МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 10 Москва 2020

ОБЩИЙ РАЗДЕЛ

УДК 81’32’37:53

А.А. Лебедев, Н.В. Максимов

Аналогии в физике и обработке информации

Рассматриваются сходства эмпирических зависимостей в языке и теоретических физических законах. Анализируется вопрос о применении квантовомеханических моделей для описания семантической составляющей текстовых документов. Предлагаются возможные аналогии между лингвистическими и физическими объектами.

Ключевые слова: *информация, компьютерная лингвистика, метод аналогий, физика, семантика*

DOI: 10.36535/0548-0027-2020-10-1

ВВЕДЕНИЕ

Информация присуща и материальному, и абстрактному миру, естественным и искусственным системам. Но к какому бы миру ни относились объекты и процессы, в мир коммуникаций и смыслов они входят посредством образов (свойств, имен, сообщений), причем в мире человека эти образы имеют преимущественно лингвистическую (семиотическую) основу.

Лингвистика и физика, представляя разные логические уровни мироздания с разными типами объектов, могут, тем не менее, иметь одинаковые закономерности. В целом, привлечение концептуальных физических аналогий, в том числе и для задач извлечения знаний, на данный момент не является чем-то уникальным. Основная проблема заключается в выработке и обосновании конкретных (измеримых)

* Работа выполнена при поддержке Министерства науки и высшего образования РФ (проект государственного задания № 0723-2020-0036)

аналогий между предметными областями. Но на этапе отбора возможных моделей, очевидно, можно и полезно опираться на физическую картину мира, проецируя свойства объектов из физики на схожие объекты в языке.

Язык – сложная самоорганизующаяся динамическая система. Нам не доступны для непосредственного наблюдения те механизмы, которые отвечают за системность языка, но мы можем пытаться их смоделировать [1], используя, в частности, аналогии.

Похожесть, аналогичность двух явлений обычно объясняется совпадением закономерностей, которым они подчиняются. Абстрактные модели (теории) двух явлений могут «перекрываться», и это приводит к похожести данных о явлениях. Поэтому наблюдая одно из них, можно высказывать суждение о другом. И одной из самых богатых на полезные аналогии областей знания является физика. Так, количественные модели из физики, полученные при исследовании окружающего мира, находят сегодня свое применение в биологии [2], экономике [3], социологии [4] и т.д. Кроме того, внутри самой физики, модели, полученные в рамках одних представлений и одной теории, могут быть использованы для описания других явлений или объектов. Поэтому интересно было бы рассмотреть аналогии из данной области, применимые к обработке текстов.

Среди работ, ассоциирующих методы и свойства таких достаточно разных дисциплин как компьютерная лингвистика и физика можно выделить три следующих распространенных методологических подхода (точек зрения).

1. Есть сходство характера зависимостей между лингвистическими и физическими объектами, но для него не существует фундаментального обоснования. Действительно, трудно убедить себя в том, что есть связь (детерминированное соответствие) между такими материальными объектами, как элементарные частицы (которые мы непосредственно не наблюдаем, но можем зарегистрировать, и объективность их существования считается неоспоримой) и гораздо более субъективными абстрактными объектами такими, как знак или смысл. Однако в частных случаях, когда на содержательном уровне сходство кажется очевидным, можно воспользоваться готовыми моделями уже хотя бы потому, что это позволяет упростить моделирование и расчеты. Большая часть исследований придерживается именно этого подхода, и главным критерием для выбора той или иной физической модели является близость выбранной зависимости рассматриваемым эмпирическим данным.

2. Свойства языка опосредованно связаны с физикой через физиологию. Все используемые и изучаемые человеком объекты и их взаимодействия, так или иначе, связаны с физическим миром. Это относится и к абстрактным объектам (но которые существуют в сознании «физического» человека), поскольку их взаимодействие с окружающим миром возможно, только если они будут представлены материальными средствами – языком коммуникаций, или будут воплощены в объектах и процессах. Например, в [5] показана аналогия алфавитов и ряда аминокислот, и как следствие – показана схожесть алфавитов основных языковых групп.

Другой, иллюстрирующий данный подход, пример приведен в [6], где высказана идея о связи частотных распределений текста (речи) с пространственно-временной организацией периодических процессов головного мозга (кодировании образов слов «пакетами волн нейронной активности»). В [7] рассмотрена связь активности областей мозга с процессами определения семантики слов в тексте. Более фундаментальная идея о врожденном характере языка представлена в трудах Н. Хомского и в его теории универсальной грамматики.

Таким образом, поскольку биологические процессы[1] в итоге реализуются физико-химическими методами, то и в лингвистике мы можем использовать закономерности, наблюдаемые в физике.

3. Наличие зависимостей сходного вида свидетельствует о принципиальной связи между процессами в языке и в физике. Все процессы так или иначе имеют информационную природу [8], а общие информационные законы и зависимости играют роль фундаментальных. Например, закон Ципфа в социальных явлениях играет, по-видимому, ту же роль, что и нормальное распределение в физике. Закономерности теории информации Шеннона сильно связаны со статистической физикой. Таким образом, здесь информация и информационные процессы (в том числе язык и формирование текстов) являются неотъемлемой частью физической картины мира [9] и, следовательно, общность (сходство) закономерностей в разных областях вполне вероятна. Данной точки зрения придерживаются и авторы настоящей статьи.

АНАЛОГИИ МЕЖДУ ФИЗИЧЕСКИМИ И ЛИНГВИСТИЧЕСКИМИ ОБЪЕКТАМИ

Одной из самых очевидных аналогий является иерархичность. Будем рассматривать тексты на естественном языке и сам язык как систему из порождающих элементов и правил порождения (они будут определять связи между элементами). В качестве таких элементов могут выступать алфавит, слова, словосочетания, предложения, тексты.

При этом на нескольких иерархических уровнях идет порождение:

- слов из элементов алфавита (букв);
- словаря из слов и порождение словосочетаний из элементов словаря;
- предложений из слов и словосочетаний;
- текстов из предложений.

На каждом из уровней есть свои правила порождения. И здесь мы можем увидеть сходство с описанием физического мира, где определяющие картину реальности взаимодействия зависят от рассматриваемого пространственно-энергетического масштаба [10]. Будем рассматривать различные правила порождения (сочетания букв, построение новых слов с

[1] Отметим, что единственная известная на сегодня «машина», реализующая преобразование (взаимодействие) абстрактных объектов в реальные – это человек.

использованием приставок и суффиксов, построение предложений и т.д.) как влияние разных сил на разных масштабах рассмотрения: на уровне кварков (сильного взаимодействия), на уровне ядра (слабых ядерных взаимодействий), на уровне элементарных частиц и вещества (электромагнитного взаимодействия), на уровне космологическом (гравитационного взаимодействия).

Уровень сильного взаимодействия

Главную роль на уровне кварков (элементов, образующих адроны, в том числе протоны и нейтроны) играет сильное взаимодействие [10]. Аналогом кварков в языке могут выступать буквы, поскольку между этими объектами есть определенное сходство. Например:

- в физике выделяют шесть типов кварков, обозначаемых символами **u**, **d**, **s**, **c**, **b**, **t** (up, down, strange, charmed, bottom, top). В языке алфавит состоит из конечного множества букв (элементов). Для современного русского языка это буквы: «а», «б», «в» … «я»;
- каждый кварк обладает набором специфических свойств (электрический заряд, спин, цвет и т.п.). Буквы можно разделить на подмножества по специфическим свойствам (гласные/согласные, звонкие/глухие и т.п.);
- кварки группируются в пары/тройки, например, выделяют пары по сходным свойствам: **ud**, **cs**, **tb**. В языке буквы группируются в слоги («но», «то» и др.);
- группировка кварков происходит с учетом их свойств (не каждая комбинация разрешена – так, в образованной группе электрический заряд должен быть целым). Легкие кварки **u** (up) и **d** (down) — самые распространенные в природе. Из них состоят протоны (**uud**), нейтроны (**udd**) [11]. И в языке одни варианты сочетания букв статистически более ожидаемы, чем другие. Например, для русского языка комбинация слога типа «согласная − гласная» составляет 54% всех встречаемых, типа «согласная – согласная − гласная» и «согласная − гласная − согласная» по 14% [12];
- в свободном состоянии кварки не встречаются. Аналогом такого явления в языке будет то, что для передачи смыслов буквы самостоятельно не используются.

Таким образом, также как кварки выступают «строительным материалом» адронов, так и алфавит является «строительным материалом» в языковых конструкциях.

Уровень слабого взаимодействия

При росте энергии взаимодействующих частиц начинают проявляться процессы, обусловленные слабым ядерным взаимодействием. Их интенсивность меньше, чем у процессов, вызванных сильным и электромагнитным взаимодействием (т.е. процессы протекают медленнее), кроме того, слабое взаимодействие имеет малый радиус действия. В физике слабое взаимодействие отвечает за радиоактивный распад атомных ядер и ядерные реакции синтеза [10]. В языке аналогом слабого взаимодействия являются правила, по которым слово изменяется во времени (переразложение, изменение длины слова или количества букв алфавита). Приведем соответствующее сравнение свойств физических и лингвистических объектов:

- в физике известно, что тяжёлые кварки или лептоны[2] могут распадаться на лёгкие и более стабильные. Здесь видна аналогия с тем, что слова изменяют свою длину под влиянием развития языка. Если учитывать очевидно неравномерное распределение слов по длине [13], то можно заметить тенденцию к уменьшению длины слова со временем;
- слабое взаимодействие имеет малый радиус действия, и в языке процессы распада (уменьшения длины) происходят на уровне слова;
- слабое взаимодействие позволяет лептонам, кваркам и их античастицам обмениваться энергией, массой, электрическим зарядом и квантовыми числами, т. е. превращаться друг в друга. И в языке слова, за счет механизмов словообразования, могут переходить из одной части речи в другую, например, «забегать — *забегаловка*» [14].

В современной физике слабое взаимодействие не всегда рассматривается отдельно, его иногда объединяют с электромагнитным, получая единое электрослабое взаимодействие. Для языка это означает что правила словообразования и правила построения словосочетаний и предложений также могут рассматриваться совместно.

Уровень электромагнитного взаимодействия.

Существенную роль на уровне элементарных частиц и макровещества играет электромагнитное взаимодействие. Как в физике электромагнитное взаимодействие отвечает за межатомное взаимодействие, так и в языке его аналог – грамматика, отвечает за словообразование. Рассмотрим иллюстрирующие данную аналогию примеры:

- в зависимости от знака электрического заряда частицы могут отталкиваться/притягиваться или не взаимодействовать. В языке в зависимости от состава и типа морфем они могут сочетаться или не сочетаться;
- в физике электромагнитное взаимодействие зависит от расстояния между частицами. В языке взаимодействие между морфемами зависит от расстояния в функциональном пространстве между ними в слове (префикс – корень связаны сильнее чем префикс – флексия);
- электромагнитное взаимодействие отвечает за строение вещества (как отдельных атомов, так и макровещества). Правила грамматики отвечают за формирование лексикона (словаря) и построение предложений.

На этом уровне в языке, как и в физике, мы наблюдаем формирование разнообразных устойчивых объектов (слов), способных сочетаться в новые устойчивые образования (словосочетания и предложения).

[2] Фундаментальные частицы с полуцелым спином, не участвующие в сильном взаимодействии.

Уровень гравитационного взаимодействия

На космологических масштабах превалирует гравитация. В языке её аналогом могут выступать правила формирования текстов из предложений и правила формирования рубрик. Приведем качественные аналогии между гравитационным взаимодействием и явлениями в языке:

- в физике гравитация отвечает за формирование масштабных пространственных структур, сила такого типа взаимодействия быстро уменьшается с расстоянием. Правила языка рекомендуют, чтобы имеющие общий предмет изложения фрагменты текста были расположены рядом. Близкие по смыслу тексты группируются в тематические рубрики;
- гравитация определяет геометрию пространства-времени, в котором движется материя, т. е. геометрия не задана изначально, а определяется распределением материи. В языке расположение текста (порядок следования его фрагментов) определяет семантический образ (пространство) документа. Изменение порядка следования текста ведет к изменению его семантического образа.

Стоит отметить, что формирование текстов определяется не столько стилистикой конкретных будущих документов, сколько целевыми потребностями субъектов, использующих данные тексты.

Меры и количественные закономерности

Другие аналогии между языком и физикой можно увидеть в количественной лингвистике.

Основываясь на статистических данных слова в тексте можно рассматривать как классический или почти классический газ, где частота встречаемости слов или их полисемия выступает в роли энергии (температуры) [15–17]. Наблюдаются зависимости между длиной слова и частотой его встречаемости [18], длиной слова и его полисемией [19]. Распределения предложений по длине (в словах) сходно с распределением Максвелла или распределением Планка для излучающего тела [20]. Ранговые распределения стоп-слов и редко используемых слов по частоте использования схожи с распределениями Бозе–Эйнштейна и Ферми–Дирака, соответственно [21, 22].

Обобщая полученные в различных работах результаты, можно привести следующие соотношения:

- в физике **энергия** – это общая количественная мера движения и/или взаимодействия. В языке аналогом энергии могут выступать: 1) частота встречаемости слова в тексте или некоторой совокупности текстов [16, 21, 22]; 2) длина слова в символах, слогах, морфемах [15]; 3) число значений одного слова [17]; 4) длина предложения в словах [20];
- материальные объекты в физике характеризуются **массой** (количество вещества материи в теле) $m = \rho \cdot V$, где m – масса тела, ρ – плотность тела, V – объем тела. Аналогом массы могут выступать длина слова или словосочетания в знаках (буквах); длина слова или словосочетания в словах; комбинация длины в знаках и длины в словах (например, их произведение [23]);
- одним из фундаментальных понятий в физике является «пространство», с помощью которого описываются свойства взаимного расположения и протяженности объектов. В качестве меры удаленности или близости двух объектов в пространстве используется **расстояние.** Для лингвистических объектов также можно ввести меры близости. Так, для слов как последовательности символов (букв) можно ввести *расстояние Левенштейна* – минимальное количество односимвольных операций (а именно вставки, удаления, замены), необходимых для превращения одной последовательности символов в другую. Если представить тексты как неупорядоченное множество слов с различной частой употребления, то можно ввести другую меру. Тогда на основе текстов можно построить матрицу термин-документ, в которой отдельные строки будут представлять собой образы текстов как векторы некоторого пространства [24], где компонентами вектора будут «веса» отдельных слов. Здесь мерой близости двух векторов принято считать косинус угла между ними, что может определять, например, близость смыслов текстов, описанных данными векторами;
- согласно работе [14], слово «…после своего появления в языке в некотором начальном значении может либо сохранять это значение в течение всей своей жизни, либо претерпевать эволюцию, последовательно рождая новые значения…». Для меры противодействия употреблению слова в новом смысле, например, в [23] рассматривается аналогия с высотой потенциального барьера – минимальной энергией классической частицы, необходимой для преодоления области пространства, разделяющей две другие области с различными или одинаковыми потенциальными энергиями.

Из приведенных сравнений можно увидеть, что как в языке, так и в физике, «энергия» и «масса» могут быть эквивалентны, т.е. выражаться через одни и те же величины (например, через длину слова). «Энергия» в языке, как и в физике, может принимать разные виды (в физике – кинетическая и потенциальная, в языке - как количество букв в слове и как количество смыслов у слова). Кроме того, появляется пространство для смысла слов, где происходит их взаимодействие – так называемое семантическое пространство, либо семантическое поле.

