4. Other security features

In addition, there are also other elements that protect the blockchain.

More than two users confirm and ensure the security of the transaction. Even in most modern processing systems, only a few levels of verification are involved in verification — as a rule, it is the seller, the buyer, and some third parties (most often a bank or a credit agency).

However, there are from several hundred to several thousand different nodes in the blockchain system, each of which contains a complete copy of the registry of records. Therefore, any of these nodes can also participate in the verification of the transaction, and if the node for some reason does not accept the transaction, it will be canceled. Such an alignment almost to a minimum reduces the possibility of creating a false or fraudulent transaction[7].

The cryptographic keys used by the system in exchange processes are also a miracle of modern cybersecurity. Each encrypted key is a long, complex sequence of data that is practically undecipherable. And if you consider that for confirmation, two such unique keys are required, the system begins to look almost like an impregnable fortress. At the same time, the blockchain is considered to have a unique security system, because with such a level of protection it is possible to retain almost complete transparency of transactions.

5. The most vulnerable places

But, as already mentioned, even the blockchain is not perfect. He, like any other system, has weak spots. So, if you plan to use cryptocurrency and invest your funds in it, or if you have to deal with the blockchain in the future, then you just need to know and understand the potential vulnerabilities of the technology. Therefore, try to remember the following features related to the safety of this technology:

System complexity

If you decide to create a system based on blockchain technology from scratch, then one small mistake can be fatal and "put" all your development. Of course, this cannot be considered a disadvantage of the blockchain itself — rather, it concerns the features of its use. In addition, the average person is much more difficult to understand the blockchain because of its complexity, which, in turn, means that many do not fully understand the risks associated with the use of the system, and do not fully use the available functionality.

Network size

The work of the blockchain requires at least several hundred, or even better, several thousand nodes that work in concert. It is because of this that the system is extremely vulnerable to attacks in the initial stages of work. For example, if a user can gain control over 51% of the system nodes, he will be able to fully control the result of work. And if there are only 20 nodes in the system, then such a scenario is more than possible.

Network speed and efficiency

The blockchain structure is also one of the reasons why the normal functioning of the system can be disrupted. So, if the system becomes too widespread, and the blockchain's infrastructure is not ready for this volume of transactions, as a result, the speed of transactions may decrease, data storage problems may occur, and this will not affect the network efficiency in the best way [8].

Usage policy

Although it cannot be said that this item is directly related to the security of the blockchain, but the policy of the system may affect its application and further development. Considering that the currency in the blockchain system is international and decentralized, this, in essence, devalues the national currency controlled by the state. And, of course, at the moment the governing bodies of some states are seeking to impose more stringent restrictions on the use of the blockchain. Governments of different countries hope to bring the system under control before it becomes a serious competitor and threatens their economy. Indirectly, this is also a significant security threat to the blockchain, which can significantly slow down the spread of technology.

Third Party Systems

For example, NiceHash — a third-party market for Bitcoin mining — was recently cracked, as a result of which cryptocurrencies worth more than $ 60 million were stolen. As it turned out, this platform was unsafe. That is, it is not a security bug of the blockchain system itself. Rather, on the contrary, cybercriminals gained access to the NiceHash system using the blockchain.

For transactions in the blockchain system, public and private cryptographic keys are used.

By themselves, such keys are almost impossible to crack, but a cybercriminal can get them in a simpler and more familiar way. For example, keys can be obtained if you store them on an unsafe or weakly secured platform. So, if someone hacks your mailbox, he will be able to get access to all your letters, and therefore to the keys of your profile in the blockchain. In this case, the attacker will be able to seize your funds, posing as you. And this is one of the main issues concerning the security of the system.

Traditional fraudulent tricks

Users of the system can also fall for other, more traditional tricks scam. In fact, such fraudulent schemes are not considered a weak point in the blockchain security system. So, for example, you can receive an e-mail in which a stranger to you will convince you that it was you who became the lucky one who won something significant. Alternatively, fraudsters may offer you to spend your cryptocurrency on some product or service that you will never receive.

6. The main risks and threats to information security of technology blockchain

At the moment, the main threat to the blockchain, relatively hypothetical, is the "51% attack", when an attacker can roll back transactions by printing alternative blocks on a side chain (branch) and is guaranteed to refute what happens in the main chain of the blockchain. In fact, it looks like a shuttle run. However, taking into account the resource-intensiveness of the hash function solution and the emission of new bitcoins, so far this option seems unlikely. The collusion of the owners of the largest mining pools also looks unconvincing (if you do

not take into account the statistics of the largest producers of bitcoins). But there were already similar examples: one of the pools — ghash.io — gained power close to 50%, after which the owners stopped accepting new users in order not to create a compromising situation.

Consider a scenario where an attacker tries to generate an alternative chain faster than honest nodes. Even if it succeeds, he will not be able to make any changes in the system, for example, to create coins from the air or take coins that no one has transferred to him. Nodes will not accept an incorrect transaction as a payment, and honest nodes will never accept a block with an incorrect transaction. The attacker can only change one of his own transactions by returning to himself the payment he recently made. The race between the honest chain and the attacker's chain can be described in terms of a binomial random walk. A successful event is an increase in the honest chain by one block with an approach to the goal by +1, an unsuccessful event is an increase by one block in the attacker's chain with a decrease in the gap by –1. The possibilities of an attacker in a race under restrictions are similar to the description of the Gambler's Ruin problem (the ruin of a gambler). And so, a gambler with unlimited credit starts the game under conditions of restriction and can potentially hold an unlimited number of games to try to achieve break-even. We can calculate the likelihood of them achieving a break-even point or the same thing that an attacker will overtake honest chain builders [9].

Let: p – the likelihood that an honest host will find the next block; q – the likelihood that the attacker will find the next block; qz – the likelihood that an attacker will win the race if he falls behind by z blocks.

Then:

$$q_z = \begin{cases} 1 & \text{if } p \le q \\ (q/p)^z & \text{if } p > q \end{cases}$$

Suppose that $p>q$, then the probability decreases exponentially with an increase in the number of blocks that the attacker is behind. Thus, if he fails to get ahead at the very beginning, then his chances of winning in the future will become vanishingly small. Consider now how long the payee must wait to be sure that the sender will not be able to change the transaction. Suppose the sender is an attacker who wants the recipient to believe that the payment has been made, but after a while to return the payment to himself. The recipient will be notified when this happens, but the sender hopes that it will be too late. The recipient generates a new key pair and gives the public key to the sender shortly after signing it. This does not allow the sender to prepare a block chain in advance working on it ahead of time to complete the transaction at the moment. Only when a transaction is sent, can a dishonest sender begin to work in secret on a parallel chain containing an alternative version of this transaction. The recipient waits until the transaction is added to the block and the Z blocks are added after that. He does not know at what stage of construction the attacker is, but assuming that honest blocks were built with the same average time per block, the expected value of the attacker's gain can be found through the Poisson distribution[10]:

$$\lambda = z\frac{q}{p}$$

To get the probability with which the attacker can still come forward, multiply the Poisson distribution of each value of the attacker's progress by the probability that he will come forward from this point:

$$\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \cdot \begin{cases} (q/p)^{(z-k)} & \text{if } k \le z \\ 1 & \text{if } k > z \end{cases}$$

Or after regrouping:

$$1 - \sum_{k=0}^{z} \frac{\lambda^k e^{-\lambda}}{k!} \left(1 - (q/p)^{(z-k)}\right)$$

By analyzing the resulting expression numerically, one can easily verify that the probability decreases exponentially with increasing z.

Threat 2. Control Package

Different products of the blockchain technology use different methods of block confirmation. For example, in Bitcoin, the proof-of-work method is used — block confirmation by computing power. Another option for closing blocks is proof-of-stake, when blocks are printed not with computing power, but with the help of money held by people in their hands. In this case, in order to conduct a "51% attack", you must have 51% of the coins wrapped around the system. As in the case of the "shuttle run", if the attacker owns more than 51% of the coins, he will also be able to create an alternative chain, which will become the main one. This situation is reminiscent of a vote at a shareholders meeting, when one of the owners has a controlling stake in his hands, blocking the votes of other holders[11].

Threat 3. Key to all doors

If the security of the blockchain causes a minimum of concern, then the safety of Bitcoins, on the contrary, raises many questions, because, like ordinary paper money, cryptocurrency can also be stolen. The key of the blockchain entry is a hash function of the public key. Uncertain or negligent storage of a private key can lead to theft or loss of bitcoins. According to the Harvard Business Review, the cost of lost bitcoins is already about $ 950 million. The easiest way to protect yourself is to create a wallet password. But if a hacker kidnaps both your wallet and your password, it will be almost impossible to recover the stolen Bitcoins, since the transactions represented with the stolen keys seem to be checking nodes indistinguishable from legitimate

transactions. Some skeptics claim that hackers will be able to crack the key using services that calculate passwords by hash. However, given the current computing power, this seems unlikely. But if suddenly an algorithm appears that allows for the effective factorization of elliptic curves, then there will be a possibility that it will be easy to find private keys to the wallet addresses from which the money was spent.