ФИЗИЧЕСКИЕ АНАЛОГИИ ПРИМЕНИТЕЛЬНО К СЕМАНТИКЕ ТЕКСТОВ

В рассмотренных выше аналогиях практически не уделялось внимание семантике. Связано это с тем, что совокупность смыслов слов и их отношений (семантическое пространство) не входит в выделенные ранее уровни иерархии, а существует практически независимо, «пронизывая» их все и позволяя осуществлять переходы между ними.

Семантика тесно связана с понятием «информации». Не будем здесь останавливаться на проблеме определения информации (с этим можно ознакомиться в [25]): достаточно определить информацию согласно ГОСТ 7.0-99 как «сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации», причём восприятие это целенаправленное и имеет прагматический характер.

В ряде публикаций [9, 26] озвучена идея, что между информационными и физическими процессами нет принципиальных различий. «Информация неизбежно связана с физическим представлением и, следовательно, с ограничениями и возможностями, связанными с законами физики...» [9]. Таким образом, информация выступает как дополнение к уже известным физическим понятиям «материя», «вещество», «энергия» и должна подчиняться общим физическим законам [25].

Квантовомеханические аналогии. Общий подход

Семантика проявляется только во взаимодействии текстов с окружением (активным приемником), и результат взаимодействия существенным образом зависит и от состояния текста, и от состояния окружения. Здесь можно увидеть аналогию с квантовой механикой, где объект изучения – волновая функция или матрица плотности, описывающая квантовомеханическую систему [11] – тесно связан с измеряющим его прибором. Следует отметить, что подобные попытки применения аппарата квантовой механики в несвойственных областях тоже не уникальны. Так, есть работы по «квантовой» экономике и даже «квантовой» истории [11, 27].

Можно выделить следующие основные положения, подкрепляющие идею использования инструментария квантовой механики.

1. Смысл слова есть вероятностная характеристика *процесса и результата соотнесения его* знакового образа с некоторым явлением окружающей действительности, и задается некоторой функцией распределения в многомерном пространстве [28].

Таким образом, слово/термин (как аналог наблюдаемой[3] в квантовой механике) до попытки определения его смысла (аналог измерения/соотнесения с эталоном) содержит в себе множество возможных смыслов, каждый из которых можно получить в результате измерения с некоторой вероятностью.

Состояние любой квантовой частицы в каждый момент времени задает волновая функция, определенная с точностью до фазового множителя [29]. По аналогии будем считать, что и состояние слова задается волновой функцией:

$$\psi(x,t) = A(x,t)\cdot\exp(-i\cdot\varphi\cdot t)$$

так, что:

$$\left|\psi(x_0,t)\right|^2 dx = P(x_0,t)$$

где: x – координата в смысловом пространстве;

t – время;

i – мнимая единица;

$A(x,t)$ – амплитуда вероятности;

φ – фаза;

$P(x_0,t)$ – вероятность того, что термин (слово) имеет смысл, соответствующий некоторой окрестности dx точки x_0 смыслового пространства [23].

2. Смысл термина можно представить вектором $\vec{T}$ в некотором пространстве:

$$\vec{T} = \Sigma_i C_i \cdot \vec{e_i},$$

где: $\vec{e_i}$ – базисные вектора выбранного пространства,

C_i – компоненты вектора $\vec{T}$.

Примером таких векторов могут служить строки матрицы термин – термин, построенной на некотором массиве текстов. Тогда векторами будут выступать отдельные слова, а компонентами векторов – частота их совместной встречаемости.

Идея векторного представления семантики предложена достаточно давно [24]. Этот подход и ныне широко используется в алгоритмах информационного поиска (например, ***Word2vec)***, безотносительно стоящей за ним парадигмы. Семантическая близость терминов при этом может определяться через угол между векторами, или расстоянием между ними. В качестве базовых векторов обычно принимают термины, наиболее часто встречающиеся с исходным. Тогда компоненты C_i будут характеризовать частоту совместной встречаемости.

Автор работы [30], где рассмотрен вопрос о свойствах пространства, которому принадлежат рассмотренные выше вектора, приходит к выводу, что это пространство, как и в квантовой механике, является гильбертовым, а не евклидовым.

3. Информация некоммутативна, т.е. важен порядок поступления информации [31]. Некоммутативность информации иногда объясняют, используя матричное представление информации [32]. Например, матрица «термин – документ» как распределение терминов по массиву документов, или тематико-статистический спектр информационного потока как распределение семантических единиц (документов или терминов) по тематическим рубрикам [33]. Матричное представление широко используется в квантовой механике, где состояние квантовой системы может быть описано через матрицу плотности [11, 29]. Применение формализма матриц плотности в задачах информационного поиска рассмотрено в [34].

В [35] предложено рассматривать информацию как суперпозицию возможных состояний информационного объекта в n предметных областях, характеризуемую функцией:

$$\Lambda = c_1E_1 + c_2E_2 + ... + c_nE_n \ ,$$

где E_i– функция распределения объекта для *i-й* предметной области;

c_i – вероятностный показатель отнесения объекта к *i-й* предметной области.

[3] Наблюдаемой в квантовой механике называется величина, значения которой измеряются в эксперименте.

Соответственно, информационное взаимодействие характеризуется функцией:

$$\Omega = c_1 E_1 \Theta_1^T + c_2 E_2 \Theta_2^T + ... + c_n E_n \Theta_n^T \ ,$$

где Θ_i – функция распределения вероятностей для информационного объекта, с которым взаимодействует исходный информационный объект, для *i-й* предметной области, т.е. информация, взаимодействуя с конкретной предметной областью (окружением), фактически принимает единственное состояние из множества возможных. Такое, случившееся в результате выбора, состояние информации, зафиксированное в виде контекстно дообусловленного информационного объекта, составляет знание. Аналогично, в квантовой механике в результате измерения (взаимодействия с измеряющим прибором) реализуется только одно из возможных квантовых состояний объекта.

4. Информация эмерджентна. Согласно [31] эмерджентными называют такие свойства сложных систем, которые порождаются взаимодействием элементов и не наблюдаются ни в одном из элементов, если они рассматриваются отдельно (система больше суммы своих частей). Так, в результате взаимодействия смыслов терминов может возникать новый смысл, несводимый к сумме составляющих, или относящийся к новой предметной области.

Данное свойство можно представить как изменение размерности одной или нескольких матриц компонентов информационного взаимодействия $E_i \Theta_i^T$ [35]. К примерам подобных явлений можно отнести формирование слов с несколькими корнями (пар-о-ход, документ-о-оборот) или устойчивых выражений (красная строка, гол как сокол).

5. Значение (смысл) термина определяется контекстом его использования, что можно сравнить с процедурой приготовления состояния, что в [36] называется *guppy effect*. Задавая контекст, мы заранее подготавливаем необходимые возможные смыслы терминов. В качестве физической аналогии можно привести пример из [11]: «Мы можем приготовить фотоны в состоянии с определённой линейной поляризацией, пропустив их через поляризатор. Часть фотонов при этом окажется забракованной (поглотится или отразится, в зависимости от устройства поляризатора)». Таким образом, контекст выступает как поляризационный фильтр, который «убирает» из рассмотрения значения, не относящиеся к целевым.

Квантовомеханические аналогии. Следствия

Приведенные ранее аналогии позволяют перейти к моделям, рассматривающим не только свойства и близость статистических характеристик, но и сходство между процессами в физике и языке.

1. **Текст документа можно представить как квантовомеханическую систему,** специально подготовленную по некоторому правилу, т. е. для поиска возможных интерпретаций свойств дуального объекта термин/документ (сочетающему в себе связку «знак»+«значение знака») могут быть использованы известные физические модели.

2. Сам процесс отождествления смысла термина (текста) может быть аналогичен феномену редукции (декогеренции) волновой функции квантового объекта. Здесь целесообразно рассмотреть теорию квантового дарвинизма [37], хотя проблема декогеренции выходит за рамки данной работы.

Согласно [37], квантовые системы, изолированные от окружающей среды, в общем случае находятся в состоянии суперпозиции базисных состояний, для которых сохраняются фазовые соотношения между базисными компонентами волновой функции системы, т. е. когерентность. Общая волновая функция системы и ее окружения может быть представлена как произведение волновой функции системы на волновую функцию окружения, что соответствует их взаимной независимости. При взаимодействии системы с окружающей средой происходит декогеренция: отдельные части исходной системы запутываются с компонентами окружения, т.е. общая волновая функция уже не может быть представлена как произведение волновых функций. В ходе декогеренции возникают корреляции между состоянием системы и ее окружения, т.е. в окружение «записывается» информация об исходной системе. Таким образом суть квантового дарвинизма состоит в том, что только те состояния, которые формируют множественные информационные всплески (*multiple informational off-spring*) – множественные отпечатки окружения, могут быть обнаружены с помощью малых фрагментов окружения. Природа возникающей при этом классичности (приборной регистрируемости) состоит в их способности «порождать», формировать множественные записи – собственные копии – с помощью окружения. Объективное существование – критерий классичности – возникает как следствие избыточности [37].

Так и в языке понимание содержания сообщения (текста) происходит за счет избыточности образующих его терминов, позволяющей эффективно взаимодействовать с окружением (т.е приемником сообщения). Чем больше избыточных частей сообщения согласовалось с окружением, тем больше вероятность того, что сообщение будет интерпретировано верно (т.е. «выживет» исходный смысл сообщения).

3. **Эффекты преобразования смыслов** можно описать, используя представление смыслов в виде как массивных тел, так и безмассовых волн.

В [23] предложена модель семантического сдвига термина по аналогии с эффектом квантового туннелирования. Причиной изменения смысла является его принципиальная нелокализуемость.

Другую аналогию можно увидеть в [38]: «Квантовые корреляции могут также рассеивать информацию через степени свободы, которые в действительности недоступны для наблюдателя. Взаимодействие со степенями свободы, внешними по отношению к системе – ее окружением – создает такую возможность». Процесс рассеяния можно рассматривать как процесс изменения смысла.

4. Приведем также аналогии между квантовой механикой и теорией информационного поиска, рассмотренные в работе [39].

Таблица 1

Аналогии между квантовой механикой и информационным поиском

Квантовая механика	Информационный поиск
a quantum system Квантовая система	*a collection of object for retrieval* Информационно-поисковый массив (например, массив документов).
complex Hilbert space Комплексное гильбертово пространство	*information space* Информационное (семантическое) пространство
state vector Вектор состояния (полное описание квантовой системы)	*objects in collection* Документы из информационно-поискового массива
observable Наблюдаемая	*query* Поисковый образ документа (например, выраженный в форме запроса)
measurement Измерение	*search* Информационный поиск
eigenvalues Собственные значения	*relevant or not for one object* Релевантные или нерелевантные документы
probability of getting one eigenvalue Вероятность получения одного из собственных значений	*relevance degree of object to query* Степень релевантности документа информационной потребности

Таблица 2

Аналогии между квантовой механикой и векторной моделью информационного поиска

Квантовая механика	Векторная модель
complex Hilbert space Комплексное гильбертово пространство	*real Euclidean space* Действительное евклидово пространство
state vector Вектор состояния	*vector* Вектор (например, векторная модель массива документов)
self-adjoint linear operator for observable Самосопряженный линейный оператор наблюдаемой	*query vector* Поисковый образ документа, выраженный в виде вектора
measurement – interaction between measurement device and quantum system Измерение – взаимодействие между измеряющим прибором и квантовой системой	*search – may involve interaction between user and system* Информационный поиск – может включать взаимодействие между пользователем и системой
eigenvalues of the operator Собственные значения оператора наблюдаемой	*relevant or not for one object* Релевантные или нерелевантные документы
probability of obtain one eigenvalue Вероятность получения одного из собственных значений	*relevance degree of object to query* Степень релевантности документа к запросу

В табл. 1 представлены аналогии между квантовой механикой и информационным поиском, а в табл. 2 – аналогии между квантовой механикой и векторной моделью информационного поиска

Представленные сопоставления (см. табл. 1 и 2) показывают подобие основных компонентов, что может свидетельствовать о сходстве природы рассматриваемых явлений и процессов.

АНАЛОГИИ В ЧАСТИ ВОЛНОВЫХ СВОЙСТВ

Информация имеет двойственность состояния: до взаимодействия – это некоторый цельный неделимый объект, во время взаимодействия – это макрообразование «квантов», которые могут быть составляющими и других, существующих или гипотетических, объектов. Информация в фазе стабильного состояния (хранимый или передаваемый информационный объект) обладает свойствами макрообъекта, а в фазе взаимодействия – волновыми.

«Волны[4] – изменения некоторой совокупности физических величин (полей), способные перемещаться (распространяться), удаляясь от места их возникновения, или колебаться внутри ограниченных областей пространства». Сложный волновой процесс может быть рассмотрен как сумма нескольких гармонических колебаний с разной амплитудой и частотой.

Волновые свойства и явления могут быть использованы и при формализации свойств документальной

[4] Здесь и далее определения для физических величин даны из Физической энциклопедии под ред. М. Прохорова. – М.: Советская Энциклопедия, 1988.

информации. Процесс информационного взаимодействия можно представить как распространение информации от источника к приемнику с дальнейшей обработкой её приемнике, что, в свою очередь, можно представить как движение некой «волны информации» (информационных объектов) от источника к приемнику.

Параметры волны

1. «Длина волны – расстояние между двумя ближайшими друг к другу точками в пространстве, в которых колебания происходят в одинаковой фазе».

Под «длиной волны» применительно к информации можно подразумевать некоторый количественный аспект, например, число символов в тексте или семантическое расстояние между документами в которых встречается заданный термин. В работе [40] длина волны определяется через частоту использования термина в тексте и связана с его положением в тексте.

2. «Амплитуда – максимальное значение смещения или изменения переменной величины от среднего значения при колебательном или волновом движении».

В физике квадрат амплитуды электромагнитной волны может характеризовать число фотонов, попадающих в определенную область, а при малых значениях – уже вероятность обнаружения фотона в заданной области.

Для рассматриваемых задач под амплитудой можно подразумевать величину, характеризующую число документов в заданной предметной области или вероятность обнаружить термин/документ в заданной предметной области.

3. «Частота – физическая величина, характеристика периодического процесса, равна количеству повторений или возникновения событий (процессов) в единицу времени».

Наиболее очевидный вариант – связать частоту колебаний с частотой употребления термина в тексте или выборке текстов [23]. В [41] частота характеризует степень близости смыслов употребления термина в разных текстах, т. е. чем менее вероятен в данном наборе текстов контекст употребления термина, тем больше связанная с ним частота излученной волны.

Частоту можно связать с длиной волны через скорость её распространения, например, время, за которое информационный объект переместится на «длину волны». Тогда частота будет характеризовать «устойчивость» смыслового содержания информационного объекта.

4. «Фаза колебаний – аргумент периодической функции, описывающей колебательный или волновой процесс». Фаза колебания тесно связана с частотой волны, но колебания с одинаковыми амплитудами и частотами могут различаться фазами.

В лингвистике фазу можно связать с предметной областью. Тогда «волна информации» от одного источника есть суперпозиция волн одной длины, но разных фаз. В таком представлении нужен фазовый фильтр в приемнике, который бы разделял исходный волновой пакет на предметные области.

Другой вариант – когда фаза связывается с наличием в приемнике информации, соответствующей излученной источником. Например, если такая информация в приемнике уже есть, то информация, испущенная источником, будет по отношению к имеющейся в противофазе. Тогда в сумме они взаимопогасятся, т.е. взаимодействие не приведет к изменениям в приемнике.