Threat 4. Exchange attacks

The reliability of cryptocurrency exchanges raises no less questions. In August 2016, 119,756 bitcoins (about $ 65 million) were stolen from the Hong Kong Stock Exchange Bitfinex, one of the four largest cryptocurrency trading sites in the world. Bitfinex has a reputation as one of the most reliable and secure organizations: most user funds were stored in multi-signature wallets and in "cold stores". Despite this, the attackers managed to bypass Bitgo protection, including two-factor authentication and a multi-signature mechanism, and to commit mass theft from individual users' wallets. The details of the hacking were never conveyed to the general public, but the media replied that the Bitfinex employees might be involved in hacking, which again raises the question of the human factor.

Conclusions

Can we now, having considered all these points, say with confidence that the blockchain is a secure and safe system?

Rather, it is worth concluding that the system will function properly if used correctly and accurately. It is also worth bearing in mind that its security depends on the presence of a sufficient number of users, and many security gaps appear trite due to the human factor. Therefore, we should not forget that any system has weak spots, and the blockchain is not an exception to this rule[12].

References

[1] http://www.blockchain4innovation.it/wpcontent/uploads/sites/4/2017/05/Blockchain-

[2] https://www.coindesk.com/information/who-created-ethereum

[3] https://www.coindesk.com/information/how-ethereum-works

[4] A Survey of blockchain security issue and challenges(Iuon-Chang Lin1,2 and Tzu-Chun Liao2)[jan-12- 2017]

[5] Public standares and patients controll:how to keep electronic medical records accessible but private(Kenneth D Mandl, Peter Szolovits, Issac S Kohane)[3 february 2001]]

[6] https://blockgeeks.com/guides/smart-contracts/

[7] *A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman,* Medrec: Using blockchain for medical data access and permission management, in 2016 2nd International Conference on Open and Big Data (OBD), Aug 2016, pp. 2530.

[8] *G. Zyskind, O. Nathan, and A. Pentland.* Decentralizing privacy:Using blockchain to protect personal data, in Security and Privacy Workshops (SPW), 2015 IEEE, May 2015

[9] https://www.researchgate.net/publication/319058582 Blockchain Challenges and Opportunities A Survey

[10] http://www.meti.go.jp/english/press/2016/pdf/053101f.pdf

[11] https://www.dotmagazine.online/issues/innovation-in-digital-commerce/what-can-blockchain-do/securityand-privacy-in-blockchain-environments

[12] https://www.business2community.com/tech-gadgets/issues-blockchain-security-02003488

The article was delivered to your editory stuff on the 20.05.2019

УДК 004.8

Г.В. Марчук[1], В.Л. Левківський[2], С.С. Каліберда[3]

[1]Державний університет «Житомирська політехніка»,
м. Житомир, Україна, mgv.555.mgv@gmail.com

[2]Державний університет «Житомирська політехніка»,
м. Житомир, Україна, levkivskyy@ztu.edu.ua

[3]Державний університет «Житомирська політехніка»,
м. Житомир, Україна, kassergey@gmail.com

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

У статті досліджено роботу методів інтелектуального аналізу даних, таких як лінійна і поліноміальна регресія та метод опорних векторів. Успіх застосування заснований на тому, що методи і технології Data mining забезпечують дослідження даних і виявлення в них прихованих закономірностей різних видів. Аналіз допомагає виявити різні ознаки і параметри даних, і тому є сильним інструментом на етапі формування моделей прогнозування.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ЛІНІЙНА РЕГРЕСІЯ, ПОЛІНОМІНАЛЬНА РЕГРЕСІЯ, МЕТОД ОПОРНИХ ВЕКТОРІВ, ХРОНІЧНЕ ЗАХВОРЮВАННЯ

В статье исследована работа методов интеллектуального анализа данных, таких как линейная и полиномиальная регрессия и метод опорных векторов. Успех применения основан на том, что методы и технологии Data mining обеспечивают исследования данных и выявления в них скрытых закономерностей различных видов. Анализ помогает выявить различные признаки и параметры данных, и поэтому является сильным инструментом на этапе формирования моделей прогнозирования.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ, ЛИНЕЙНАЯ РЕГРЕССИЯ, ПОЛИНОМИНАЛЬНОЙ РЕГРЕССИЯ, МЕТОД ОПОРНЫХ ВЕКТОРОВ, ХРОНИЧЕСКИЕ ЗАБОЛЕВАНИЯ

The main research of the article is the data mining methods, such as linear and polynomial regression and the support vector machine. The application success is based on the fact that the methods and technologies of Data mining ensure the study of data and the research of hidden patterns in them. The analysis assists in identification of various features and data parameters, and therefore it is a powerful tool in the stage of forming forecasting models.

INTELLECTUAL ANALYSIS OF DATA, LINEAR REGRESSION, POLYNOMYNAMIC REGRESSION, SUPPORT VECTOR MACHINE, CHRONIC DISEASES

Вступ

Інтелектуальний аналіз даних це процес визначення нових, коректних і потенційно корисних знань на основі даних, які представлені великими об'ємами. Крім того, аналіз включає в себе безліч різних підходів і методів для дослідження і перетворення даних.

Основна мета інтелектуального аналізу даних полягає в тому, щоб створити модель, що дозволяє ефективно інтерпретувати і використовувати ті дані, якими володіє дослідник на даний час, і ті дані, які отримає в майбутньому. Оскільки аналіз даних включає в себе безліч методів, то основний етап створення моделі даних – це вибір методу аналізу, що буде використаний в цій моделі. Для правильного вибору методу потрібен практичний досвід. Далі модель потрібно доопрацювати, щоб зробити її більш ефективною.

1. Постановка проблеми

Хронічне захворювання – це стан здоров'я або хвороба людини, яка є стійкою або іншою довготривалою в її наслідках або хворобою, що настає з часом. Термін хронічний часто застосовується, коли перебіг захворювання триває більше трьох місяців. Часті хронічні захворювання включають артрит, астму, рак, діабет, вірусні захворювання, такі як гепатит С та ВІЛ / СНІД тощо[1].

Багато людей з хронічними захворюваннями можуть навіть не підозрювати, що вони хворіють, симптоми хвороби можуть бути не помітні. Часто це є причиною відсутності розуміння та підтримки лікарів, родичів, друзів та колег.

Частка людей які в Україні мають хронічні захворювання станом на 2017 рік складала 37,75%, а у 2018 році – 38,94%. Найбільш поширеними були гіпертонія та серцеві захворювання – про наявність однієї з цих хвороб повідомили відповідно 42,8% і 25,1% осіб (2017р.) та 46% і 26,1% осіб (2018р.), які мають хронічні захворювання [2, 3].

У країнах Європейського Союзу більше третини осіб віком 16 років і старше мали хронічні захворювання або проблеми зі здоров'ям. Найбільша кількість таких осіб у Фінляндії – 47%, Естонії – 44%, по 42% – у Німеччині та Португалії, а найменша в Італії – 15%, Румунії – 19% та Болгарії – 21%. В Україні повідомили, що мають хронічні захворювання або проблеми зі здоров'ям 44% осіб цієї вікової категорії [2, 3].

Особливе занепокоєння викликає проблема передчасної смертності чоловіків, тривалість життя яких за 1991—2010 роки зросла лише на один рік, тоді як у країнах, які приєдналися до ЄС у 2004—2007 роках, за аналогічний період тривалість життя чоловіків зросла на чотири роки. У 2009 році в Україні тривалість життя жінок була на вісім років коротша, ніж в середньому у країнах ЄС, а чоловіків — на 12 років.

Соціально-економічними наслідками передчасної смертності є не лише зменшення років потенційного життя та збільшення величини безповоротних втрат унаслідок смерті, а і значні економічні збитки. Через передчасну смертність населенням України лише щороку втрачається близько 4 млн. років потенційного життя, відповідно обсяг недовиробленого національного продукту становив від 47,9 до 89,1 млрд. гривень, причому левова частка втрат була зумовлена смертністю чоловіків.

Таким чином можна зробити висновки що в Україні хронічні захворювання мають великий вплив на здоров'я людей та на економіку. Ця проблема потребує ретельного вивчення та проведення досліджень, які в подальшому можуть допомогти при проведенні медичної реформи та прийняті рішень про превентивні міри.

2. Опис джерела даних

На жаль поки що медична статистика в Україні не може надати необхідного об'єму статистичної інформації. Для проведення наукових досліджень і подальшого впровадження результатів в Українську медичну сферу будемо використовувати статистичні дані по хронічним хворобам 500 міст США. Одним з джерел цієї інформації є американська організація Centers for Disease Control and Prevention (скорочено CDC) – Центри з контролю та профілактики захворювань.

Ці дані самі по собі унікальні, тому що вони охоплюють 103 млн. осіб в віці від 18 років, мають в своєму складі 27211 тисяч записів по різним територіям статистичної звітності, населення котрих складає від 50 чоловік до 26980 чоловік. Також серед даних є код штату, округу, міста, географічні координати, що дозволяє в подальшому розширити аналіз на основі інших статистичних даних як середній дохід домогосподарства, рівень безробіття та інше.