Как правило, для вычислений принципиально знать не абсолютную величину фазы, а разность фаз между двумя взаимодействующими объектами [36, 42].

5. «Поляризация волн – характеристика поперечных волн, описывающая поведение вектора колеблющейся величины в плоскости, перпендикулярной направлению распространения волны».

Если на пути распространения такой волны поставить поляризационный фильтр, то в зависимости от величины угла между направлением поляризации в волне и в фильтре, волна может либо пройти полностью, либо не пройти совсем, либо пройти частично. Таким образом какие-то процессы в приемнике могут играть роль поляризационного фильтра.

Так, в работе [43] предложена квантовомеханическая аналогия, где в процессах информационного поиска документы рассматриваются как фотоны, поисковый запрос – как поляризационный фильтр, а поляризация – как степень релевантности документа поисковому запросу. Расширение запроса выступает в качестве дополнительного фильтра.

Виды взаимодействия волн

1. Интерференция волн – это явление наложения когерентных волн, свойственное волнам любой природы (механическим, электромагнитным и т.д.).

Под «интерференцией информации» можно понимать такой тип информационных взаимодействий, в которых совместное воздействие нескольких «слабых» источников приводит к возникновению «сильного» отклика на приемнике.

Вопрос использования интерференции для объяснения некоторых результатов в задачах информационного поиска рассмотрен в работах [36, 43–45].

2. Дифракция волн — явление, которое проявляет себя как отклонение от законов геометрической оптики при распространении волн. Она представляет собой универсальное волновое явление и характеризуется одними и теми же законами при наблюдении волновых полей разной природы.

Степень проявления дифракции зависит от соотношения длины волны и ширины волнового фронта, либо от размера непрозрачного экрана на пути распространения фронта, либо от неоднородностей структуры самой волны.

В информационных взаимодействиях можно видеть следующие аналогии дифракции:

- тип информационных взаимодействий, в котором распределение информации (в источнике) в одной предметной области влияет на распределение информации (в приемнике) в другой предметной области;
- тип информационных взаимодействий, в которых при «перекрытии» части текста, несущего информацию, возможно восстановить первоначальный смысл сообщения по оставшейся части, т. е. «волна»,

несущая информацию, как бы огибает препятствие и проникает в области за ним.

3. Дисперсия волн в физике – это зависимость фазовой скорости волны от её частоты.

Под «дисперсией информации» можно подразумевать, что при распространении информации от источника к приемнику более короткие сообщения будут обрабатываться приемником раньше. Это приводит нас к следующей аналогии: изначальная информация, излучаемая источником, есть суперпозиция волн разной длины (каждая из набора длин волны относится к какой-либо предметной области).

4. Совместное действие «дисперсии» и «интерференции» может привести к возникновению «экстремальных волн», которые систематически возникают при распространении волновых пакетов в нелинейной среде. Они представляют собой самопроизвольную концентрацию волновой энергии в малых областях пространства. Возникнув, экстремальные волны могут либо существовать некоторое время в виде движущихся фокусов, либо привести к своеобразной «вспышке», приводящей к быстрой диссипации волновой энергии. Экстремальные волны весьма разнообразны. Они широко распространены в природе (самофокусировка света, волны-убийцы) и наблюдаются в лабораторных экспериментах.

Подобными «вспышками» можно описать возникновение инсайтов (внезапных озарений, нового понимания, постижения существенных отношений, задач, проблем и структуры ситуации в целом). Широко известным примером внезапного озарения может служить легенда об Архимеде, когда он смог решить проблему измерения объема короны увидев, как вода выливается из ванны во время купания. Считается, что для возникновения инсайтов требуется, с одной стороны, накопить критический объем информации о проблеме, с другой стороны – отвлечься от сознательного поиска решения (переключиться на другую деятельность). Возможно, что отвлечение позволяет накопленной информации более эффективно распространяться в семантическом пространстве, вступая в большее количество взаимодействий. А поскольку изначально информация подбиралась под конкретную задачу, то возникает эффект самофокусировки, направленности на конкретную предметную область.

ЗАКЛЮЧЕНИЕ

Из представленного материала видно, что различные физические аналогии находят свое применение при описании широкого круга информационных и лингвистических явлений, а это позволяет говорить о принципиальной адекватности метода аналогий.

В реальном мире, включающем и физический, и лингвистический компоненты, наблюдается возрастание сложности. Поведение каждого элемента в системе определяется не рамками этого элемента, а его связями со всей системой. И чем сложнее объект, тем в большей степени его свойства зависят от окружения. Кроме того, возрастание сложности приводит к возникновению новых типов структур или поведения, появлению свойства эмерджентности. То есть эмерджентность – это следствие существования «логического» уровня: имеется некоторое правило выделения системы – внешнее свойство, позволяющее считать целостным то, что изначально является множеством элементов. Кроме того, язык – это не только средство отображения, но и инструмент познания (текст как модель описываемого объекта), т. е. он обладает свойствами, обеспечивающими развитие предметной области, синтез новых структур и свойств.

Однако в отличие от «уровней» физического мира, где законы физики определяют пространство возможностей для всего, что появляется в природе, семантические конструкции определяются той логикой, которая *удобна* субъекту, т. е. локально замкнутой и компактной, и потому конструктивной для вариантного и быстрого (по сравнению с природой) создания новых структур и новых качеств.

Логика физических теорий опирается и отражает законы природы. Каждая из теорий может иметь ту или иную область существования, но она всегда детерминирована. Физические теории носят вычислительный характер и зависимости заданы на точных переменных. Адекватность теории подтверждается не только внутренней согласованностью, но и количественно измеряемыми результатами экспериментов. В отличие от этого семантические структуры (и теории) заданы на лингвистических переменных, и их смысл существенным образом зависит от контекста[5]. Но при всей схожести механизмов (например, интерференции или дифракции) провести экспериментальное исследование любых свойств, аналогично физическим, удается только для конкретного случая, что скорее заставит читателя усомниться в адекватности модели и целесообразности применении аналогии.

Действительно, не представляется возможным рассчитать значения свойств объектов, поведения, явлений, представленных в текстовой форме. Тем не менее, информационные коммуникации нуждаются в конструктивном определении и вычислении количественных оценок таких свойств, как новизна, ценность, актуальность, которые статистически (на уровне ансамблей, а не отдельных объектов) отражают связи и закономерности предметной области.

Кроме того, любые целесообразные (разумные) абстрактные[6] построения (и текст в том числе) так или иначе, но неминуемо связаны с физической реальностью и, соответственно, эффективность взаимосвязи зависит от адекватности метода сопряжения (интерфейса), т.е. нахождение общности, «общего знаменателя» позволит не только унифицировать и упростить такой механизм, но и построить в достаточной мере абстрактную теорию информационных коммуникаций.

---

[5] В классической физике это принципиально недопустимо. В квантовой механике теоретически допустимо, но связано с множественным (вариантным) заданием *изменяющейся* волновой функции.

[6] По словам Д. Гильберта «Строгая формализация теории предполагает полную абстракцию от смысла». Естественный же язык за счет нечеткости имеет практически неограниченное число степеней свободы, но, фиксируя тот или иной смысл, мы тем самым строим в той или иной степени *формальное* описание системы, которое можно преобразовывать и анализировать, в том числе методами математики.

СПИСОК ЛИТЕРАТУРЫ

1. Поветкина Ю.В. Моделирование как метод лингвистического исследования // Филологические науки. Вопросы теории и практики. – 2012. – № 6(17). – С. 132–136.
2. Минеев А.Б. Некоторые аналогии плазма – биология // Computational nanotechnology. – 2018. – №1. – С. 91–107.
3. Чернавский Д.С., Старков Н.И., Малков С.Ю., Косе Ю.В., Щербаков А.В. Об эконофизике и её месте в современной теоретической экономике // УФН. – 2011. – Т. 181, №7. – С. 767–773.
4. Словохотов Ю.Л. Физика и социофизика. Ч. 3. Квазифизическое моделирование в социологии и политологии. Некоторые модели лингвистики, демографии, математической истории // Проблемы управления. – 2012. – №3. – С. 2–34.
5. Длясин Г.Г. Азбука Гермеса Трисмегиста или молекулярная тайнопись мышления. – М.: Белые альвы, 1998. – 144 с.
6. Лебедев А.Н. Закономерности повторения слов в речи // Психологический журнал. – 1983. – Т. 4, № 5. – С. 11–23.
7. Lau E.F, Phillips C., Poeppel D. A cortical network for semantics: (de)constructing the N400 // Nature reviews. Neuroscience. – 2008. – Vol.9, №12. – P. 920-933.
8. Гуревич И.М. Законы информатики – основа строения и познания сложных систем. – М.: ТОРУС ПРЕСС, 2007. – 400с.
9. Landauer R. The physical nature of information // Physics Letters A. – 1996. – Vol. 217. – P. 188-193.
10. Вайнберг С. Единые теории взаимодействия элементарных частиц // Успехи физических наук. – 1976. – Т. 118, Вып. 3. – С. 505–521.
11. Иванов М.Г. Как понимать квантовую механику. Изд. 2-е, испр. и доп. – Москва – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2015. — 552 с.
12. Елкина В.М., Юдина Л.С. Статистика слогов русской речи // Вычислительные системы. – 1964. – Вып.10. – С. 58–78.
13. Меркулова И.А. Квантитативные характеристики русской лексики на общеславянском фоне // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2014. – №3. – С.100 – 107.
14. Поддубный В.В., Поликарпов А.А. Диссипативная стохастическая динамическая модель развития языковых знаков // Компьютерные исследования и моделирование. – 2011. – Т. 3, №2. – С. 103–124.
15. Шрейдер Ю.А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа) // Проблемы передачи информации. –1967. – Т. 3, Вып.1. – С. 57–63.
16. Miyazima S., Yamamoto K. Measuring the temperature of texts // Fractals. – 2008. Vol.16. – №1. – P. 25–32.
17. Селезнев Г.Д. Природа экспоненциального распределения слов по числу значений // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2007. – №2(1). – С. 42-45.
18. Leopold E. Frequency spectra within word length classes // Journal of Quantitative Linguistics. – 1988. – Vol. 5, №3. – P. 224-231.
19. Келер Р. Синергетическая лингвистика: Структура и динамика лексики. – URL: http://ubt.opus.hbz-nrw.de/volltexte/2007/413/pdf/synling.pdf (дата обращения 12.06.2020).
20. Ji S. Waves as the Symmetry Principle Underlying Cosmic, Cell, and Human Languages // Information. – 2017. – Vol. 8. – P. 1–25;
21. Маслов В.П. Бозе-газ ангармонических осцилляторов и уточнение закона Ципфа // Теоретическая и математическая физика. – 2006. – Т. 148, №3. – С. 495–496.
22. Маслов В.П. Связь распределения Ферми–Дирака с лингвостатистическими распределениями // Математические заметки. – 2017. – Т.101, Вып.4. – С. 531–548.
23. Лебедев А.А., Максимов Н.В., Смирнова Е.В. Семантический сдвиг термина: анализ зависимостей и квантомеханическая модель // Научно-техническая информация. Сер. 2. − 2016. − № 2. − С. 14−22.
24. Salton G., Wong A., Yang C. A vector space model for automatic indexing // Communications of the ACM. − 1975. – Vol.18, №11. − P. 613-620.
25. Урсул А.Д. Природа информации: философский очерк. – Челябинск: Челяб. гос. акад. культуры и искусств, 2010. – 231 с.
26. Берг А.И., Спиркин А.Г. Кибернетика и диалектико-материалистическая философия // В кн.: Проблемы философии и методологии современного естествознания. – М.: Наука, 1973. – С. 139-146.
27. Schaden M. Quantum finance // Physica A: Statistical Mechanics and its Applications. – 2002. – Vol. 316. – P. 511–538.
28. Налимов В.В. Спонтанность сознания. Вероятностная теория смыслов и смысловая архитектоника личности. − М: Прометей, 1989. − 288 с.
29. Ландау Л.Д., Лифшиц Е.М. Теоретическая физика: учебное пособие для вузов. В 10 т. Т. III. Квантовая механика (нерелятивистская теория). 4-е изд., испр. – М.: Наука, 1989. – 768 с.
30. van Rijsbergen C. The Geometry of Information Retrieval. – Cambridge: Cambridge University Press, 2004. – 150 p.
31. Муранивский Т.В. Методы обработки документов на основе использования свойств и закономерностей информации: учеб. пособие. − М.: МГИАИ, 1984. − 103 с.
32. Kitto K., Ramm B., Sitbon L., Bruza P. Quantum Theory Beyond the Physical: Information in Context // Axiomathes. – 2011. – Vol. 21, №2. – P. 331–345.
33. Попов И.И., Романенко А.Г., Сумароков Л.Н. Теоретико-множественное моделирование систем научно-технической информации // Вопросы информационной теории и практики. – 1978. – Вып. 33-34. – С. 16-63.

34. Balkır E. Using Density Matrices in a Compositional Distributional Model of Meaning. Master's thesis. – University of Oxford, 2014. – URL:http://www.cs.ox.ac.uk/people/bob.coecke/Esma.pdf (дата обращения 12.06.2020).
35. Максимов Н.В. Информация и знания: природа, концептуальная модель // Научно-техническая информация. Сер.2. – 2010. – №7. – С.1-10.
36. Aerts D., Gabora L., Sozzo S. Concepts and Their Dynamics: A Quantum-Theoretic Modeling of Human Thought // Topics in cognitive science. – 2013. – Vol. 5, №4. – P. 737–772.
37. Zurek W.H. Quantum Darwinism // Nature Physics. – 2009. – Vol. 5, №3. – P. 181–188.
38. Zurek W.H. Decoherence and the Transition from Quantum to Classica – Revisited // Los Alamos Science. – 2002. – №27. – P. 86–109.
39. Li Y., Cunningham H. Geometric and quantum methods for information retrieval // ACM SIGIR Forum. – 2008. – Vol. 42, №2. – P. 22–32.
40. Krylov J.K. Synergetic Models and Methods in Quantitative Linguistics // Journal of Quantitative Linguistics. – 2002. – Vol. 9, №2. – P. 125-185.
41. Wittek P., Darányi S. Spectral Composition of Semantic Spaces // Proc. Of 5th International Quantum Interaction Symposium (QI, 26-29 June, 2011, Aberdeen, UK). – Berlin: Springer, 2011. – P. 60–70.
42. Sordoni A., He J., Nie J. Modeling latent topic interactions using quantum interference for information retrieval // Proc. of the 22nd ACM international conference on Information & Knowledge Management (CIKM'13, 27 October - 01 November, 2013, San Francisco, USA). – New York: Association for Computing Machinery, 2013. – P. 1197–1200.
43. Zhang P., Song D., Zhao X., Hou Y. Investigating query-drift problem from a novel perspective of photon polarization // Proc. of the 3rd International Conference on the Theory of Information Retrieval (ICTIR, 12-14 September 2012, Bertinoro, Italy). – Berlin: Springer, 2011. – P. 332– 336.
44. Bruza P., Kitto K., Nelson D., McEvoy C. Is there something quantum-like about the human mental lexicon? // Journal of mathematical psychology. – 2009. – Vol. 53, №5. – P. 363– 377.
45. Melucci M., van Rijsbergen K. Quantum Mechanics and Information Retrieval // In: Advanced topics in information retrieval. The Information Retrieval Series. – Berlin: Springer, 2011. – P. 125–155.

Материал поступил в редакцию 03.07.20.