Показники поділяються на три основні групи:

– нездоровий спосіб життя (5 показників);

– хронічні захворювання (13 показників);

– охоплення населення превентивними методами (9 показників).

Усі показники представлені у співвідношенні відсотку населення та діапазону похибки. Дані представлені у форматі csv файлу.

Таким чином можна зробити висновки, що обрані статистичні дані підходять для проведення різних наукових досліджень.

3. Результати досліджень

Відомо багато експертних систем для постановки медичних діагнозів, але технології інтелектуального аналізу даних дають можливість виявляти закономірності отримання різних захворювань, створювати умови найефективнішого лікування, передбачати результати призначеного курсу лікування тощо.

Знання, що добуваються методами Data mining, прийнято представляти у вигляді закономірностей. Серед таких виступають: асоціативні правила, дерева рішень, кластери, алгоритми машинного навчання, тощо.

Одним з алгоритмів машинного навчання для класифікації та прогнозування є наївний Баєсів класифікатор. Слід зауважити, що алгоритм Баєса найчастіше використовується для класифікації, але в нашому випадку він буде використовуватися для прогнозування можливого значення індикатора в залежності від іншого індикатора. Показники в тестових даних представлені у відсотках від 0 до 100%. Кожен інтервал відсотків буде представляти окремий клас. Наприклад якщо прогнозований показник розбити на 25 класів, то виходить що значення в діапазонах [0%, 4%), [4%, 8%), [12%, 16%) ... [92%, 96%), [96%, 100%] будуть представляти окремі класи. Розрахувати діапазон відсотків можна шляхом ділення числа 100 на кількість класів (наприклад 100%/25 класів = 4% довжини діапазону на один клас). У випадку коли треба точність прогнозування в 1%, можна проводити класифікацію даних для 100 класів (100%/100 класів = 1% довжини діапазону на один клас). Таким чином буде створена модель, котра буде прогнозувати значення одного показника в залежності від іншого (інших).

Недоліком описаного вище підходу може бути випадок, коли дані розташовані в невеликому діапазоні. Наприклад коли є 25 класів та 80% деяких показників розташовано в невеликому діапазоні, наприклад в [48%, 52%) та [52%, 56%), то підхід може повертати в деяких випадках точний прогноз, але відносна точність буде недостатньою.

Для дослідження даних використовувалися наступні бібліотеки pandas, scikit, numpy, matplotlib.

Найпростішими методами аналізу даних є візуалізація та лінійна регресія. Слід зауважити, що основні дані зберігаються в стовпчиках з назвою «*CrudePrev», де * – будь які символи. Також серед даних наявні комірки з незаповненими значенням. Для зручності роботи ці комірки були заповнені

нулями. Сирцевий код для імпорту даних для навчання представлено нижче.

```
import pandas as pd
data = pd.read_csv("../data.csv")
# print('headers: ', data.columns.values)
def data_only_crude():
   data_only_crude_prev=data.loc[:,data.
   columns.str.contains('CrudePrev')]
   return data_only_crude_prev.fillna(0)
data_only_crude_prev = data_only_crude()
```

Візуальний аналіз на всій вибірці проводити було складно, тому що наявно 28 стовпців даних і при попарному порівнянні виходить 784 діаграми, в кожні з яких по 27210 точок. Тому було вирішено для початку використати лінійну регресію для визначення кореляції показників.

В таблиці 1 представлено деякі показники кореляції, а в таблиці 2 наведено опис цих показників.

Таблиця 1

Кореляція показників отримана шляхом використання лінійної регресії та оцінки результатів лінійної регресії коефіцієнт детермінації (R-квадрат)

	ARTHRITIS_CrudePrev, %	BINGE_CrudePrev, %	BPHIGH_CrudePrev, %	BPMED_CrudePrev, %	CHD_CrudePrev, %
ARTHRITIS_CrudePrev	0,00	34,19	74,76	64,95	80,75
BINGE_CrudePrev	34,10	0,00	50,12	22,15	45,42
BPHIGH_CrudePrev	74,85	50,29	0,00%	57,12	73,26
BPMED_CrudePrev	65,03	22,21	57,10	0,00	50,89
CASTHMA_CrudePrev	24,79	14,89	33,43	3,61	20,42
CHD_CrudePrev	79,42	44,71	71,98	49,98	0,00

Таблиця 2

Опис стовпців в джерелі даних

Назва поля	Опис
ARTHRITIS_CrudePrev	Модельна оцінка поширеності захворювання артритом серед дорослих у віці від 18 років, 2016 р.
BINGE_CrudePrev	Модельна оцінка поширеності споживання алкоголю серед дорослих у віці до 18 років, 2016 р.
BPHIGH_CrudePrev	Модельна оцінка поширеності захворювання високим кров'яним тиском серед дорослих у віці 18 років, 2015 р.
BPMED_CrudePrev	Модельна оцінка поширеності прийому ліків для контролю високого кров'яного тиску серед дорослих у віці від 18 років, 2015 р.
CASTHMA_CrudePrev	Модельна оцінка поширеності захворювання астмою серед дорослих у віці від 18 років 2016 р.
CHD_CrudePrev	Модельна оцінка поширеності захворювання ішемічною хворобою серця серед дорослих у віці від 18 років 2016 р.

Більшість відносин були логічними при використані лінійної регресії, але і на її основі можна побачити наступні цікаві залежності між показниками:

Поширеність втрати всіх зубів після 65 років на дільницях, в яких високий показник курців. Але при подальшому аналізі було виявлено що курці рідше ходять до стоматологів (рис. 1).

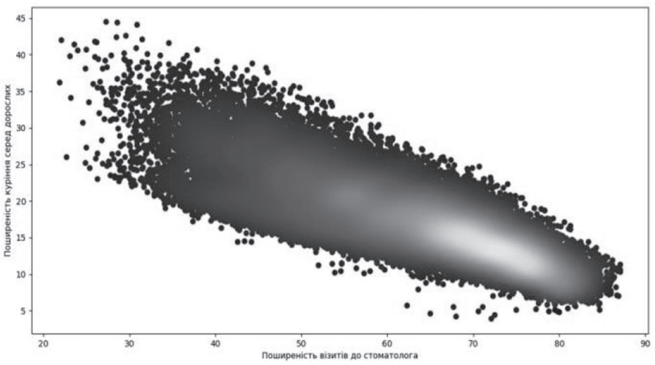

Рис. 1. Залежність візитів до стоматолога від кількості курців на дільниці

Наявність залежності між низькою фізичною активністю серед населення та втратою зубів після 65 років (рис. 2). При цьому також на дільницях, де люди не займаються фізичною активністю, менша статистика візитів до стоматолога.

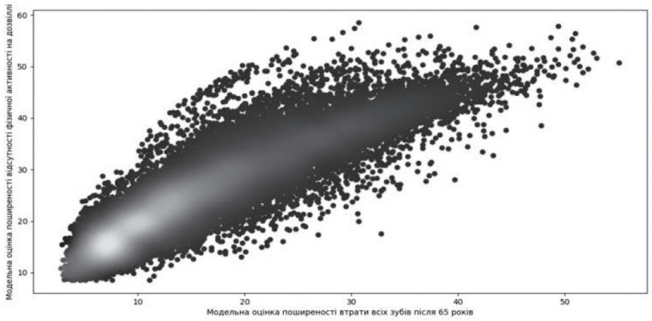

Рис. 2. Залежність між втратою зубів та фізичною активністю

Залежність між поширеністю хвороби нирок та інсультом (коефіцієнт кореляції 85,9%).

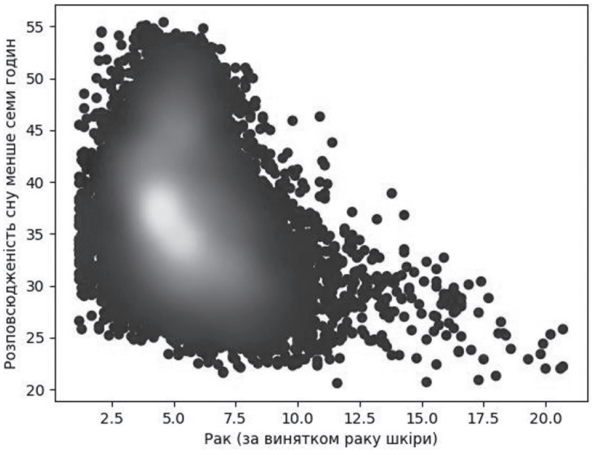

Рис. 3. Залежність між захворюванням на рак та сном менше 7 годин на добу

Серед показників було визначено, що на хронічні захворювання майже ніяк не впливають такі

показники як сон менше 7 годин на добу, різні види профілактики після 65 років (можливо цей показник треба аналізувати тільки на віковій категорії 65 років і старше), тест Папаніколау (тест на визначення раку матки, можливо цей показник має вплив тільки в категорії жінок).

В деяких випадках метод опорних векторів зміг визначити залежності, які слабше визначалися лінійною регресією (світлий колір значить більшу щільність точок, темний меншу), найбільш цікавими були:

Зворотна залежність між поширеністю захворювань на рак та поширеністю сну менше 7 годин на добу. Точність прогнозування 63%. Допустиме відхилення прогнозування 1% (рис.3).

Зворотна залежність між поширеністю захворювання на рак та поширеністю курців на дільниці.

Залежність між населенням які роблять тест на холестерин та кількістю хворих на рак (можливо це пов'язано з віковою категорією, яка мешкає на дільниці). Точність прогнозування 80% (рис.4).

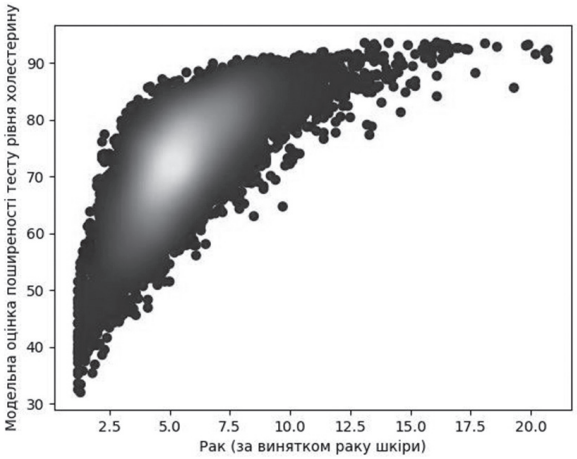

Рис. 4. Залежність між населенням які роблять тест на холестерин та кількістю хворих на рак

Більш сильна залежність між розповсюдженістю захворювань на рак та прийомом ліків для контролю тиску (рис.5) в порівнянні з розповсюдженістю захворювань на рак та високим тиском (рис.6).

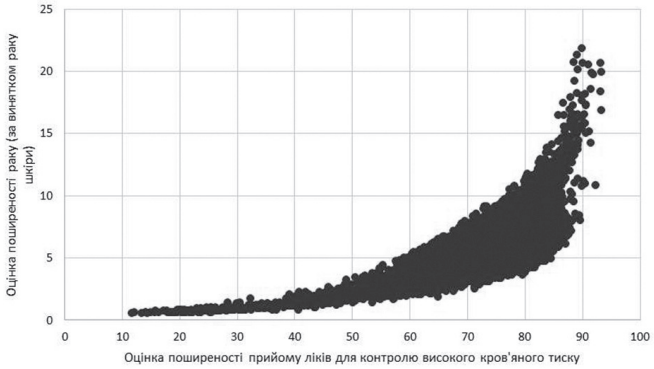

Рис. 5. Залежність між захворюваннями на рак та прийомом ліків для контролю тиску

Після повторного проходження навчання моделі на основі комбінації найбільш впливових показників, були покращені результати прогнозування і виявлені наступні залежності:

Поширеність артриту та розладів стану психічного здоров'я корелює з легеневою обструктивною хворобою на 91%, хоча по одинці ці параметри корелюють відповідно на 64.9% та 53.6%.

Оцінка поширеності аналізу крові, сигмоїдоскопії або колоноскопії та оцінка поширеності хронічної обструктивної легеневої хвороби корелюють з поширеністю кількості курців на 80,6%, хоча по одинці ці параметри корелюють відповідно на 28% та 66,5%.

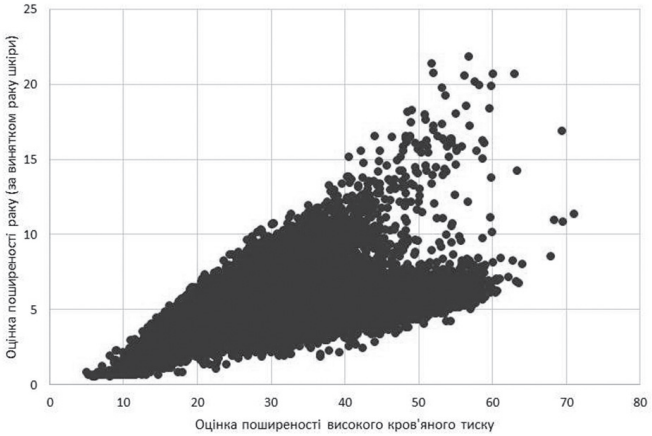

Рис. 6. Залежність між захворюваннями на рак та високим кров'яним тиском

Оцінка поширеності раку (за винятком раку шкіри) та оцінка поширеності відсутності фізичної активності і дозвілля корелюють з поширеністю високого рівня холестерину на 79%, хоча по одинці ці параметри корелюють відповідно на 35% та 30,8%.

Серед моделей були обрані моделі, які давали точність по методу найменших квадратів більше 70%. Деякі варіанти результатів прогнозування різних показників наведені у таблиці 1.

Таблиця 3

Результати прогнозування різних показників

PredictBy	PredictTo	Linear-Regression, %	Polynomial Regression, %	SVM, %
артрит	ішемічна хвороба	80,92	81,37	80,72
споживання алкоголю	артрит	74,90	75,64	75,12
споживання алкоголю	цукровий діабет	80,04	80,23	78,87
споживання алкоголю	інсульт	82,83	84,93	84,13

Продовження табл. 3

PredictBy	PredictTo	Linear-Regression, %	Polynomial Regression, %	SVM, %
ішемічна хвороба	артрит	80,92	82,08	81,65
ішемічна хвороба	хронічна обструктивна легенева хвороба	77,09	77,33	78,60
ішемічна хвороба	високий рівень холестерину	77,32	82,68	85,74
ішемічна хвороба	хвороба нирок	81,00	81,13	80,65
хронічна обструктивна легенева хвороба	ішемічна хвороба	77,10	77,42	76,79
хронічна обструктивна легенева хвороба	інсульт	83,19	83,39	82,74
цукровий діабет	хвороба нирок	91,37	91,37	90,92
цукровий діабет	інсульт	88,84	90,04	89,16
хронічна хвороба нирок	ішемічна хвороба	81,00	81,09	81,02
хронічна хвороба нирок	цукровий діабет	91,37	91,41	91,23
хвороба нирок	інсульт	93,23	94,03	93,23
інсульт	цукровий діабет	88,84	89,86	89,36
інсульт	хвороба нирок	93,23	94,24	94,24

Для тестування використовувалися записи з навчальної вибірки. Дані було розбито на дві вибірки: навчальна(80%) та тестова (20%).

В бібліотеці scikit майже у кожної моделі є метод score і для кожної моделі він використовує різні підходи. В даному випадку для порівняння результатів використання різних моделей було розроблено окремий метод. Метод прогнозування повертає відношення кількісті правильно прогнозованих значень з урахуванням похибки до усіх значень. Нижче наведено приклад програмного коду.

```
from math import isclose
def score_prediction(model, test, column_to_predict, column_by_predict, accuracy=0.0, y_column_preparer= None):
  count_of_test_rows = len(test)
  count_of_close_predicted = 0
  test_by = test[column_by_predict].values
  test_to = None
  if y_column_preparer is None:
    test_to = test[column_to_predict].values
  else:
    test_to = y_column_preparer(test[column_to_predict]).values
  predicted_values = model.predict(test_by)
  for i in range(0, count_of_test_rows):
    if isclose(predicted_values[i], test_to[i], abs_tol=accuracy/2):
      count_of_close_predicted = count_of_close_predicted + 1
  return count_of_close_predicted/count_of_test_rows
```

Метод на вхід отримує модель, тестові дані, назву колонки для прогнозування, назви колонок на основі яких відбувається прогнозування, бажана точність прогнозування (в абсолютних величинах), метод для підготовки тестових даних.

На основі тестових даних можна зробити висновки, що найбільш точним методом прогнозування є метод опорних векторів. Але слід зауважити що складність методу опорних векторів складає $O(n^2)$, де n – кількість записів в вибірці, що може бути недоліком у випадку наявності великої кількості навчальної вибірки.

Менш якісні результати демонструє метод поліноміальної регресії, але цей вид регресії здатен знаходити залежності в даних, які не описуються лінійним рівнянням. При цьому використання цього методу вимагає попередньої обробки даних, на основі яких буде відбуватися прогнозування.

Найменш точні результати демонструє метод лінійної регресії, але цей метод є найбільш простим та швидким, тому використання лінійної регресії для розвідувального аналізу можна вважати обґрунтованим.

Висновки

Стосовно використаних алгоритмів можна зробити наступні висновки: для більшості випадків прогнозування зв'язків між величинами достатньо більш простих моделей, наприклад лінійної регресії, яка дозволила швидко провести розвідувальний

аналіз. Серед використаних алгоритмів, найбільш ефективним виявився метод опорних векторів, але цей метод також вимагає витрат часу на проведення аналізу. Альтернативним методом виявився метод поліноміальної регресії.