Сведения об авторах

ЛЕБЕДЕВ Александр Анатольевич – ведущий математик кафедры финансового мониторинга Национального исследовательского ядерного университета МИФИ, Москва
e-mail: lebedevalex@live.ru

МАКСИМОВ Николай Вениаминович – доктор технических наук, профессор, профессор кафедры финансового мониторинга Национального исследовательского ядерного университета МИФИ, Москва
e–mail: NV-MAKS@YANDEX.RU

УДК 004.89:510.64:[159.9.07:343.982.4]

Е.Р. Гросс, С.М. Гусакова, Н.В. Огорельцева, А.Н. Охлупина

ДСМ-система психолого-почерковедческих исследований подписи

Описываются принципы разработки, логико-алгебраические методы, а также программная реализация ДСМ-системы, позволяющей решать задачи почерковедческих исследований подписи с привлечением данных психологии. Приводятся результаты экспериментов.

Ключевые слова: *ДСМ-система, психологические характеристики, особенности выполнения подписи, потенциальные гипотезы, дополнительные параметры*

DOI: 10.36535/0548-0027-2020-10-2

ВВЕДЕНИЕ

Почерковедческое исследование подписи – один из наиболее востребованных и вместе с тем сложных и трудоемких видов криминалистической экспертизы. В то время как рукописный текст встречается все реже, подпись остается объектом, удостоверяющим документ и придающим ему юридическую силу. Еще более производство почерковедческой экспертизы усложняет наблюдающаяся в последнее время тенденция к упрощению подписи. Все это приводит к выводу о необходимости применения компьютерных технологий в этой области криминалистических исследований. Такие технологии уже существуют, но, в основном, работают в автоматическом режиме. Автоматические системы для идентификации подписи (см., например, [1-3]) не используются в судебной практике, так как либо рассматривают только динамическую подпись, выполненную на планшете, либо не объясняют полученный результат, в то время как судебная экспертиза предполагает обязательную аргументацию предъявляемого экспертом вывода. Поэтому очевидно, что в помощь эксперту должна быть создана автоматизированная система, позволяющая аргументировать сделанный ею вывод.

Проблемой в почерковедении является также определение факторов, влияющих на формирование подписи. Психологи и криминалисты не отрицают связи между психологическими особенностями личности и признаками почерка, проявляющимися в рукописном тексте и в подписи, но, как справедливо замечено в [4]: «разработок в научной литературе в данной области сегодня представлено крайне мало». При этом существующие исследования направлены на определение характеристики психологического портрета по почерку (см., например, [5, 6]), в то время как каузальность в этом вопросе имеет обратное направление. Кроме того, практически нет компьютерных систем, выполняющих анализ данных для решения этой задачи.

Поэтому необходимая автоматизированная система, состоящая из двух подсистем, была создана как интеллектуальная ДСМ-система поддержки почерковедческих исследований. Первая подсистема, решающая задачу идентификации подписи и позволяющая исследовать информативность значений частных признаков подписи и групп частных признаков, описана в [7]. Вторая подсистема нацелена на решение задачи выявления влияния психологических характеристик исполнителя подписи на ее особенности.

Опишем основные принципы, заложенные в реализацию ДСМ-системы, ее теоретические и программные средства, проведенные эксперименты.

ОСНОВНЫЕ ПРИНЦИПЫ РЕАЛИЗАЦИИ ДСМ-СИСТЕМЫ

Интеллектуальная система поддержки научных исследований в области почерковедения основывается на ДСМ-методе, направленном на получение на основе информации, содержащейся в базе фактов и базе знаний, нового знания в виде эмпирических закономерностей вида «подобъект V является причиной проявления эффекта Y» и «объект X проявляет эффект Y». ДСМ-метод включает ДСМ-рассуждение и ДСМ-исследование [8, 9].

При проведении ДСМ-рассуждения осуществляется синтез познавательных процедур: индукции, аналогии и абдукции. В результате индуктивного вывода порождаются гипотезы о причинах проявления эффектов, с помощью процедур аналогии – гипотезы о проявлении или не проявлении эффекта, абдуктив-

ное рассуждение позволяет определить – на достаточном ли основании получены указанные гипотезы.

При проведения ДСМ-исследования гипотезы проверяются на устойчивость при последовательном расширении базы фактов, в результате чего выделяются эмпирические законы и тенденции.

Конструирование интеллектуальной системы, основанной на ДСМ-методе поддержки научных исследований, начинается с создания модели предметной области, включающей аксиомы предметной области, язык представления данных, операцию сходства, необходимую для проведения индуктивного рассуждения, и отношение вложения, участвующее в рассуждении по аналогии. Модель предметной области влияет на формулировку предикатов, реализующих правдоподобные рассуждения, которые используют индукцию и аналогию, поэтому существуют различные варианты ДСМ-метода [10]. Такой подход позволяет создавать интеллектуальные системы, учитывающие особенности предметной области и решаемых в них задач, а также давать объяснение полученных результатов.

В соответствии с решаемыми задачами система разделена на две подсистемы – «Подпись» и «Психология», связанные через респондентов, данные о которых содержатся в системе.

Первая подсистема включает данные о признаках подписей, причем в базе фактов содержатся как описание подлинных подписей, выполненных респондентами, так и признаки поддельных подписей, выполненных от лица этих респондентов другими лицами. В подсистеме есть ДСМ-решатель, реализующий ДСМ-рассуждение и позволяющий находить гипотезы о причинах, вызывающих эффект[1], а также гипотезы о подлинности или поддельности исследуемого образца подписи. Важной частью подсистемы «Подпись» является подсистема ввода, позволяющая эксперту однозначно описывать одинаковые признаки подлинных и поддельных образцов подписей, что существенно для автоматизации почерковедческих исследований [11], вводить новые признаки и формировать базу фактов. Архитектура подсистемы «Подпись» позволяет проводить исследования информативности значений и групп частных признаков в автоматическом режиме. Возможность такого исследования обеспечивает наличие в базе данных признаков не только подлинных, но и поддельных подписей. Необходимо отметить, что информативность групп признаков в почерковедении ранее не исследовалась.

В базе фактов второй подсистемы, созданной для решения задачи выявления связи психологических характеристик и особенностей подписи, через респондента связаны психологические характеристики, хранящиеся в этой подсистеме, с признаками подписи из подсистемы «Подпись». Эти характеристики представляют собой заполненные респондентами специально подобранные психологические опросники: структуры темперамента; черт характера В.М. Русалова [12]; «Самочувствие, активность, настроение»; уровня агрессивности Басса-Перри.

Опросник структуры темперамента выявляет такие психологические черты, как стремление к умственному или физическому труду, лидерству, а также характеризует быстроту переключения с одного вида деятельности на другой – моторно-двигательных и речедвигательных актов, – и предоставляет возможность оценивать уровень эмоциональности.

Опросник черт характера позволяет определить *деакцентуации*, *норму* или *акцентуации* испытуемого, т.е. выяснить – проявляются ли некоторые черты характера преувеличенно в ущерб другим.

Опросник «Самочувствие, активность, настроение» фиксирует психологическое состояние в момент его заполнения, которое должно осуществляться одновременно с выполнением образцов подписи. Как отмечается в [13]: «Графологический анализ отражает не только индивидуальные психические особенности личности, но и ее морально-психологическое состояние на текущий момент времени».

Опросник Басса-Перри диагностирует агрессивность респондента, выявляя такие характеристики как *физическая агрессия*, *гнев* и *враждебность*.

Предполагается, что психологические черты, выявляемые с помощью опросников, влияют на особенности подписи.

Подсистема «Психология» включает подсистему ввода, позволяющую респонденту в удобном формате заполнять ответы на вопросы, содержащиеся в опросниках, автоматически подсчитывать баллы по каждой шкале опросника и присваивать значения в соответствии с интервалом, в который попал каждый полученный бал.

При построении ДСМ-системы разработка модели предметной области происходит в тесном контакте с экспертом и пользователями системы, а полученные аксиомы предметной области влияют на выбор варианта ДСМ-метода.

Сложная структура частных признаков нашла отражение в языке представления данных, разная информативность признаков определила выбор варианта ДСМ-метода для атрибутов с весами для решения задачи идентификации подписи, необходимость учитывать не только входящие, но и не входящие в исследуемый образец признаки, содержащиеся в подлинных и поддельных образцах подписи, повлияла на особенности формулировки предиката вывода по аналогии. При решении идентификационной задачи набор истинностных значений гипотез о доопределении соответствует вариантам выводов, один из которых должен представить эксперт в заключении.

В связи с тем, что не выработана обоснованная научно гипотеза о том, какие психологические факторы определяют особенности подписи, а в психологических опросниках отражены разные стороны психологии личности, для исследования привлекаются несколько психологических опросников с возможностью их замены. Система должна предоставить средства для различных вариантов работы с выбранными опросниками.

[1] Эффектом в рассматриваемой задаче является подлинность подписи. Термин *причина*, принятый в ДСМ-методе, используется в том смысле, что особенности подписи, выделенные ДСМ-решателем, позволяют сделать вывод о её подлинности.

Особенности структуры данных и содержание базы фактов в подсистеме «Психология» требуют привлечения специальных логико-алгебраических средств.

Программная реализация системы, использующая современные средства создания интеллектуальных систем, должна учитывать открытость системы и предоставлять возможность расширения ее функциональности, а также баз фактов и знаний.

Теоретические принципы, выбор варианта ДСМ-метода, формулировка предикатов для реализации индуктивного вывода и вывода по аналогии с учетом особенностей модели предметной области для задачи идентификации подписи и исследований информативности признаков были изложены в [7]. Средства, основанные на указанных принципах и реализованные в подсистеме «Психология», изложены далее.

ЛОГИКО-АЛГЕБРАИЧЕСКИЕ СРЕДСТВА, РЕАЛИЗОВАННЫЕ В ПОДСИСТЕМЕ «ПСИХОЛОГИЯ»

Для выявления зависимости особенностей подписи от психологических характеристик её исполнителя, как было уже отмечено, нужно использовать разные психологические опросники, состоящие из некоторого количества вопросов, ответы на которые выбираются из заданного набора и оцениваются числовыми баллами. Список вопросов вместе со спектром баллов описывает шкалы, соответствующие психологическим характеристикам, выявляемым опросником. Сумма балов, полученных за ответы на относящиеся к определенной шкале опросника вопросы, может принимать значения в диапазоне от 0 (нуля) до максимального суммарного балла.

Диапазон разбивается на части, которым присваиваются имена. В подсистеме «Психология» такое разбиение осуществляется одним из двух способов: *стандартным* (как принято у психологов), т.е. делением на три уровня (низкий, средний, высокий) и *дробным* – дополнительным членением каждого стандартного уровня на два или три подуровня в зависимости от величин диапазонов в разных опросниках.

Таким образом, в результате заполнения респондентом опросника образуется кортеж, состоящий из названий шкал (психологических характеристик) с их значениями, где значение – это имя части, в которую попал суммарный по ответам бал по каждой шкале.

Обозначим через T^j множество кортежей, состоящих из названий и всевозможных значений шкал опросника с номером j. Среди значений может быть пустое значение – λ, но надо иметь в виду, что поскольку при заполнении опросников должен быть дан ответ на каждый вопрос, в базе фактов значение λ не встречается[2]. Однако его присутствие в базе знаний не исключено в результате применения операции сходства.

При наличии нескольких опросников возможны различные варианты исследования. Первый – работать с каждым опросником отдельно. Второй – выбрав один из опросников как главный, одну или несколько психологических характеристик других опросников использовать для разделения на (+)- и (–)-примеры. Третий вариант – использовать ДСМ-метод с параметром ситуации [14], расширив его на случай нескольких дополнительных параметров.

В первом и втором случаях работы с опросниками база фактов (БФ) состоит из примеров вида $J_{<\nu,\,0>}(X_n \Rightarrow_1 Y_n)$, где X_n – кортеж значений шкал, полученных в результате заполнения выбранного опросника респондентом с номером n, Y_n – признаки подписи респондента с номером n, $X_n \in T^j$, $Y_n \in \mathscr{B}(H)$, $\mathscr{B}(\cdot)$ – булеан, H – множество всех признаков подписей со всеми значениями. Оператор Россера–Тюркетта $J_{\langle \nu,k \rangle}\varphi$ равен истинностному значению двузначной логики «истина», если $\nu[\varphi] = \langle \nu, k \rangle$, значению «ложь», если $\nu[\varphi] \neq \langle \nu, k \rangle$, где $\nu[\varphi]$ – функция оценки, $\langle \nu, k \rangle$ – представляет «внутренние» истинностные значения фактов и гипотез, $\nu \in \{1, -1, 0, \tau\}$ (“фактическая истина”, “фактическая ложь”, “фактическое противоречие”, “неопределенность”, соответственно), k – номер шага ДСМ-рассуждения. Для начального состояния базы фактов $k = 0$.

При работе с расширенным ситуационным методом пример базы фактов имеет вид: $J_{<\nu,\,0>}((X_n^1; X_n^2, X_n^3, X_n^4) \Rightarrow_1 Y_n)$. Здесь X_n^1 – кортеж значений шкал, полученных в результате заполнения выбранного опросника респондентом с номером n; X_n^2, X_n^3, X_n^4 – кортежи значений шкал остальных опросников, заполненных этим респондентом. $(X_n^1; X_n^2, X_n^3, X_n^4)$ можно рассматривать как расширенный объект, при этом X_n^2, X_n^3, X_n^4 являются дополнительными параметрами, влияние которых на зависимость между X_n^1 и Y_n предстоит выяснить.

Следует заметить, что почерковеды выделяют общие и частные признаки подписи. Общие признаки характеризуют подпись в целом, они не связаны с конкретными буквами. Набор этих признаков должен присутствовать в описании подписи полностью. Из частных признаков наличествуют только те, которые проявились в соответствии с присутствующими в подписи буквами. Общие признаки описываются кортежем, каждый элемент которого – номер признака и номер значения этого признака. Частные признаки имеют иерархическую структуру, которая отражает группу признаков, одну или две буквы/безбуквенные элементы возможно с конкретизацией части буквы, а также значение признака, тоже возможно с конкретизацией. Частные признаки занумерованы в соответствии с этой структурой, что позволяет использовать разные варианты операции существенного сходства в зависимости от решаемой задачи. В качестве Y_n могут быть взяты общие и частные признаки, только общие или только частные признаки, а также обобщенные частные признаки, относящиеся к одинаковым элементам разных букв или относящиеся только к группе признаков независимо от буквы.

Если система работает с одним опросником, отношение $\Rightarrow_1$ между T^j – множеством кортежей значений шкал выбранного опросника и $\mathscr{B}(H)$ – булеаном множества значений общих и/или частных признаков читается как «респондент с номером n и психологическими характеристиками X_n имеет признаки подписи Y_n».

[2] Значение λ – это отсутствие значения для данной характеристики. Не следует путать со значением, соответствующим числу баллов, равным нулю.

Описание признаков подписи не может быть разделено на отдельные признаки так, что про каждый можно сказать, присутствует он в подписи респондента или нет. Весь набор признаков подписи рассматривается в совокупности. Из этого следует, что представление данных в решаемой задаче соответствует неатомистическому варианту ДСМ-метода, предполагающему разделение базы фактов на положительную ($БФ^+$) и отрицательные ($БФ^-$) части целиком по всему примеру. В рассматриваемом случае это значит, что пример с номером n, психологическими характеристиками X_n и признаками подписи Y_n целиком относится либо к $БФ^+$, либо к $БФ^-$, а не так, как в атомистическом случае, когда один и тот же объект по одним свойствам относится к положительным примерам, а по другим – к отрицательным. Неатомистический вариант требует введение фактора, разделяющего примеры на положительные и отрицательные. Подсистема «Психология» предоставляет возможность вводить различные факторы для такого разделения. Ими могут служить подлинность подписи, значение одной или нескольких шкал опросника, используемого в исследовании в качестве дополнительного, пол респондентов или другие факторы.