Завдяки використанню методів інтелектуального аналізу даних можна поясняти деякі закономірності. Наприклад при аналізі даних було виявлено, що на дільницях де мала фізична активність та розповсюджено куріння більш високий показник втрати всіх зубів після 65 років та менша статистика візитів до стоматолога. Таким чином можна зробити висновки, що основною причиною втрати зубів було не відсутність фізичної активності та куріння, а скоріше нерегулярні візити до стоматолога.

Також не виявлено сильної залежності між розповсюдженістю хронічних захворювань та сном менше ніж 7 годин на добу, але слід задуматися над залежністю між захворюванням на рак та сном менше 7 годин на добу.

Література:

[1] Use Your Words Carefully: What Is a Chronic Disease? [Електронний ресурс] // Front. Public Health. – 2016. – Режим доступу до ресурсу: https://www.frontiersin.org/articles/10.3389/fpubh.2016.00159/full

[2] Самооцінка населенням стану здоров'я та рівня доступності окремих видів медичної допомоги у 2017 році. // Державна служба статистики України. – 2018. – С. 7.

[3] Самооцінка населенням стану здоров'я та рівня доступності окремих видів медичної допомоги у 2018 році. // Державна служба статистики України. – 2019. – С. 7.

[4] 500 Cities: Local Data for Better Health [Електронний ресурс] // Centers for disease and control prevention. – 2018. – Режим доступу до ресурсу: https://www.cdc.gov/500cities/about.htm

Поступила до редколегії 06.06.2019

УДК 004.75

С.Г. Удовенко[1], Л.Е. Чала [2], Є.С. Кушвід[2]
[1]ХНЕУ ім. С. Кузнеця, м. Харків, Україна, serhiy.udovenko@hneu.net
[2]ХНУРЕ, м. Харків, Україна, larysa.chala@nure.ua

МЕТОД ПОРІВНЯННЯ ТЕКСТОВО-ГРАФІЧНИХ ФРАГМЕНТІВ В ЕЛЕКТРОННИХ ДОКУМЕНТАХ ЗА ГІБРИДНИМ КРИТЕРІЄМ

Розглянуто метод порівняння текстово-графічних фрагментів в електронних документах за гібридним критерієм. Цей метод дозволяє визначати інтегроване значення подібності між запитом, пов'язаним з зображенням в запиті, та текстово-графічним зображенням в базі даних документів. Критерій, що застосовується, передбачає використання вагових коефіцієнтів для зображень з анотаціями та для зображень без анотацій. Визначено перспективи використання запропонованого методу.

КОМБІНОВАНИЙ КРИТЕРІЙ, ПОРІВНЯННЯ ТЕКСТОВО-ГРАФІЧНИХ ФРАГМЕНТІВ, ХЕШ-МЕТОД ОБРОБКИ ЗОБРАЖЕНЬ, ГІБРИДНИЙ ЗАПИТ

Удовенко С.Г., Чалая Л.Э., Кушвид Е.С. Метод сравнения текстово-графических фрагментов в электронных документах по гибридному критерию. Рассмотрен метод сравнения текстово-графических фрагментов в электронных документах по гибридному критерию. Этот метод позволяет определять интегрированное значение сходства между запросом, связанным с изображением в запросе, и текстово-графическим изображением в базе данных документов. Применяемый критерий предусматривает использование весовых коэффициентов для изображений с аннотациями и для изображений без аннотаций. Определены перспективы использования предложенного метода.

КОМБИНИРОВАННЫЙ КРИТЕРИЙ, СРАВНЕНИЕ ТЕКСТОВО-ГРАФИЧЕСКИХ ФРАГМЕНТОВ, ХЭШ-МЕТОД, ОБРАБОТКИ ИЗОБРАЖЕНИЙ, ГИБРИДНЫЙ ЗАПРОС

S.G. Udovenko, L.E. Chala, Ye.S. Kushvid. A method for comparing text and graphic fragments in electronic documents using a hybrid criterion. A method for comparing text and graphic fragments in electronic documents by a hybrid criterion is considered. This method allows you to determine the integrated similarity value between the request associated with the image in the request and the text-graphic image in the document database. The criterion used provides for the use of weights for images with annotations and for images without annotations. The prospects for using the proposed method are determined.

COMBINED CRITERION, COMPARISON OF TEXT-GRAPHIC FRAGMENTS, HASH METHOD OF PROCESSING IMAGES, HYBRID REQUEST

Вступ

Останнім часом отримали розповсюдження цифрові технології, що сприяють стрімкому збільшенню кількості інформації, яка зберігається в базах даних різного тематичного призначення. Для швидкого доступу до текстово-графічних електронних документів (ТГ-документів) виникає потреба у розробці інформаційних систем з ефективним керуванням цими базами зображень та можливістю оперативного пошуку зображень за запитами користувачів. Відзначимо, що такі системи об'єднує необхідність аналізу подібності зображень для подальшого анотування або класифікації [1, 2]. На сьогодні є актуальною проблема побудови мультимодальної системи пошуку та порівняння текстово-графічних (ТГ) фрагментів електронних документів, що містять зображення та текст. Важливими етапами роботи такої системи є індексація зображень та формування запитів. Індексація зображень – це операція, яка полягає у вилученні текстового підпису зображення, що описує його семантичний вміст, для можливості ефективного пошуку у базі даних. Формування запитів – операція, що дозволяє представляти інтереси користувача. Реалізація цих етапів потребує дослідження процесів пошуку і порівняння ТГ-фрагментів електронних документів за характеристиками зображення та текстовими підписами. На рис. 1 наведено загальну схему порівняння ТГ-фрагментів за комбінованими запитами.

Рис. 1. Схема порівняння ТГ-фрагментів за комбінованими запитами

У даній роботі досліджуються основні етапи схеми порівняння текстово-графічних фрагментів електронних документів, а також пропонується метод такого порівняння за комбінованим мультимодальним критерієм.

1. Характеристика основних етапів схеми порівняння текстово-графічних фрагментів

Розглянемо деякі аспекти реалізації схеми, що наведено на рис. 1. Відзначимо, що база ТГ-документів цієї схеми корегується в режимі реального часу. Обсяг зображень цієї бази не є фіксованим, адже інформація про зображення постійно розвивається, в залежності від нових зовнішніх подій, які можуть виникнути в будь-який час. Зазначимо припущення, які є важливими для використання та розвитку системи порівняння ТГ-фрагментів на етапі індексації зображень:

– екземпляри зображень бази не завжди містять повну текстово-графічну інформацію;

– загальна кількість зображень не є фіксованою і обробка нових зображень здійснюється в режимі реального часу;

– база знань системи ґрунтується на анотаціях зображень, доповнених текстом, а також на їх візуальних характеристиках (колір, текстура, форма. просторова диспозиція);

– навчання є інтерактивним і може бути виконане шляхом використання методів навчання з підкріпленням. Взаємодія між користувачами та експертами системи має здійснюватися у простий спосіб для фільтрації не релевантних запитам зображень під час пошуку;

– навчання здійснюється на малих вибірках навчальних даних (до 20 екземплярів).

– система має формувати (аналізувати) анотації зображень в режимі реального часу.

З оглядом на ці гіпотези, в кожний поточний момент можуть розглядатися три типи зображень в пропонованій системі: зображення без анотації; зображення з додатковою інформацією (наприклад, додатковий опис або довідкові дані); зображення з автоматично доданою інформацією (з «розширеними» анотаціями).

Можливість еволюції бази даних зображень і відповідної бази знань визначила такі сценарії витягу ТГ-фрагментів та формування запитів:

– сценарій 1: зображення без анотації використовується як запит, спрямований на пошук близьких зображень з бази даних;

– сценарій 2: набір слів в текстово-графічних документах використовується як повноцінний запит;

– сценарій 3: зображення та слова в текстово-графічних документах використовуються як повноцінний запит.

Проблема застосування двох останніх сценаріїв виникає, коли частина бази даних зображень не анотована, що робить цю частину недоступною через текстові запити. У цьому випадку необхідно використовувати зв'язок між текстовими словами та візуальними характеристиками для мультимодального пошуку та анотування зображень, які містять текстову частину.

В значній мірі швидкодія методів порівняння ТГ-фрагментів залежить від тривалості попередньої обробки, що дозволяє представити анотовані зображення в придатному для швидкого порівняння вигляді. На етапі попередньої обробки цих зображень доцільно використовувати хеш-методи в комбінації з деякими більш точними методами [3, 4].

Метод мультимодального визначення подібності ТГ-фрагментів має поєднувати візуальний пошук зразків зображень та текстовий пошук за ключовими словами. Візуальна подібність між запитом, пов'язаним з зображенням в запиті, та текстово-графічним зображенням в базі даних оцінюється за величиною скалярного добутку відповідних векторів з використанням комбінованого критерія.

2. Попередня обробка зображень в ТГ-документах з використанням комбінованого хеш-методу

Визначальні ознаки зображень, які використовуються для попередньої обробки зображень в ТГ-документах, можуть бути розбиті на кілька груп [5]. До першої групи таких ознак віднесемо набір гістограм, для побудови яких використовуються палітри HSV і RGB. При формуванні гістограм для RGB-палітри колірний простір розбивається на 32 рівні частини і визначається кількість пікселів зображення в кожній з частин. Для HSV-палітри гістограми будуються окремо по кожній з компонент. Ознаки, отримані з різних гістограм, доповнюють одна одну, так як палітри HSV і RGB пов'язані нелінійним перетворенням і відповідають за різні властивості об'єктів. До другої групи можна віднести текстурні ознаки, засновані на підрахунку граничних пікселів. Для обчислення цих ознак зображення переводяться в чорно-білий формат, визначаються різкі характеристики яскравості переходи, а потім з верхньої і нижньої областей зображень зображення виділяються двадцятивідсоткові смуги пікселів. При пошуку зображень за зразком аналізуються пари зображень, які порівнюються. У дослідженнях, присвячених використанню текстурних ознак при порівнянні зображень, виділяються ознаки контрастності, грубості, спрямованості, лінійних образів, регулярності, однорідності і шорсткості текстур [6].

Перспективним є їх комбіноване використання з наступною розробленням відповідної сигнатури, придатної для ефективної реалізації алгоритму порівняння аналізованого і базового зображень.

Аналіз показав, що для створення таких сигнатур попередньої обробки зображень в ТГ-документах доцільно використовувати ідею хеш-методу в комбінаціях з деякими більш точними методами. До таких методів належать, перш за все, детектор Харріса (ДХ), метод LoG (Laplacian-of-Gaussian), SIFT-дескриптори (Scale Invariant Feature Transform), метод TI (TinyImages) і метод МППО (метод пошуку за зразком). Розглянемо деякі особливості цих методів.

Метод ДХ заснований на оцінці зміни інтенсивності світла з подальшим визначенням опорних точок зображень. До переваг методу відносяться інваріантність до поворотів та зрушення інтенсивності. У той же час він не є інваріантним до зміни масштабу зображення. Метод LoG дозволяє вирішити проблему порівняння зображень при зміні масштабу зображення. SIFT-дескриптор є ефективним засобом формування системи інваріантних структурних ознак. Він заснований на використанні сучасних базових принципів локальної обробки, що включають в комплексі локальну фільтрацію, формування значущих ознак, аналіз простору перетворень та апроксимацію координат ознак. У той же час широке застосування SIFT-дескрипторів обмежується їх обчислювальною складністю і високими вимогами до технічних засобів їх реалізації. До гібридних методів порівняння зображень відноситься метод TI. Тут спочатку генерується зменшена копія зображення з роздільною здатністю 32x32 пікселя. На першому кроці, незалежно від пропорцій, зображення стискується до розміру 20x20 пікселів і приводитися до сірого. У якості першої сигнатури використовується центральна частина зображення розміром 16x16 пікселів. В якості другої сигнатури використовуються дескриптори цікавих точок, координатами яких є екстремумами DoG-перетворення (Difference of Gaussian). Вводиться емпіричний поріг, нижче якого сигнатури вважаються близькими. Аналіз координат близьких сигнатур на парі зображень дозволяє виділити область їх перетину, для якої обчислюється попіксельна різниця і різниця карт градієнтів. Якщо такі різниці менше порогових, то пара зображень вважається близькою. Для порівняння зображень за зразком може бути використаний також метод МППО, в основі якого лежить представлення зображення у вигляді нечіткої колірної гістограми [7].

Відмінною рисою методу МППО є те, що гістограми будуються тут без використання процедури дефаззифікації. Для цього в кожній точці зображення визначається вектор 75-вимірного простору, компоненти якого є середнім арифметичним значень трьох функцій належності, після чого здійснюється усереднення даних векторів. Відстань між подібними векторами, що описують зображення, обчислюється на основі функції перетину гістограм.

Виділимо основні моменти, які необхідно враховувати при створенні комбінованого методу порівняння зображень:

– для підвищення швидкості обробки вихідне зображення доцільно зменшити;

– при комбінуванні хеш-методу з іншими методами доцільно використовувати відповідні вагові коефіцієнти;

– на етапі попередньої обробки має формуватися сигнатура, яка дозволяла б здійснювати оперативне визначення близьких об'єктів.

У таблиці 1 наведені результати порівняльного аналізу швидкості обробки зображень із застосуванням різних методів (SIFT, ДХ, LoG, TI, МППО, хеш-метод).

Таблиця 1

Оцінка середньої швидкості попередньої обробки одного зображення (300 px)

Метод попередньої обробки зображень	Середній час (сек) попередньої обробки одного зображення
SIFT	5.6
ДХ	3.1
LoG	4.9
TI	2.2
МППО	2.6
хеш-метод	0.2

Очевидно, що в якості базової процедури для вирішення поставленого завдання доцільно вибрати обчислювальну процедуру хеш-методу, що має суттєві переваги в швидкості обробки зображень в порівнянні з іншими методами.

Розглянемо особливості формування хешу зображення із застосуванням хеш-методу. Відповідно до теорії перетворень Фур'є зображення є двовимірним (залежність яскравості від горизонтальної та вертикальної координат) неперіодичним сигналом. Для RGB зображення необхідно розглядати яскравість в каналах Red, Green і Blue. Розглянемо декомпозицію зображення на фрагменти, що відповідають різним діапазонам частот:

– на низьких частотах будуть міститися найбільші деталі, що задаються загальним розподілом яскравості і кольору, і, отже, визначають форму об'єкта;

– на середніх частотах формується середня і дрібна деталізація, яка задає «локальний контраст» і для знятих крупним планом об'єктів є фактурою поверхні;

– на високих частотах формується наддрібна деталізація («мікроконтраст»), яка задає різкість зображення.

Очевидно, що для порівняння зображень необхідно, перш за все, використовувати низькі частоти. Розглянемо приклад роботи алгоритму попередньої обробки вхідного зображення, представленого на рис. 2 [5].

Рис. 2. Вхідне зображення

Для позбавлення від високих частот зменшимо вихідне зображення (відповідно до процедури хеш-методу). В даному прикладі зображення зменшується до розмірності 8x8. Зменшене зображення потім знебарвлюється (переводяться в градації сірого, що істотно скорочує розмір хешу). Зменшене і знебарвлене зображення представлено на рис. 3.

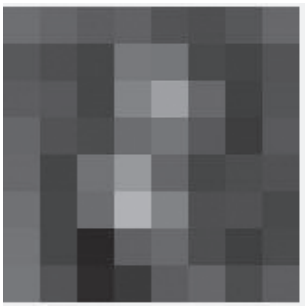

Рис. 3. Зменшене та знебарвлене зображення

Далі для кожного з кадрів зображення обчислюється середнє значення пікселів. Кожен піксель порівнюється із середнім значенням і, якщо він більше середнього значення, то в комірку хеша записується 1 (інакше 0). В результаті формується підсумкове хеш-зображення (рис. 4).

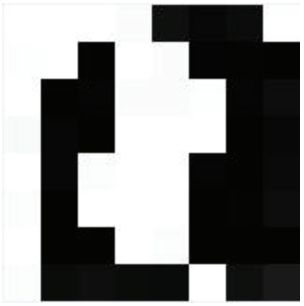

Рис. 4. Вихідне зображення

Для підвищення точності базової хеш-процедури, додамо вагові характеристики для чорно-білого рисунка 8 * 8. При цьому будемо вважати, що ступінь «значущості» точки на зменшеному зображенні збільшується в залежності від кількості сусідів іншого кольору. Таким чином, коефіцієнт змінюється від 0 до 8. Наведемо опис відповідної процедури:

Крок 1. На етапі первинної обробки вихідне зображення спочатку зменшується до розміру 10 * 10, після чого крайові значення обрізаються до розмірності 8 * 8 із загальним числом пікселів 64. Це дозволяє позбутися від рамок на зображенні та виділити його основну частину. Таким чином, хеш буде відповідати всім варіантам зображення, незалежно від розміру та співвідношення сторін.

Крок 2. Видаляється колір зображення. Зменшене зображення переводиться в градації сірого, що зменшує хеш з 64 пікселів (64 значення червоного, 64 зеленого і 64 синього) всього до 64 значень кольору.

Крок 3. Здійснюється приведення зображення до чорно-білих бітів. Для кожного з кадрів обчислюється середнє значення пікселів, а потім кожен піксель порівнюється із середнім значенням (якщо він більше середнього значення, то в клітинку хеша записується 1, інакше 0).

Крок 4. Проводиться побудова хеша: 64 окремих біта переводяться в одне 64-бітове значення (порядок не має значення, якщо він зберігається постійним). Підсумковий хеш не зміниться, якщо зображення стиснути або розтягнути. Зміна яскравості або контрасту, а також маніпуляції з квітами також істотно не впливають на підсумковий результат.