При таком подходе появляется возможность исследовать влияние разделяющего фактора на получаемые из примеров базы фактов зависимости. Так, если при разделяющем факторе *пол респондента* полученная в $БФ^+$ зависимость $V \Rightarrow_2 W$, где $V \in T^j$ $W \in \mathscr{B}(H)$, $\Rightarrow_2$ – отношение между T^j и $\mathscr{B}(H)$ «психологические характеристики, выражаемые подмножеством значений шкал опросника влияют на проявление признаков подписи», получается и в $БФ^-$, то эта зависимость с полом респондента не связана. Если в $БФ^+$ получена зависимость $V \Rightarrow_2 W_1$, а $БФ^-$ зависимость $V \Rightarrow_2 W_2$ и $W_1 \neq W_2$, то это значит, что психологические характеристики, проявленные в V, поразному влияют на особенности подписи у мужчин и женщин. То же можно сказать, если в $БФ^+$ и в $БФ^-$ получены зависимости $V_1 \Rightarrow_2 W$ и $V_2 \Rightarrow_2 W$ и $V_1 \neq V_2$.

Для нахождения зависимостей между психологическими характеристиками и особенностями подписи была выбрана однородная стратегия ДСМ-метода с предикатами простого сходства. Ее применение предполагает выбор между прямым и обратным вариантами. Содержательно *прямой* метод означает, что сходство объектов влечет сходство эффектов, а *обратный* – сходство эффектов есть следствие сходства объектов [15], т.е. эти варианты отличаются каузальной направленностью. Логически это выражается в формулировке условия исчерпываемости, выполнение которого предполагает рассмотрение всех примеров, дающих некоторый результат операции сходства. Для прямого метода – это сходство объектов, для обратного – эффектов. Прямой предикат сходства выбирается, когда более информативно описание объектов, обратный – в случае более информативного описания эффектов. Однако, поскольку описание и психологических характеристик, и признаков подписи достаточно информативно, возникает трудность в выборе между прямым и обратным вариантами стратегии ДСМ-метода.

При работе с дополнительными параметрами (опросниками) добавляются варианты с исчерпываемостью по одному или нескольким дополнительным параметрам, имеющим существенное влияние, если исчерпываемость по ним и по основному опроснику или по свойствам совпадает. Это еще больше усложняет выбор методов решения и их интерпретацию.

В результате применения прямого или обратного предикатов сходства индуктивный вывод дает разные гипотезы. В некоторых случаях они совпадают. Способ, позволяющий обозреть гипотезы и прямого, и обратного методов, был описан в [16]. Он заключается в нахождении потенциальных гипотез.

Потенциальная гипотеза, порождаемая примерами с номерами $\{1,\ldots,k\}$ – это пара (C, Q), где:

$$((C = (X_1 \sqcap \ldots \sqcap X_k)) \,\&\, (Q = (Y_1 \sqcap \ldots \sqcap Y_k)) \,\&$$
$$\&\; (C \neq \varnothing \,\&\, Q \neq \varnothing) \,\&$$
$$\&\forall X_m \,\forall Y_m \,((C \sqcap X_m \neq C) \vee (Q \sqcap Y_m) \neq Q)) \,\&$$
$$\&(m \notin \{1, \ldots, k\}),$$

X_i – кортеж значений шкал опросника респондента с номером i, Y_i – набор признаков подписи этого респондента, $\sqcap$ – операция сходства. На данный момент в системе операция сходства для кортежей определена следующим образом: если $X_1 = (x_{11},\ldots,x_{1n})$ и $X_2 = (x_{21},\ldots,x_{2n})$, то $X_1 \sqcap X_2 = (x_1,\ldots,x_n)$, где $x_i = x_{1i}$, при $x_{1i} = x_{2i}$ и $x_i = \lambda$ при $x_{1i} \neq x_{2i}$ (другой вариант определения операции сходства для кортежей приведен в [17]). Операция сходства для признаков подписи определяется в зависимости от содержания множества H. Если в H только общие признаки, то $\sqcap$ – операция сходства для кортежей, если H состоит из частных признаков, то, во-первых, операция сходства определяется на каждом признаке отдельно, т.е. $Y_1 \sqcap Y_2 = \{ Y_1/q \sqcap Y_2/q \mid Y_1/q, Y_2/q$ – проекции объектов Y_1, Y_2 на признак с именем $q\}$, во-вторых, операция $\sqcap$ в зависимости от решаемой задачи может быть определена на разных уровнях – уровне значения признака (без учета буквы), уровне элемента буквы (в случае обобщенных признаков), с учетов всех уровней. Поскольку все уровни описания признака учтены в иерархической нумерации признаков, операция сходства имеет непустой результат при совпадении значений уровней иерархии в этой нумерации от начала до выбранного в определении операции сходства уровня.

Если $\forall X_m \,\forall Y_m \,((C \sqcap X_m = \varnothing) \vee (Q \sqcap Y_m = \varnothing)) \,\&\, (m \notin \{1, \ldots, k\})$, то (C, Q) – минимальная потенциальная гипотеза.

Обозначим через P множество всех потенциальных гипотез для БФ. Из определения потенциальной гипотезы следует, что P содержит все гипотезы прямого и обратного методов. Однако в этом множестве могут содержаться и пары (C, Q), реальными гипотезами не являющиеся.

Если представить базы фактов $БФ^+$ и $БФ^-$ как пространства сходства $\Sigma^\sigma = \langle \boldsymbol{U}^\sigma, Cov^\sigma \rangle$, где $\sigma \in \{+,-\}$, $\boldsymbol{U}^\sigma = \cup_k (X_k \times Y_k)$, $Cov^\sigma = \{(X_k \times Y_k)\}$ – покрытие множества $\boldsymbol{U}^\sigma$, $(X_k \Rightarrow_1 Y_k) \in БФ^\sigma$, $k = 1,\ldots,N$, N – количество примеров в $БФ^\sigma$, то множество потенциальных

гипотез совпадает с системой замыканий Галуа, ассоциированной с Σ^{σ}.

Для нахождения множества потенциальных гипотез разделим примеры базы фактов $БФ^{\sigma}$ на части, состоящие из кортежей шкал опросника и совокупностей признаков подписи. Обозначим эти части через $БФ^{\sigma}(T^{j})$ и $БФ^{\sigma}(H)$. Найдем все результаты сходства в $БФ^{\sigma}(T^{j})$ и $БФ^{\sigma}(H)$.

Введем следующие обозначения:

$\check{N}_1$ – множество наборов имен респондентов, породивших все сходства в $БФ^{\sigma}(T^{j})$;

$\check{T}^{j}$ – множество всех результатов сходств в $БФ^{\sigma}(T^{j})$;

$\check{N}_2$ – множество наборов имен респондентов, породивших все сходства в $БФ^{\sigma}(H)$;

$\check{H}$ – множество всех результатов сходств в $БФ^{\sigma}(H)$.

Тогда $\check{T}^{j} = \{(N':C) | C = \Pi_{i \in N'} X_i\}$,

где $N' \in \check{N}_1$, $\check{H} = \{(N'':Q) | Q = \Pi_{k \in N''} Y_k\}$, $Q \in \mathfrak{B}(H)$, $N'' \in \check{N}_2$. Множество $\{(\tilde{N}: (C,Q)) \mid \tilde{N} \in \check{N}_1 \cap \check{N}_2, C = \Pi_{i \in \tilde{N}} X_i, Q = \Pi_{j \in \tilde{N}} Y_j\}$ обозначим через $(\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$.

Можно показать, что множество P всех потенциальных гипотез базы фактов совпадает с $(\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$. Действительно, пусть $p \in P$ и $p = (N':(C, Q))$, $N' = 1, ..., k$. По определению для потенциальной гипотезы $p = (C = (X_1 \Pi \dots \Pi X_k)) \& (Q = (Y_1 \Pi \dots \Pi Y_k)) \& (C \neq \varnothing \& Q \neq \varnothing)$ имеет место один из вариантов:

1. $(C = (X_1 \Pi \dots \Pi X_k)) \& (Q = (Y_1 \Pi \dots \Pi Y_k)) \& (C \neq \varnothing \& Q \neq \varnothing) \& \forall X_m \forall Y_m ((C \Pi X_m \neq C) \& (Q \Pi Y_m \neq Q)) \& (m \notin \{1, ..., k\})$.

2. $(C = (X_1 \Pi \dots \cap \Pi X_k)) \& (Q = (Y_1 \Pi \dots \Pi Y_k \Pi Y_{k+J})) \& (C \neq \varnothing \& Q \neq \varnothing) \& \forall X_m \forall Y_m (C \Pi X_m \neq C) \& (m \notin \{1, ..., k, k+J\}) \& (J = j_1, \dots, j_s)$.

3. $(C = (X_1 \Pi \dots \Pi X_k \Pi X_{k+I})) \& (Q = (Y_1 \Pi \dots \Pi Y_k)) \& (C \neq \varnothing \& Q \neq \varnothing) \& \forall X_m \forall Y_m (Q \Pi Y_m \neq Q) \& (m \notin \{1, ..., k, k+I\}) \& (I = i_1, \dots, i_r)$.

4. $(C = (X_1 \Pi \dots \Pi X_k \cap X_{k+I})) \& (Q = (Y_1 \Pi \dots \Pi Y_k \cap Y_{k+J})) \& \forall i \in I \forall j \in J (i \neq j) \& \forall X_m \forall Y_m ((C \Pi X_m \neq C) \vee (Q \Pi Y_m \neq Q)) \& (m \notin \{1, ..., k\})$.

<u>В случае 1</u> $((N': C) \in \check{T}) \& (N': Q) \in \check{H})$, $N' \in \check{N}_1 \cap \check{N}_2$. Тогда $p = (N': (C,Q))$, $N' \in \check{N}_1 \cap \check{N}_2$, $C = \Pi_{i \in N'} X_i$, $Q = \cap_{j \in N'} Y_j$.

<u>В случае 2</u> $((N': C) \in \check{T}) \& (N'': Q) \in \check{H})$, $N' = 1, ..., k$, $N'' = 1, ..., k, k+J$. Здесь $p = (N' \cap N'': (C, Q))$, $N' \cap N'' \in \check{N}_1 \cap \check{N}_2$, $N' \cap N'' = N'$, $C = \Pi_{i \in N'} X_i$, $Q = \Pi_{j \in N'} Y_j$.

<u>В случае 3</u> $((N': C) \in \check{T}) \& (N'': Q) \in \check{H})$, $N' = 1, ..., k, k+I$, $N'' = 1, ..., k$. В этом случае $p = (N' \cap N'': (C, Q))$, $N' \cap N'' \in \check{N}_1 \cap \check{N}_2$, $N' \cap N'' = N''$, $C = \Pi_{i \in N''} X_i$, $Q = \Pi_{j \in N''} Y_j$.

<u>В случае 4</u> $((N':C) \in \check{T}) \& (N'': Q) \in \check{H})$, $N' = 1, ..., k, k+I$, $N'' = 1, ..., k, k+J$. $p = (N' \cap N'': (C, Q))$, $N' \cap N'' \in \check{N}_1 \cap \check{N}_2$, $N' \cap N'' = 1, ..., k$, $C = \Pi_{i \in N' \cap N''} X_i$, $Q = \Pi_{j \in N' \cap N''} Y_j$.

Следовательно,

$\forall p \in P$ $(p \in (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}})$ и $P \subseteq (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$.

Пусть $(N' \cap N'':(C, Q)) \in (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$. Если $N' \subseteq N''$, либо $N'' \subseteq N'$, то $(N' \cap N'':(C, Q)) = (N': (C,Q))$, $N' \in \check{N}_1 \cap \check{N}_2$, $C = \Pi_{i \in N'} X_i$, $Q = \Pi_{j \in N'} Y_j$ или $(N' \cap N'':(C, Q)) = (N'': (C,Q))$, $N'' \in \check{N}_1 \cap \check{N}_2$, $C = \Pi_{i \in N''} X_i$, $Q = \Pi_{j \in N''} Y_j$. Очевидно, что в этом случае имеет место один из вариантов 1, 2 или 3, приведенных выше.

Если $\neg((N' \subseteq N'') \vee (N'' \subseteq N'))$ и $N' \cap N'' \neq \varnothing$, то имеет место вариант 4.

Таким образом, $(N' \cap N'':(C, Q)) \in P$ и $(\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}} \subseteq P$. Из $P \subseteq (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$ & $(\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}} \subseteq P$ следует $P = (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$.

Нетрудно заметить, что в случае 1 потенциальная гипотеза p удовлетворяет прямому и обратному предикатам простого сходства, в случаях 2 и 3 – прямому и обратному предикатам соответственно, а в случае 4 p не является реальной гипотезой. Отсюда следует, что если $(N' \cap N'':(C, Q)) \in (\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$ и $N' = N''$, то $(N' \cap N'':(C, Q))$ – гипотеза и прямого, и обратного методов одновременно, если $N' \subset N''$, $(N' \cap N'':(C, Q))$ – гипотеза прямого метода, если $N'' \subset N'$, $(N' \cap N'':(C, Q))$ – гипотеза обратного метода, если $\neg((N' \subseteq N'') \vee (N'' \subseteq N'))$ и $N' \cap N'' \neq \varnothing$, то $(N' \cap N'':(C, Q))$ реальной гипотезой не является. Это замечание дает указание к реализации алгоритма классификации множества $(\check{N}_1 \cap \check{N}_2)_{\check{T}\check{H}}$ всех потенциальных гипотез по методу. Такая классификация позволяет сделать выбор между прямым и обратным вариантами ДСМ-метода.

Помимо варианта работы с каждым опросником отдельно можно, выбрав один опросник как основной, использовать для разделения на положительные и отрицательные примеры одну или несколько шкал других опросников.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Программная реализация ДСМ-системы предполагает разработку комплекса программ, позволяющего хранить и обрабатывать данные и проводить необходимые эксперименты и исследования с применением ДСМ-метода.

К программной реализации описываемой ДСМ-системы были предъявлены следующие требования:

- интерфейс должен быть интуитивно понятным эксперту-почерковеду, эксперту-психологу, респондентам и представлен в привычных им категориях;
- подсистема ввода информации о подписи должна учитывать особенности структуры заполняемых признаков подписи;
- поскольку с ДСМ-системой могут работать несколько пользователей, должна быть реализована возможность ее одновременного использования с разных компьютеров независимо от их местонахождения и программного обеспечения.

Исходя из выдвинутых требований, было решено разрабатывать веб-приложение и хранить необходимую информацию в удаленной базе данных.

Реализация системы имеет двухуровневую структуру:

1) база данных, расположенная на удаленном сервере – хранит информацию о респондентах, признаках подписи и опросниках;

2) интерфейс и логика ДСМ-метода – содержится в веб-приложении.

После проведенного исследования технологий, соответствующих этим требованиям, в качестве средства разработки был выбран веб-фреймворк (программная платформа, определяющая структуру системы) *Django*, с использованием языков программирования *Python*, *JavaScript* и языков разметки *HTML* и *CSS* (+ *CSS*-фреймворк *Bootstrap*), а в качестве системы управления базами данных – *MySQL*.