Крок 5. Для порівняння аналізованого зображення з базовими обчислюється відстань Хеммінга (підраховується кількість різних бітів) з урахуванням ваг. У табл. 2 наведені співвідношення між класичною (D) і модифікованою (M) відстанями Хеммінга для різних значень вагових коефіцієнтів W. При порівнянні аналізованого і базового зображень нульова відстань означає, що це однакові зображення (або варіації одного зображення). При відстані від 0 до 5 зображення в цілому досить близькі один до одного (неповні дублікати). При відстані від 6 до 9 зображення характеризуються окремими загальними ознаками, але дублікатами не є. Якщо відстань більше 9, то зображення вважаються різними.

Таблиця2

Корекція відстані між зображеннями з урахуванням вагових коефіцієнтів

W	0	1	2	3	4	5	6
D	1	1	1	1	1	1	1
M	0,8	0,85	0,9	0,95	1	1,05	1,1

3. Метод визначення подібності текстово-графічних фрагментів

Розглянемо спочатку векторну модель визначення подібності текстово-графічних фрагментів за текстовими підписами. Текстову частину документа D, що має порівнюватися з текстовою частиною запиту Q, представимо як зважений вектор термінів:

$$d_i = (w_{i,1}, w_{i,2}, ...w_{i,t}), \quad (1)$$

Для обчислення подібності між запитом Q і документом D визначається скалярний добуток між відповідними векторами:

$$S(d_i, q) = \sum_j w_{ij} \times w_{qj} S(d_i, q) = \sum_j w_{ij} \times w_{qj} . \quad (2)$$

Різні методи оцінки вагових коефіцієнтів дозволяють створити чимало функцій ранжирування для моделі векторного простору. Розглянемо підхід з використанням класичної моделі TF-IDF та техніки зважування і нормалізації. Класична векторна модель обчислює вагу терміна в документі як добуток терміну частоти (TF) і зворотної частоти документів (IDF):

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}} ; \quad idf_i = \log \frac{N}{|\{d : t_i \in d\}|}, \quad (3)$$

де $n_{i,j}$ – число входжень розглянутих термінів t_i в документ d_j, а знаменник – сума числа входжень усіх членів документа d_j; N – загальна кількість документів в корпусі, $|\{d_j : t_i \in d\}|$ є числом документів, де з'являється термін t_i (який має $n_{i,j} = 0$).

Така модель дозволяє отримати функцію подібності між запитом Q і документом D у наступному вигляді:

$$S(d_i, q) = \frac{\sum_j D_i \times D_q}{|D_i||D_q|}, \quad (4)$$

де $D_k = tf_k \cdot idf_k$ та $|D_i|$ $|D_q|$ – знаменник для стандартизації.

В імовірнісній моделі оцінюється ймовірність того, що документ d_j має відношення до конкретного запиту q, тобто $P(R|q, d_j)$.

Сукупність термінів, що використовуються в документі d_j, можна представити у вигляді бінарного вектора $x = (x_1, x_2, ..., x_n)$ з $x_i = 1$, якщо термін i присутній у d_j, та $x_i = 0$ в іншому випадку. Потім документи ранжируються в порядку убування відповідно до наступного виразу:

$$S(d_j, q) = \sum_{i=1}^{n} \log \frac{P(x_i|R)(1 - P(x_i|\overline{R}))}{(1 - P(x_i|R))P(x_i|\overline{R})}, \quad (5)$$

де R – сукупність релевантних (позитивних) результатів і $\overline{R}$ – сукупність нерелевантних (негативних) результатів. $P(x|R)$ та $P(x|\overline{R})$ є відповідно ймовірностями релевантних або нерелевантних елементів вектора $x = (x_1, x_2, ..., x_n)$.

Розглянемо далі векторну модель визначення подібності текстово-графічних фрагментів за візуальними характеристиками зображень, автоматично витягнутих з візуального вмісту. Така модель залежить від функції подібності та використовуваних візуальних сигнатур, що отримуються на розглянутому вище етапі хешування зображень, які можуть бути векторами ознак, регіональними характеристиками або узагальненими локальними характеристиками [8]. При цьому функція ранжирування залежить від використовуваних характеристик. Наприклад, подібність текстурних ознак часто вимірюється за допомогою відстані Мінковського або відстані Махалонобіса (що враховує співвідношення різних характеристик). Розглянемо два зображення, індексовані відповідними векторами $I = (I_1, I_2, ..., I_n)$ та $J = (J_1, J_2, ..., J_n)$. Оцінювання подібності між двома зображеннями полягає в обчисленні подібності між I та J. Метрики Мінковського L_p, які є найбільш поширеними геометричними відстанями, мають такий загальний вигляд:

$$L_p = \sqrt[p]{\sum_{i=1}^{n} (I_i - J_i)^p} . \quad (6)$$

Відстань Махалонобіса враховує співвідношення різних характеристик:

$$d = \sqrt{(\vec{I} - \vec{J})^T C^{-1} (\vec{I} - \vec{J})} . \quad (7)$$

де C – матриця коваріації розподілу I та J.

Розглянемо спочатку представлення загальної характеристики зображення у вигляді набору векторів ознак $((z_1, p_1), (z_2, p_2), ..., (z_n, p_n))$, де z_i – вектори ознак, а p_i – відповідні ваги цих ознак.

Нехай ми маємо дві таких характеристики $I_m = ((z_1^m, p_1^m), (z_2^m, p_2^m), ..., (z_n^m, p_n^m))$, $m = 1, 2$. Природним підходом до визначення міри подібності є вимірювання подібності між z_i^1 та z_i^2 з подальшим об'єднанням відстаней між цими векторами та відстані між наборами векторів.

Застосуємо підхід до такого вимірювання, заснований на привласненні кожній парі z_i^1 та z_i^2 таких ваг $s_{i,j}$ $(1 \le i \le n,\ 1 \le j \le n)$, щоб вони відповідали на важливості асоціації між елементами цієї пари. Ваги мають відповідати таким обмеженням: $\sum_i S_{i,j} = p_j^2$; $\sum_j S_{i,j} = p_i^1$. Після визначення ваг відстань між I_1 та I_2 агрегується з відстаней між різними парами векторів:

$$D(I_1, I_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s_{i,j} d(z_i^{(1)}, z_j^{(2)}) . \quad (8)$$

Відстань $d(x, y)$ може бути визначена різними способами. Наприклад, можна використовувати відстань Хаусдорфа, де кожен z_i^1 пристосований до свого найближчого вектора в I_2, тобто z_i^2, а відстань між I_1 та I_2 є максимальною серед усіх $d(z_i^1, z_i^2)$.

Розглянемо етап поєднання процесів порівняння текстово-графічних фрагментів електронних документів за характеристиками (ознаками) зображення та текстовими підписами (ключовими словами).

Візуальна подібність V_d між запитом, пов'язаним з зображенням Q, та текстово-графічного зображення I_i оцінюється за величиною скалярного добутку відповідних векторів:

$$V_d(\overline{I}_i, \overline{Q}) = \sum_j w_{ij} \times w_{qj} \ . \quad (9)$$

Текстова подібність S_d між запитом, пов'язаним з зображенням Q, та текстово-графічного зображення I_i розраховується наступним чином:

$$S_d(I_i, q) = \frac{|K_{i,q}|}{|K_q|}, \quad (10)$$

де $|K_{i,q}|$ – кількість однакових ключових слів для зображення Q та зображень I_i; $|K_q|$ – загальна кількість ключових слів ключових слів у зображенні Q.

Інтегроване значення подібності визначається наступним чином:

$$S_{final} = r * S_q + (1-r) * V_d \quad S_{final} = r * S_q + (1-r) * V_d, \quad (11)$$

де r – ваговий коефіцієнт.

4. Тестування та перспективи розвитку методу

При тестуванні розглянутого вище методу значення r приймались від 0,1 до 0,8 для різних типів текстово-графічних фрагментів електронних документів: зображення з розширеними підписами: r = 0,8; зображення зі короткими підписами: r = 0,1. При цьому було використано колекцію авторефератів дисертаційних робіт. Експериментальне тестування запропонованого методу підтвердило його працездатність. Зокрема, для декількох вибірок текстово-графічних фрагментів електронних документів (з повним обсягом бази в 300 рефератів) були отримані наступні середні значення оціночних характеристик: повнота – 0.85; точність – 0.87; F-міра – 0.85.

Перспективним розвитком системи є дослідження можливості урахування додаткових метаданих зображення (наприклад, інформації про розташування об'єктів, дату, час формування зображень).

Висновки

У статті був розглянутий новий метод порівняння текстово-графічних фрагментів в електронних документах за гібридним критерієм. Цей метод дозволяє визначати інтегроване значення подібності між запитом, пов'язаним з зображенням в запиті, та текстово-графічним зображенням в базі даних пошукової системи. Критерій, що застосовується, передбачає використання вагових коефіцієнтів для зображень з розширеними анотаціями та для зображень без анотацій. Схема запропонованого методу передбачає можливість постійної корекції елементів бази зображень з текстовими анотаціями та збільшення її розмірності. До основних етапів цієї схеми належать індексація зображень, попередня обробка та формування комбінованих запитів та процедура гібридного порівняння текстово-графічних фрагментів електронних документів.

Результати експериментального тестування підтверджують, що такий метод буде ефективний для розробки і модернізації систем аналізу електронних документів, що містять зображення різних типів.

Список літератури:

[1] Quack T. A System for Largescale, Contentbased Web Image Retrieval / T. Quack, U. Monich, L. Thiele, B. Manjunath // MM'04, 2004, New York, USA. – P. 120 –123.

[2] Романюк В.А. Обробка графічної інформації / В. А. Романюк, О.М. Сальніков, В. Г. Малюк та ін.; – Х. : Акад.ВВ МВСУ, 2013. – 112 с.

[3] Kulis B. Kernelized locality-sensitive hashing. Pattern Analysis and Machine Intelligence / B. Kulis, K. Grauman /, IEEE Transactions on, 34(6):1092–1104, 2012. – P. 1092–1104.

[4] Бойцов Л.М. Использование хеширования по сигнатуре для поиска по сходству / Л.М. Бойцов // Прикладная математика и информатика. – М. Изд-во факультета ВМиК, МГУ. – 2001. – № 8. – С. 135 – 154.

[5] Чалая Л. Э. Поиск неполных дубликатов в системах анализа цифровых изображений / Л.Э. Чалая, П.Ю. Попаденко // Вісник КрНУ імені Михайла Остроградського. Вип. 5/2014 (88). – 2014. – С.42 – 47.

[6] Zhang D. Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study / D. Zhang, G. Lu // In IEEE International Conference on Multimedia and Expo. – 2001. – P. 289 – 293.

[7] Волосных Д.Ф. Использование визуальных особенностей восприятия компонент цветовой модели HSI при поиске изображений по содержанию / Д.Ф. Волосных /Труды РОМИП. – 2010 – Режим доступа : http://romip.ru/ru/2010/

[8] Удовенко С. Мультимодальна система порівняння текстово-графічних фрагментів в електронних документах / С. Удовенко, Л. Чала, Є. Кушвід // Матеріали Міжнародної наукової конференції «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту (ISDMCI'2019)». – Залізний Порт, 2019 – С. 184-186.

Поступила до редколегії 10.06.2019

ПРАВИЛА
оформлення рукописів для авторів науково-технічного журналу «БІОНІКА ІНТЕЛЕКТУ»

Науково-технічний журнал «Біоніка інтелекту» приймає до друку написані спеціально для нього оригінальні рукописи, які раніше ніде не друкувались. Структура рукопису повинна бути такою: індекс УДК, відомості про авторів, заголовок, анотації (на трьох мовах), ключові слова, вступ, основний текст статті, висновки, список використаної літератури, резюме.

Відповідно до Постанови ВАК України від 15.01.2003 №7-05/1 (Бюлетень ВАК, №1, 2003, с. 2), стаття повинна мати такі необхідні елементи: постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями; аналіз останніх досліджень і публікацій і виділення не вирішених раніше частин загальної проблеми в даній області; формулювання цілей та завдань дослідження; виклад основного матеріалу досліджень з повним обґрунтуванням отриманих наукових результатів; висновки з даного дослідження та перспективи подальших досліджень у даному напрямку.

Статті мають бути виконані в редакторі Microsoft Word. Формат сторінки – А4 (210×297 мм), поля: верхнє – 25 мм, нижнє – 20 мм, ліве, праве – 17 мм. Кількість колонок – 2, з інтервалом між ними 5 мм, основний шрифт Times New Roman, кегль основного тексту – 10 пунктів, міжрядковий інтервал – множник (1,1), абзацний відступ – 6 мм. Обсяг рукопису – від 6 до 12 сторінок (мови: українська, англійська, російська та мовою оригінала).

УДК друкується з першого рядка, без відступів, вирівнювання по лівому краю.

ПІБ автора (-ів), назва статті, назва та адреса учбового закладу необхідно надати повністю російською, українською та англійською мовами.

Назва статті друкується прописними літерами; шрифт прямий, напівжирний, кегль 12.

Назви розділів нумерують арабськими цифрами, виділяють жирним шрифтом. Відступи для назви статті, ініціалів та прізвищ авторів, відомостей про авторів, назв розділів, вступу та висновків, списку літератури: зверху – 6 пт, знизу – 3 пт.

Анотації (мовою статті, абзац 6–12 рядків, кегль 9) розміщують на початку статті, в ній має бути розміщена інформація про очікувані результати описаних досліджень (на трьох мовах).

Ключові слова (4–10 слів з тексту статті, які з точки зору інформаційного пошуку несуть змістовне навантаження) наводять мовою рукопису, через кому в називному відмінку, кегль 9.

Рисунки та *таблиці* (чорно-білі, контрастні) розміщуються у тексті після першого посилання у вигляді окремих об'єктів і нумерують арабськими цифрами наскрізною нумерацією за наявності більше ніж одного об'єкта. Невеликі схеми, що складаються з 3–4 елементів виконують, використовуючи вставку об'єкта Рисунок Microsoft Word. Більш складні виконують у графічних редакторах у вигляді чорно-білих графічних файлів форматів .tif, .jpg, .wmf, .cdr із розділенням 300 dpi. Рисунки мають міститися у текстовому файлі й обов'язково подаватися окремими файлами з відповідними назвами (наприклад, рис1.jpg).

Усі елементи рисунка, включаючи написи, повинні бути згруповані. Усі написи в рисунках і таблицях мають бути виконані шрифтом Times New Roman, кегль у рисунках – 10, у таблицях – 9.

Рисунок повинен мати центрований підпис (поза рисунком), шрифт 9, відступи зверху і знизу по 6 пт. Ширина рисунка має відповідати ширині колонки (або ширині сторінки).

Формули, символи, змінні повинні бути набрані в **редакторі формул MathType**. Формули розміщують посередині рядка й нумерують за наявності посилань на них у рукописі. Шрифт – Times New Roman. Висота змінної – 10 пунктів, великих і малих індексів – 8 пт, основний математичний символ – 12 (10) пт. Змінні, позначені латинськими літерами, набирають курсивом, грецькі літери, скорочення російських слів і цифри – прямим написанням. Змінні, які є в тексті, також набирають у редакторі формул.

Список літератури вміщує опубліковані джерела, на які є посилання в тексті, укладені у квадратні дужки, друкують без абзацного відступу, кегль 9 пт, відступ зверху – 6 пт.

Після списку літератури з відступом зверху 6 пт зазначають *дату подання статті до редколегії*. Число та місяць задають двозначними числами через крапку. Розмір шрифта – 9 пт, курсив, вирівнювання по правому краю.

Резюме (Times New Roman, кегль — 10 пунктів,) подають англійською мовою: обсяг резюме до 2000 знаків (бажаний переклад). *Структура резюме:* **Background, Materials and methods, Results, Conclusion.**

Разом із рукописом (на аркушах білого паперу формату А4 щільністю 80-90 г/м^2, надрукований на лазерному принтері) необхідно подати такі документи:

1. Заяву, яку повинні підписати всі автори.
2. Акт експертизи про можливість опублікування матеріалів у відкритому друці (якщо потрібно).
3. Рецензію, підписану доктором чи кандидатом наук.
4. Відомості про авторів.
5. Електронний варіант рукопису, резюме та відомостей про авторів на e-mail: bionics@nure.ua.
6. Зробити оплату публікації.

Необхідно також зазначити один з наступних тематичних розділів, якому відповідає рукопис:

1. Теоретичні основи інформатики та кібернетики. Теорія інтелекту.
2. Математичне моделювання. Системний аналіз. Прийняття рішень.
3. Інтелектуальна обробка інформації. Розпізнавання образів.
4. Інформаційні технології та програмно-технічні ком-плекси.
5. Структурна, прикладна та математична лінгвістика.
6. Дискусійні повідомлення.

СОДЕРЖАНИЕ

Наукове видання

БІОНІКА ІНТЕЛЕКТУ

інформація, мова, інтелект

Науково-технічний журнал

№ 1 (92)

2019

Головний редактор — *Г. Г. Четвериков*

Відповідальний редактор — *І. Д. Вечірська*

Комп'ютерна верстка — *О. Б. Ісаєва*

Рекомендовано Вченою Радою
Харківського національного університету радіоелектроніки
(протокол № 8/21 от 03.07.2019)

Адреса редакції:

Україна, 61166, Харків-166, просп. Науки, 14,
Харківський національний університет радіоелектроніки, к. 127
тел. 702-14-77, факс 702-10-13,
e-mail: bionics@nure.ua

Підписано до друку 05.07.2019. Формат 60 × 84 $^{1}/_{8}$. Друк ризографічний.
Папір офсетний. Гарнітура Newton. Умов. друк. арк. 15,4. Обл.-вид. арк. 15,0.
Тираж 100 прим.

Віддруковано в редакційно-видавничому відділі ХНУРЕ
61166, Харків, просп. Науки, 14.