Фреймворк *Django* используется для создания клиент-серверных веб-приложений. Его выбор определен тем, что он удовлетворяет требованиям, предъявляемым решаемыми задачами и обеспечивает удобство работы с сущностями базы данных. Это определяется встроенным в фреймворк механизмом описания сущностей, называемым Моделью. Параметры Модели сохраняются в базе данных с помощью технологии *ORM* (*Object-Relation Mapping*), позволяющей работать с сущностями базы данных в концепции объектно-ориентированного программирования.

Для развертывания приложения на продуктовой среде (окружение, в котором развернуто программное обеспечение, где продукт доступен пользователям) необходимо было использовать веб-сервер и сервер приложений. Инфраструктура *Django* уже включает сервер разработки, предназначенный для работы в локальном окружении, но он не способен выдерживать нагрузку, возникающую при реальной работе, когда веб-приложение используется одновременно многими пользователями. Поэтому необходимо было использовать веб-сервер, обладающий высокоэффективными механизмами обработки соединений и удобными функциями безопасности. В качестве такого сервера в представленной архитектуре выступает *Nginx*. Однако он не может напрямую работать с *Python* (*Django*) приложениями, поэтому должен проксировать (перенаправлять) запросы на сервер, который умеет это делать. В роли такого сервера, который работает непосредственно с Django-приложением выступает *WSGI* сервер *Gunicorn*.

В архитектуре присутствует также супервайзер, т.е. менеджер систем и служб, который следит за тем, чтобы *Gunicorn* был запущен и, в случае падения, перезапущен.

Для организации интерфейса пользователя был применен встроенный в *Django* язык шаблонов *DTL* (*Django Template Language*). Он позволяет использовать *html*-шаблоны для генерации конечных *html*-страниц, что дает возможность динамически формировать интерфейс пользователя в зависимости от контекста.

Логика генерации веб-страницы и обработки *GET/POST* запросов к ней прописывается внутри функции представления. Функция представления, или коротко представление (*view*) – это функция языка *Python*, которая принимает на вход веб-запрос и возвращает веб-ответ. Каждое представление является простой функцией языка *Python* или методом в случае представлений, основанных на классах объектно-ориентированного программирования.

Интерфейс подсистемы ввода удовлетворяет потребностям эксперта, работающего с конкретной подсистемой. Так, например, ввод пользователем общих и частных признаков подписи реализован посредством связанных между собой выпадающих списков. Тем самым учитывается их иерархическая структура.

Для корректной работы системы было реализовано несколько алгоритмов на языке *Python*:

- автоматический подсчет суммарных баллов по шкалам после заполнения опросников респондентами;
- идентификация образца подписи;
- определение информативности признаков и групп признаков подписи;
- автоматическое заполнение технических таблиц базы данных;
- нахождение сходств по опросникам и по признакам подписей (за основу был взят алгоритм Норриса [18]);
- разделение базы фактов на «+»- и «-»-примеры;
- нахождение потенциальных гипотез;
- классификация потенциальных гипотез.

При создании удобного пользовательского интерфейса нужно было обеспечить необходимую функциональность системы: удобство заполнения, редактирования, а также анализа базы фактов. Пользователи – эксперт-почерковед и эксперт-психолог, респонденты, заполняющие опросники, работают в части системы, соответствующей их специализации.

ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ И ИХ АНАЛИЗ

Тестирование разработанной интеллектуальной ДСМ-системы подтвердило ее работоспособность и позволило сделать некоторые предварительные выводы. Эксперименты, проведенные с использованием опросника структуры темпарамента (ОСТ) и разделения на $БФ^+$ и $БФ^-$ по признаку пола, показали, что нужно выбирать вариант ДСМ-метода с предикатами прямого простого сходства, так как гипотез, удовлетворяющих обратному прямому предикату сходства, очень мало.

В эксперименте с использованием дробного разбиения со значениями шкал из множества: {*Низкий-низкий*, *Низкий-высокий*; *Средний-низкий*, *Средний-высокий*; *Высокий-низкий*, *Высокий-высокий*} в $БФ^-$ для респондентов с номерами n_1 и n_2 была найдена зависимость, представленная в таблице.

Результат эксперимента с опросником структуры темперамента

Респонденты	Шкалы и их значения	Признаки подписи
n_1, n_2 из $БФ^-$	Предметная эргичность: *Высокий-высокий* Социальная эргичность: *Высокий-высокий* Пластичность: *Высокий-высокий* Темп: *Высокий-высокий* Социальная эмоциональность: *Низкий-низкий*	Степень выработанности: *от средней до выше средней* Координация: *от средней до выше средней* Темп: *выше среднего* Преобладающая протяженность движений по вертикали: *от средней до большой*

Поскольку в $БФ^+$ такой зависимости не обнаружено, последнюю можно рассматривать как гипотезу $V \Rightarrow_2 W$ в $БФ^-$, где V – результат операции сходства значений шкал опросника структуры темперамента для респондентов n_1 и n_2 , W – результат операции сходства признаков подписей этих респондентов. Причем это гипотеза как прямого, так и обратного методов.

Респондентов, обладающих психологическими чертами из этой гипотезы, можно охарактеризовать как активных, деятельных, подвижных. Они легко переключаются с одного рода деятельности на другой, у них быстрые моторно-двигательные акты. Особенности подписей этих респондентов характеризуются подвижностью, убыстрением двигательных актов, отсутствием устойчивости. Такая зависимость выглядит логично.

При работе с опросником «Самочувствие, активность, настроение» у этих же респондентов все три шкалы имеют значение *Высокий – высокий.* Однако, чтобы определить, влияет ли на полученную при работе с ОСТ зависимость самочувствие, активность и настроение, надо получить образцы подписей и ответы на вопросы по опроснику «Самочувствие, активность, настроение» у респондентов, когда значения шкал этого опросника у них даст другие результаты.

В экспериментах с опросником уровня агрессивности Басса-Перри, имеющим три шкалы – агрессивность, гнев, враждебность – при разделении на $БФ^+$ и $БФ^-$ по фактору значение шкалы *возбудимость* опросника черт характера найдены две зависимости, полученные и прямым, и обратным методами. У респондентов с низкой возбудимостью значения шкал агрессии и враждебности среднее, а у респондентов со средней возбудимостью значения этих шкал низкие. При этом и у тех, и у других выделены четыре признака, три из которых имеют одинаковые значения, а четвертый признак – *форма основания подписи* – разные значения. Из этого можно сделать вывод, что уровень агрессии и враждебности оказывает влияние именно на этот признак. Кроме того, уровень возбудимости тоже может оказывать влияние на выработку именно такого значения этого признака.

Можно разделить респондентов по типам темперамента, обобщив значения шкал опросника структуры темпарамента, и искать зависимости в такой форме.

Проведенные эксперименты послужили для тестирования ДСМ-системы и указали направления исследований, которые можно проводить для решения поставленной задачи.

Малый объем обучающей выборки не позволяет рассматривать полученные зависимости как окончательный вариант. Для нахождения зависимостей, которые можно назвать эмпирическими законами или тенденциями, необходимо провести последовательное расширение базы фактов и выявить в этой последовательности зависимости, сохраняющие истинностную оценку, следовательно, провести ДСМ-исследование.

Напомним, что ДСМ-рассуждение включает нахождение гипотез о причинах эффектов и гипотез о доопределении объектов, про которые неизвестно, относятся они к $БФ^+$ или к $БФ^-$. Однако, следует заметить, что при решении задач с помощью подсистемы «Психология» интерес, в основном, представляет нахождение гипотез о причинах, т.е. выявление закономерностей влияния психологических характеристик на особенности подписи. Нахождение гипотез о доопределении важно в том случае, когда в качестве фактора, разделяющего базу фактов на положительную и отрицательную части, выбрана подлинность образцов подписи. В этом случае гипотезы о доопределении могут дать дополнительные аргументы *за* или *против* подлинности исследуемой подписи при решении задачи ее идентификации. Это подчеркивает единство двух частей описываемой интеллектуальной системы и дает возможность эксперту делать более обоснованный вывод при проведении почерковедческой экспертизы подписи.

Отсутствие обоснованных предположений о том, какие именно стороны психологического портрета личности влияют на особенности подписи дает основания предположить, что самый перспективный способ работы с несколькими опросниками – использование ситуационного варианта ДСМ-метода с расширением его на несколько дополнительных параметров. Это позволяет рассмотреть различные условия исчерпываемости в предикатах сходства. Работа в таком режиме должна позволить определить: оказывает ли влияние каждый из дополнительных параметров на особенности подписи и сопоставимо ли это влияние с влиянием основного опросника. Может быть, что анализ результатов приведет к тому, что основной и один из дополнительных опросников должны будут поменяться местами или какой-либо опросник, являющийся дополнительным параметром, должен быть исключен.

ЗАКЛЮЧЕНИЕ

Проведенные с помощью разработанной ДСМ-системы предварительные исследования показали её работоспособность. Интеллектуальная ДСМ-система не только строит гипотезы о причинах, позволяющих делать вывод о подлинности подписи и находит эмпирические закономерности, но и обладает средствами оценки их качества. Гибкая структура этой системы дает возможность учитывать аксиомы предметной области и особенности решаемых задач. Созданная интеллектуальная система, опирающаяся на совместную работу экспертов в предметной области и специалистов в области искусственного интеллекта, будет способствовать выявлению механизмов, влияющих на формирование подписи, а также преодолению субъективизма при описании данных и повышению объективности выводов.

Принципы, заложенные в построение интеллектуальной системы, позволяют расширять ее функциональность для ДСМ-исследования, введения расшифровок к подписям, что должно повысить точность и обоснованность их идентификации, а также реализации ситуационного ДСМ-метода с расширением его на несколько дополнительных параметров. Возможно также подключение новых подсистем для других криминалистических исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Дорошенко Т.Ю., Костюченко Е.Ю. Система аутентификации на основе динамики рукописной подписи // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2014. – № 2(32). – С. 219-223.

2. He S., Schomaker L. Writer identification using curvature-free features // Pattern Recognition – 2017 – Vol. 63. – P. 451-464.
3. Аникин И.В., Анисимова Э.С. Распознавание рукописной подписи на основе нечеткой логики // Вестник Казанского государственного энергетического университета. – 2016. – № 3(31). – С. 48-64.
4. Петрова С.И. Диагностика психологических свойств по почерку (криминалистические, физиологические и психологические аспекты) // Известия Тульского государственного университета. Экономика и юридические науки. – 2016. – № 1-2. – С. 1-11.
5. Бубнова И.С., Шестеперова Е.Л. Некоторые теоретические и практические проблемы установления психологических свойств по почерку // Сборник материалов III международной научно-практической конференции «Актуальные вопросы судебно-психологической экспертизы и комплексной экспертизы с участием психолога. Перспективы научного и прикладного исследования почерка» (г. Калуга, 16–19 мая 2019 г.). – Калуга: Изд-во КГУ им. К.Э. Циолковского, 2019. – С. 27-34.
6. Pratiwi D., Santoso G.B., Saputri F.H. The application of graphology and enneagram techniques in determining personality type based on handwriting features // Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). – 2017. – Vol.10, № 1. – P. 11-18.
7. Гусакова С.М., Охлупина А.Н. Интеллектуальная ДСМ-система как средство автоматизированной поддержки научных исследований в почерковедении // Научно-техническая информация. Сер. 2. – 2019. – № 6. – С. 1-8; Gusakova S.M., Okhlupina A.N. Intelligent DSM systems as an automated support tool for scientific research on handwriting // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol. 53, № 3. – P. 114-121.
8. Арский Ю.М., Финн В.К. Принципы конструирования интеллектуальных систем // Информационные технологии и вычислительные системы. – 2008. – № 4. – С. 4-36.
9. Финн В.К. Обнаружение эмпирических закономерностей в последовательностях баз фактов посредством ДСМ-рассуждений // Научно-техническая информация. Сер. 2. – 2015. – № 8. – С. 1-29; Finn V.K. Selection of an algorithm for the parallel implementation of the similarity method in intelligent DSM systems // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, № 4. – P. 122-151.
10. Финн В.К. Эпистемологические основания ДСМ-метода автоматического порождения гипотез // Научно-техническая информация. Сер. 2. – 2013. – № 9. – С. 1–30 (Ч I); № 12. – С. 1-29 (Ч. II).
11. Охлупина А.Н. Проблема однозначности выделения признаков подписей и ее влияние на процесс автоматизации экспертных исследований // Алтайский юридический вестник. – 2019. – №1(26). – С. 115-120.
12. Русалов В.М. О природе темперамента и его месте в структуре индивидуальных свойств человека // Вопросы психологии. – 1985. – №1. – С. 19-32.
13. Глебов В.М. Актуальность использования анализа подписи при составлении психологического портрета личности военнослужащего графологическим методом // Вестник Института мировых цивилизаций. – 2016. – № 12. – С. 81-84.
14. Климова С.Г., Михеенкова М.А. Формальные средства ситуационного анализа: опыт применения // Научно-техническая информация. Сер.2. – 2012. – №10. – С. 1-13; Klimova S.G., Mikheyenkova M.A. Formal methods of situational analysis: experience from their use // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 5. – P. 183-194.
15. Гусакова С.М., Михеенкова М.А., Финн В.К. О логических средствах автоматизированного анализа мнений // Автоматическое порождение гипотез в интеллектуальных системах / под ред. проф. В.К. Финна. – М.: Книжный дом «Либроком», 2009. – С. 446-484.
16. Гусакова С.М., Михеенкова М.А. Интеллектуальный анализ данных как инструмент формирования структуры социума // Научно-техническая информация. Сер. 2. – 2016. – № 8. – С. 9-18.
17. Гусакова С.М. Зависимость особенностей подписи от психофизиологических характеристик ее исполнителя: подход к решению задачи // Труды Шестнадцатой Национальной конференции по искусственному интеллекту с международным участием – КИИ-2018 (24–27 сентября 2018 г., г. Москва, Россия). В 2-х томах. Т 1. – М.: РКП, 2018. – 308 с.
18. Norris E.M. An Algorithm for computing the maximal rectangles in a binary relation // Revue Roumaine de Mathématiques Pures et Appliquées. – 1978. – № 23(2). – P. 243- 250.

Материал поступил в редакцию 31.07.20

Сведения об авторах

ГРОСС Екатерина Романовна – младший системный аналитик АО «Райффайзенбанк», Москва
e-mail: katya-gross@yandex.ru

ГУСАКОВА Светлана Марковна – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: svem45@yandex.ru

ОГОРЕЛЬЦЕВА Наталия Владимировна – младший технический писатель Mail.Ru Group, Москва
e-mail: ogoreltseva.nat@gmail.com

ОХЛУПИНА Анастасия Николаевна – кандидат юридических наук, преподаватель кафедры исследования документов Московского университета МВД РФ им. В.Я. Кикотя, Москва
e-mail: stasya.zharova@inbox.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 007.51–051–047.36

Т.И. Булдакова, А.В. Соколова, А.К. Халайджи

Мониторинг состояния человека–оператора киберфизической системы

Рассмотрена задача обработки разнородной информации, регистрируемой в процессе мониторинга и используемой для оценки состояния оператора киберфизической системы. Отмечено, что деятельность человека-оператора сопровождается высоким психоэмоциональным напряжением, что может негативно сказаться на его работоспособности и привести к ошибкам. В связи с этим обосновывается необходимость мониторинга его функционального состояния. Выделены особенности систем мониторинга и приведены примеры оценки состояния оператора по разнородным данным.

Ключевые слова: *человек-оператор, оценка состояния, мониторинг, разнородные данные, структурирование, скрытые закономерности*

DOI: 10.36535/0548-0027-2020-10-3

ВВЕДЕНИЕ

Сегодня киберфизические системы, как новый класс информационно-управляющих систем, используются во многих прикладных областях, в том числе применяются на критически важных объектах национальной инфраструктуры (в здравоохранении, теплоэнергетике, обороне, на транспорте, в нефтегазовой сфере и др.). Основу для построения подобных систем обеспечили технологический прогресс в области микроэлектроники и сенсорной техники, а также развитие информационно-коммуникационных технологий.

Киберфизические системы – это робастные системы реального времени с высокими требованиями к производительности и надежности, характеризующиеся большой насыщенностью сенсорами и исполнительными устройствами. Эти системы обеспечивают автоматический режим работы сложных технических объектов, особенно на нижних иерархических уровнях реализации производственного процесса, и сокращают до минимума обслуживающий персонал [1–4].

Несмотря на высокий уровень автоматизации обработки информации и поддержки принятия управленческих решений, центральным компонентом в киберфизических системах является человек-оператор, принимающий ответственные решения [5–7]. Его взаимодействие с автоматизированными системами в подобной высокотехнологичной среде становится всё более сложным и многообразным. Это создает значительные нагрузки и может приводить к ошибочным действиям, поэтому деятельность оператора, управляющего сложными техническими устройствами и решающего важные задачи, характеризуется высоким психоэмоциональным напряжением, а это может негативно сказаться на его работоспособности [8–10].

Таким образом, в киберфизических системах необходимо осуществлять мониторинг функционального состояния человека-оператора. При этом важной задачей является обработка разнородной информации, которая регистрируется в процессе мониторинга и используется для оценки и прогнозирования изменения состояния человека-оператора. В настоящей работе приведены примеры применения структурированной информации для оценки функционального состояния по биосигналам и косвенным данным.

ОСОБЕННОСТИ СИСТЕМ МОНИТОРИНГА

В общем случае системы мониторинга состояния человека позволяют интегрировать различную информацию о факторах, влияющих на его здоровье. Сгруппируем эти факторы следующим образом.

В первую очередь, – это результаты медицинских обследований, которые непосредственно характеризуют состояние организма. Обычно формирование этого набора данных происходит при медосмотре перед допуском оператора к производственной деятельности, когда врач фиксирует состояние человека, вводя медицинскую информацию в электронную историю болезни. Эта информация в достаточной степени изменчива и требует периодического обновления. Кроме того, она не гарантирует отсутствие проблем во время работы оператора.

Другую группу составляют факторы, характеризующие внешнюю среду. К ним можно отнести, например, характер производственной деятельности человека, условия труда, условия быта и питания, экологию и др. Обычно подобная информация более стабильна и служит для формирования общестатистических показателей. Она может использоваться для оценки влияния внешней среды на состояние человека-оператора.

В зависимости от целей мониторинга могут быть выделены и другие факторы. Например, для человека-оператора, управляющего сложными техническими устройствами и принимающего ответственные решения в киберфизических системах, важны психологические показатели, характеризующие его устойчивость к стрессу, эмоциональность и т.д.

Для оценки функционального состояния человека-оператора и прогнозирования его изменения можно также использовать биосигналы, которые обладают довольно большой предсказательной способностью.

Таким образом, кроме оценки текущего состояния необходимо прогнозировать изменение функционального состояния заранее, чтобы суметь предупредить его ухудшение и вовремя предпринять меры. Этим целям и служит мониторинг состояния оператора киберфизической системы.

В процессе мониторинга реализуются информационные процессы сбора, передачи, хранения и обработки данных, необходимых для формирования и принятия решений о состоянии человека-оператора (рис. 1).

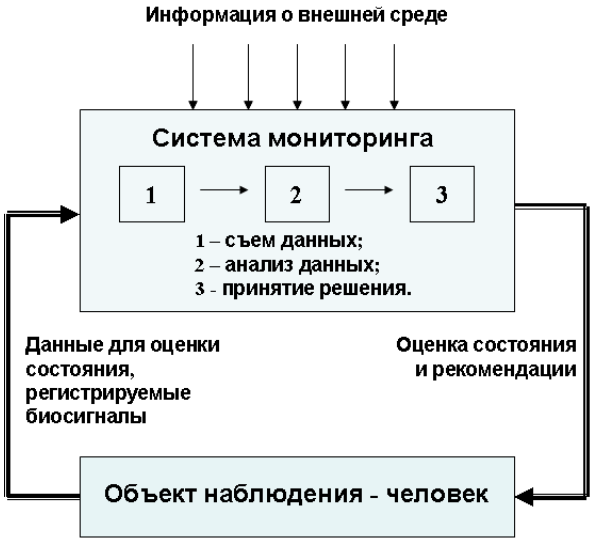

Рис. 1. Информационные процессы при мониторинге функционального состояния человека-оператора

Независимо от используемого канала передачи данных (выделенный канал *DSL*, проводной канал, радиоканал, *GSM*-канал, *Internet*, *Ethernet*) в систему мониторинга поступает огромный объем информации разного типа (количественной, качественной, интервальной, бинарной и т.д.). Весь этот массив данных характеризует (возможно, с разных сторон) одного и того же человека и содержит потенциальное знание о состоянии его организма [11, 12].

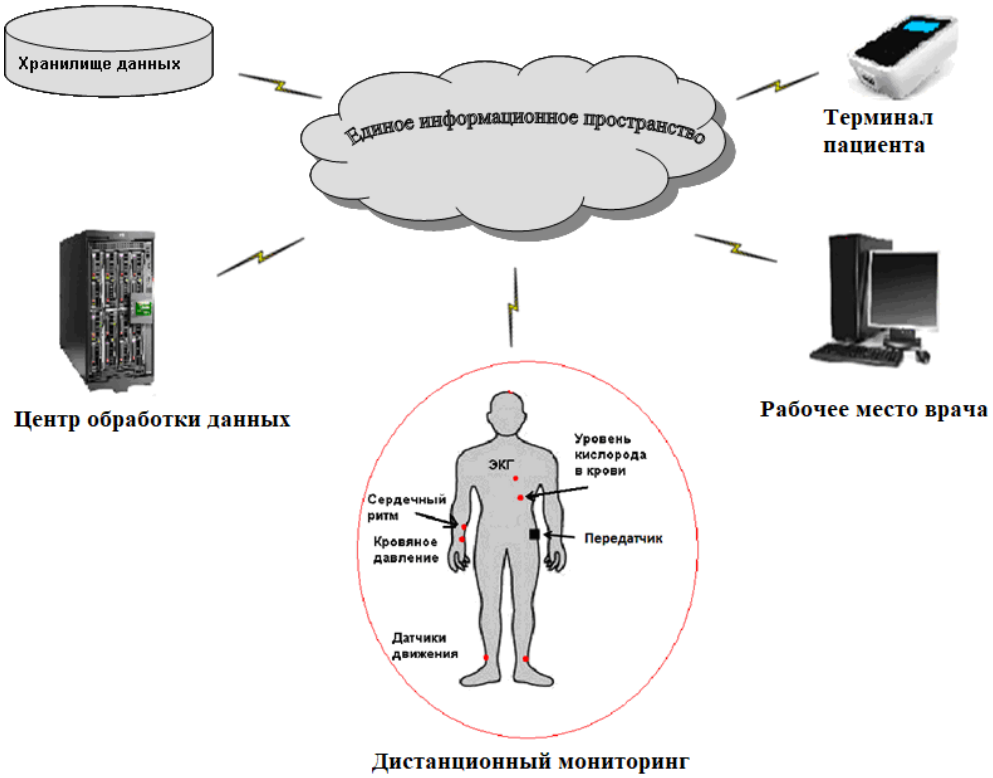

Рис. 2. Основные компоненты системы дистанционного мониторинга функционального состояния человека-оператора

В связи с развитием технологий Интернета вещей появляется всё больше «умных» устройств, которые пересылают данные в единую систему мониторинга, что позволяет решать задачу дистанционной оценки состояния человека, в том числе при выполнении им производственных обязанностей. Для сбора информации могут использоваться различные группы устройств, например: 1) устройства, косвенно проверяющие состояние человека на основе измерения скорости реакции на выдаваемые задания; 2) устройства, измеряющие такие характеристики, как пульс, давление, электрическое сопротивление руки. Однако подобные устройства являются вторичными и зачастую не отражают реальных изменений состояния человека. Хотя они и могут применяться при мониторинге, но их использование для оценки работоспособности оператора киберфизической системы малоэффективно.

Многообещающим направлением развития дистанционного биомониторинга является применение нательных датчиков, которые устанавливаются на теле человека (интегрируются в одежду, различные аксессуары, мобильные телефоны). Они регистрируют физиологическую информацию (в основном, биосигналы) и передают её по беспроводным каналам связи на сервер (рис. 2). Источниками первичной информации для оценки функционального состояния выступают датчики, регистрирующие различные биосигналы (электрическую активность и сокращения сердца, пульсовой сигнал, электрическую активность мозга, функцию внешнего дыхания и т.д.).

Таким образом, системы мониторинга функционального состояния человека-оператора позволяют собирать большой объем разнородных данных, включая различные биосигналы, и поэтому они являются основой для интеграции знаний о функциональном состоянии здоровья оператора.

НЕОБХОДИМОСТЬ СТРУКТУРИРОВАНИЯ ИНФОРМАЦИИ ДЛЯ ОЦЕНКИ СОСТОЯНИЯ ЧЕЛОВЕКА

Оценка функционального состояния в системах мониторинга может проводиться различными способами, например: 1) под наблюдением медицинского работника; 2) на основе анализа контролируемых параметров (норма – патология); 3) автоматизировано по вычислительной модели человека («виртуальной физиологии»), описывающей физиологическую активность подсистем организма человека [13-16].

Обычно в процессе мониторинга значимые показатели функций и адаптационных резервов организма сравниваются с нормативами, и далее дается оценка состояния в виде «здоров», «практически здоров», «относится к группе риска», «нуждается в наблюдении и коррекции». Кроме того, используя результаты анализа собранных данных, системы дистанционного мониторинга способны обеспечивать поддержку принятия решений при нештатных ситуациях и выработку рекомендаций по организации труда людей, управляющих сложной техникой.

Для получения целостного суждения о функциональном состоянии человека-оператора информация должна быть структурирована, чтобы облегчить обнаружение в ней скрытых закономерностей и взаимосвязей, характеризующих процессы в организме. При этом необходимо учитывать всю доступную (по результатам мониторинга) информацию для принятия решения о возможном состоянии и прогнозе его изменения.

С целью структуризации и выявления закономерностей из потока разнородной информации проводится слияние или интеграция данных. Сложность заключается в том, что измеряемые (количественные) показатели имеют разную природу, единицы измерений и диапазоны изменения. Также используются различные средства измерений, дающие разные значения точности результатов. Некоторые показатели могут иметь качественный вид.

Задача облегчается при наличии нормированных шкал с обеспечением определённой степени достоверности результатов контроля. Поэтому собранные данные должны подвергаться первичной обработке, которая включает их нормирование и кодирование. Также для повышения наглядности могут применяться как различные способы группирования информации, так и модели представления данных.

Основными этапами обработки данных для оценки состояния человека-оператора в киберфизической системе являются:

1) регистрация текущих данных по результатам мониторинга;

2) первичная обработка разнородных данных (удаление шумов, пропусков, трендов);

3) группировка и упорядочение данных, включая их нормирование и кодирование;

4) выбор информативных (для оценки состояния) показателей, в том числе оказывающих наибольшее влияние на работоспособность человека-оператора;

5) выбор технологии анализа структурированных данных;

6) установление взаимосвязи между информативными показателями и классами функционального состояния человека (с помощью обученной модели);

7) оценка функционального состояния человека-оператора и прогноз его изменения.

Таким образом, при оценке состояния человека-оператора киберфизической системы необходимо структурировать разнородную информацию, выявлять в ней скрытые закономерности и вырабатывать решения [17, 18]. Кроме того, требуется прогнозировать изменение состояния человека и, что особенно важно, выявлять ситуации, приводящие к ухудшению работоспособности оператора.

Выбор метода или модели оценки на этапе 5 зависит от собранной при мониторинге информации и решаемой задачи (текущая оценка или прогнозирование изменения состояния). Приведем примеры оценки и прогнозирования функционального состояния.

ОЦЕНКА СОСТОЯНИЯ ЧЕЛОВЕКА–ОПЕРАТОРА ПО БИОСИГНАЛАМ

В настоящее время известны различные подходы к оценке функционального состояния человека-оператора, важные для определения его работоспособности. Однако основную группу составляют методы, использующие в качестве первичной информации ЭКГ и вариабельность сердечного ритма [19].

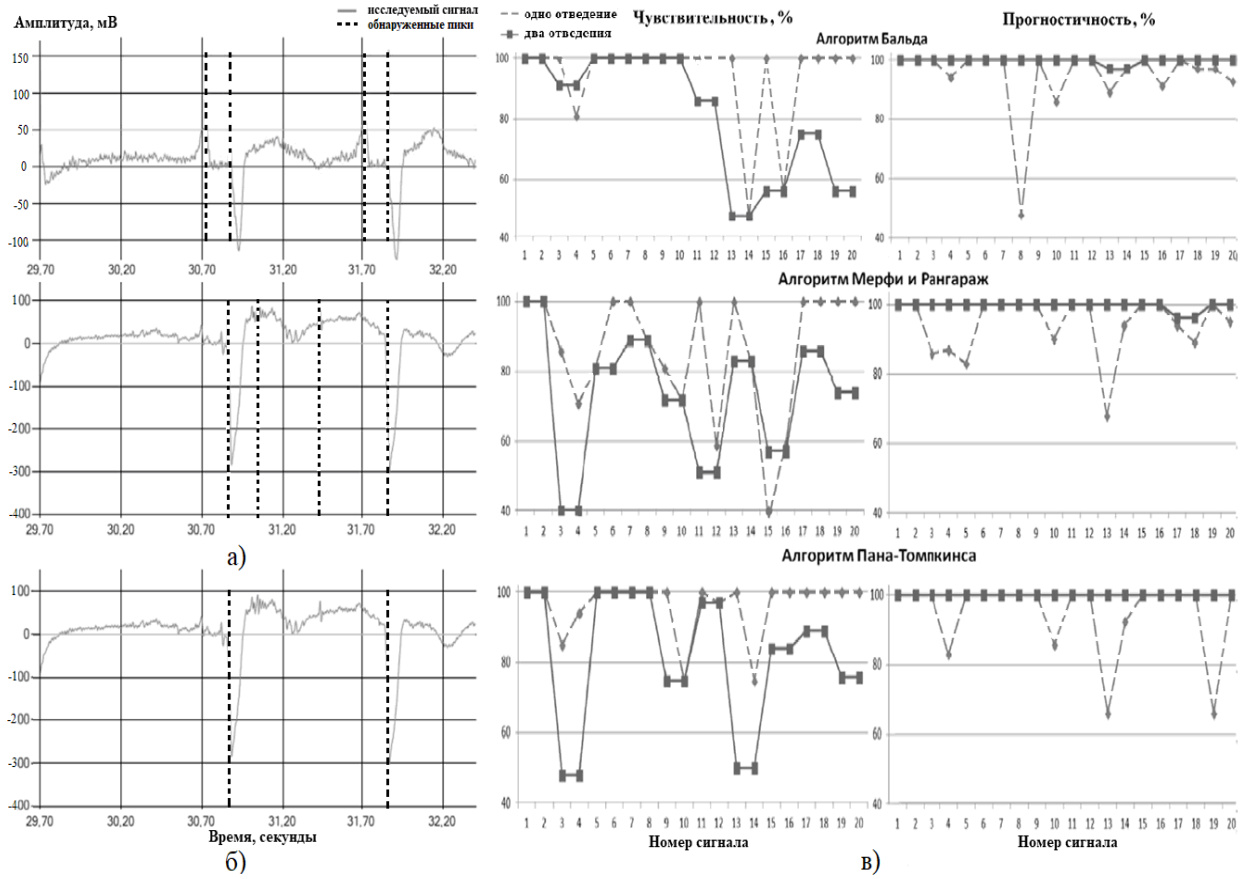

Рис. 3. Результаты тестирования одноминутных записей сигналов ЭКГ:
а) – независимый поиск в двух отведениях; б) – совместный поиск; в) – показатели чувствительности и прогностичности для трех алгоритмов

Это связано с тем, что приспособление психофизиологических функций к рабочей деятельности человека-оператора обеспечивается свойством адаптивности организма, индикатором которой как раз и является сердечно-сосудистая система.

Для повышения достоверности оценки функционального состояния целесообразно учитывать более одного биосигнала, поскольку совместная обработка синхронно зарегистрированных биосигналов повышает надежность оценки [9]. Продемонстрируем это на примере обнаружения *R*-пиков при наличии нескольких отведений ЭКГ, используя простой эвристический подход – пик фиксируется, если он есть в большинстве сигналов, при этом его местоположение центрируется по всем сигналам.

Результаты тестирования одноминутных записей сигналов ЭКГ из базы данных *MIT* для разных алгоритмов извлечения R-зубцов представлены на рис. 3.

Видно, что при обработке двух отведений показатели прогностичности, учитывающие количество ложно обнаруженных пиков, стали выше, однако для некоторых сигналов при наличии отведения, в котором пики были обнаружены хуже, – чувствительность понизилась. Это связано с тем, что в используемом подходе при обработке двух отведений обязательно наличие пика в двух сигналах. Если в паре отведений обнаружение пиков не может быть произведено качественно, то показатели чувствительности уменьшаются, но при увеличении числа сигналов такие ситуации меньше влияют на общие результаты. Таким образом, обработка даже двух отведений одновременно повышает точность обнаружения R-пиков.

Рассмотрим пример оценки функционального состояния с использованием двух биосигналов. Взаимосвязь исходных данных и классов состояний опишем с помощью математической модели, параметры которой настраиваются на биосигналы человека-оператора.

Предлагается для создания математической модели использовать сфигмограмму, которая регистрирует колебания стенки кровеносного сосуда, обусловленные выбросом ударного объема крови в артериальное русло. На основе принципа базовых моделей колебательных систем и с учетом биомеханики сосуда, можно описать динамические свойства сосудистой стенки уравнением Ван-дер-Поля – Релея:

$$\ddot{x}+\left[\varepsilon_1\left(x^2-r^2\right)+\varepsilon_2\left(\dot{x}^2-\omega_0^2\cdot r^2\right)\right]\cdot\dot{x}+ax=P(\omega_0 t), \quad (1)$$

где: x – перемещение стенки артерии, регистрируемое датчиком;

$P(\omega_0 t)$ – воздействие сердечной активности на динамику стенки сосуда;

ε_1, ε_2, a, ω_0 и r – параметры модели, определяющие колебания стенки кровеносного сосуда (частота, амплитуда и др.). При этом сигнал ЭКГ является входом модельной системы, а сфигмограмма – выходом.

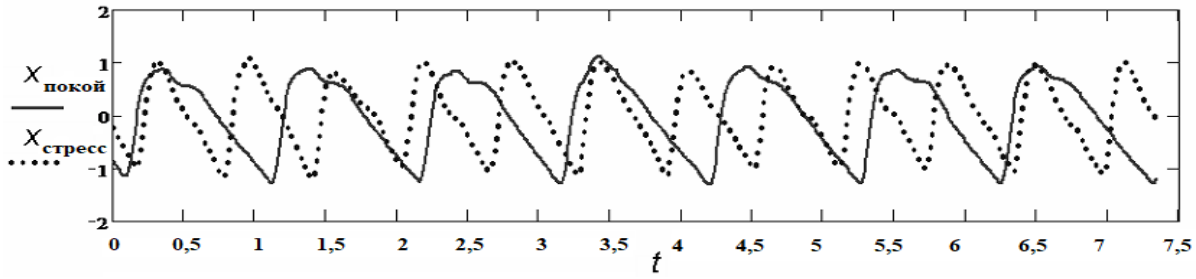

Рис. 4. Пульсовые сигналы в покое и при стрессовой нагрузке

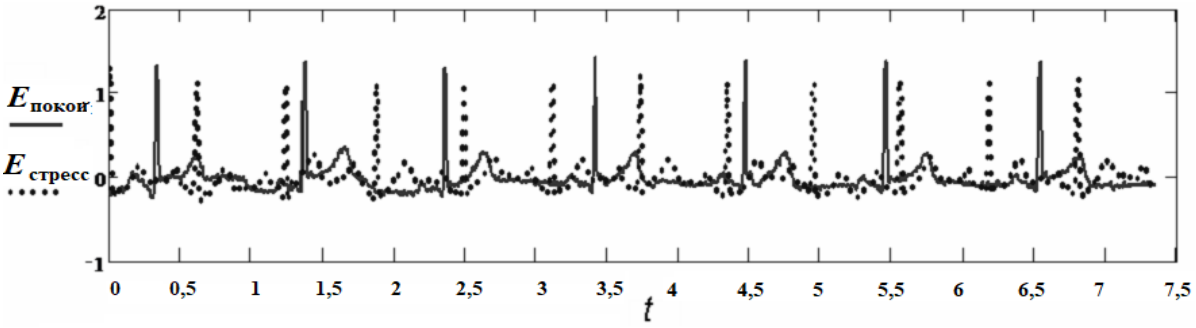

Рис. 5. Сигналы электрической активности сердца в покое и при стрессовой нагрузке

В представленной модели параметр a имеет физический смысл, так как он отражает механические свойства сосудов, в частности «жесткость» их стенок. При здоровых (эластичных) сосудах адаптационный механизм человека стремится изменить их «жесткость» с увеличением нагрузки за счет работы обволакивающих гладких мышц. Поэтому представленная модель содержит потенциальное знание о состоянии сосудов, а, значит, и о состоянии сердечно-сосудистой системы конкретного человека-оператора.

Поскольку система «сердце-сосуды» функционирует в режиме предельного цикла, то неизвестные параметры ω_0 и r модели (1) можно определить по экспериментальным данным, после чего с использованием измеренных значений $x_u(t)$ и вычисленных значений $\dot{x}(t)$ и $\ddot{x}(t)$, находятся значения p_i, a, ε_1 и ε_2. Здесь p_i – это коэффициенты разложения функции P в ряд Фурье, $i = 1, \ldots, N$.

На рис. 4 и 5 показаны сигналы пульсовой и электрической активности сердца в покое (сплошная линия) и при психоэмоциональном напряжении (пунктирная линия), зарегистрированные синхронно. Из рисунков видно, что помимо увеличения частоты изменяется также форма биосигналов.

Предложенная модель (1) позволяет различить эти состояния, им соответствуют разные значения параметров модели a, ε_1 и ε_2. К примеру, для человека в состоянии покоя получены значения: $\varepsilon_1 = -0{,}3$; $\varepsilon_2 = -3{,}37$; $a = 36{,}15$. В стрессовом состоянии при интенсивной нагрузке для того же человека параметры получили значения: $\varepsilon_1 = -1{,}07$; $\varepsilon_2 = -8{,}31$; $a = 95{,}5$.

Эффективность оценки состояния человека с помощью подобных базовых моделей, созданных на основе двух синхронно зарегистрированных биосигналов, отмечена в [8, 9].

ПРОГНОЗИРОВАНИЕ СОСТОЯНИЯ ЧЕЛОВЕКА–ОПЕРАТОРА ПО КОСВЕННЫМ ДАННЫМ

Последствия высокого психоэмоционального напряжения работы человека-оператора нельзя недооценивать. Общеизвестно, что стрессы приводят к различным заболеваниям [20, 21]. При этом патологические изменения в организме происходят постепенно и незаметно, проявляясь на поздних стадиях заболевания, когда требуются значительные усилия врачей. Это приводит к тому, что опытные специалисты надолго выбывают из строя. Поэтому весьма важно оценить подверженность сотрудников заболеваниям, обусловленным напряженным характером работы, а также идентифицировать заболевание на ранней стадии и сделать прогноз его развития в зависимости от различных факторов. Это позволит своевременно разработать и провести в жизнь необходимые социальные программы и, в конечном счете, сохранить работоспособность квалифицированных специалистов.

Поэтому при оценке состояния человека-оператора киберфизической системы важно учитывать не только биосигналы, но и нейрогуморальные воздействия, влияющие на психологическое поведение и реакцию человека. Важность учета нейрогуморальных регуляций состоит в том, что они приспособли-

вают работу сердечно-сосудистой системы к потребностям организма в условиях внешних и внутренних воздействий. Результатом регулирования является изменение свойств сердечно-сосудистой системы, влияющих на функциональное состояние человека. Более того, скрытые закономерности, содержащиеся в информации о психофизиологическом поведении оператора и степени напряженности его производственной деятельности, могут использоваться для прогнозирования ухудшения состояния здоровья человека-оператора. Поэтому структурирование разнородной информации и применение интеллектуальных технологий для извлечения скрытых знаний позволяют оценить вероятность возникновение и развитие ряда тяжелых хронических заболеваний.

В частности, известно, что основной причиной такого серьезного недуга как язвенная болезнь являются психологическое напряжение и стрессовые ситуации. Поэтому в качестве примера рассмотрим задачу прогнозирования язвенной болезни на основании косвенных факторов (психологических характеристик, характера работы и т.п.).

Представим организм человека сложной динамической системой в виде $S = \{X, A, V\}$, где X – вектор переменных состояния системы, или вектор симптомов; A – набор параметров, характеризующих свойства системы; V – вектор возмущающих воздействий. При этом имеется группа параметров a_i, отражающих чувствительность системы к одному определенному классу внешних возмущений $V_1 \subseteq V$. В данном случае a_i – это психологические и другие показатели, характеризующие реакции организма на стрессовые возмущения V_1. Задача состоит в том, чтобы при действии постоянных возмущений V_1 на основании показателей a_i, которые измерены в разных шкалах, сделать прогноз о вероятности возникновения язвенной болезни.

В качестве параметров a_i используем результаты тестирования операторов киберфизической системы при помощи универсальных психологических тестов. Выделим шесть показателей, которые закодируем следующим образом: 1) сила нервных процессов (0 – слабая, 1 – средняя, 2 – сильная); 2) уравновешенность нервных процессов (0 – возбуждение, 1 – торможение, 2 – норма); 3) выраженность эмоционального стресса (0 – максимальная, 1 – сильная, 2 – умеренная, 3 – отсутствует); 4) уровень снижения работоспособности (0 – высокий, 1 – умеренный, 2 – отсутствует); 5) шкала социальной адаптации Холмса (0 – низкая, 1 – пороговая, 2 – высокая); 6) Торонтская шкала алекситимии (0 – неврозы, 1 – психосоматические заболевания, 2 – здоровые). Кроме того, учтем характер и стаж работы, а также наследственность (генетическую устойчивость человека). Отметим, что все эти данные могут быть легко получены по результатам простого общения с человеком-оператором, без специального медицинского обследования у врача. Подобная информация важна при оценке функционального состояния человека-оператора и при выборе характера его производственной деятельности.

Поскольку сведения о взаимосвязи между функциональным состоянием и совокупностью параметров a_i, собранных по результатам тестирования людей в процессе мониторинга состояния здоровья, отсутствовали, то в качестве модели применялась обученная нейронная сеть (рис. 6). Для обучения сети использовалась информация о здоровых людях с различной интенсивностью производственной деятельности, а также данные о больных язвенной болезнью. Процент правильного распознавания с помощью нейронной сети составил свыше 85% от общего количества прогнозов.

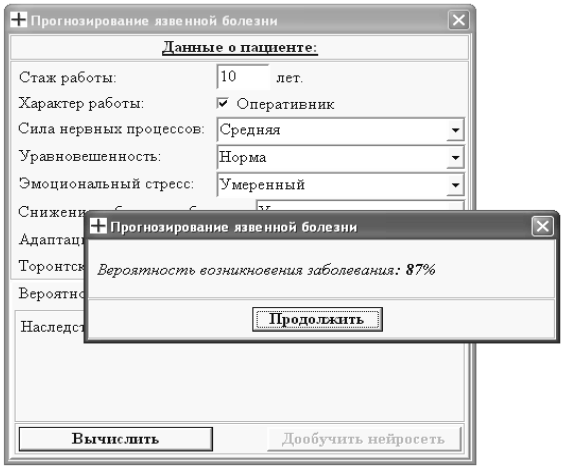

Рис. 6. Фрагмент работы нейросетевой модели

Таким образом, по косвенной информации (включающей психологические показатели) можно прогнозировать вероятность возникновения и развитие этого тяжелого заболевания. Более того, универсальные психологические тесты с помощью нейросетевых технологий полезны при организации адресных профилактических мероприятий, поскольку язвенную болезнь с полным основанием можно отнести к разряду профессиональных заболеваний человека-оператора киберфизической системы.

ЗАКЛЮЧЕНИЕ

В настоящей статье было показано, что информация, собранная при мониторинге, содержит потенциальное знание о состоянии здоровья человека, для его извлечения необходимо структурировать разнородную информацию, которая затем используется для оценки и прогнозирования состояния человека-оператора. Кроме того продемонстрированы два подхода к оценке функционального состояния: по биосигналам и по косвенным факторам, к которым относятся психологические характеристики человека и продолжительность напряженной работы. Представлена также базовая модель биосистемы «сердце–сосуды», состояние которой является индикатором функционального состояния оператора. Для прогнозирования состояния здоровья по косвенным факторам применена обученная нейронная сеть.

СПИСОК ЛИТЕРАТУРЫ

1. Lu Y. Industry 4.0: A survey on technologies, applications and open research issues // Journal of Industrial Information Integration. – 2017. – Vol. 6 – P. 1-10. DOI: org/10.1016/j.jii.2017.04.005.
2. Koval' V.A., Osenin V.N., Suyatinov S.I., Torgashova O.Y. Synthesis of discrete controller for construction of a distributed controller of temperature conditions of steam oil heater // Journal of Computer and Systems Sciences International. – 2011. – Vol. 50(4). – P. 638-653. DOI: 10.1134/ S1064230711040125.
3. Herwan J., Kano S., Ryabov O., Sawada H., Kasashima N. Cyber-physical system architecture for machining production line // IEEE International Conference on Industrial Cyber-Physical Systems – ICPS. (St. Petersburg, Russia. 15-18 May 2018). – P. 387-391. – URL: https://ieeexplore.ieee.org/ document/ 8387689. DOI: 10.1109/ ICPHYS.2018.8387689.
4. Булдакова Т.И., Суятинов С.И. Информационно-аналитическая система управления снабжением и производством инструмента // Информационные технологии. –2002. – № 11. – С. 28-33.
5. Sowe S.K., Zettsu K., Simmon E., de Vaulx F., Bojanova I. Cyber-Physical Human Systems: Putting People in the Loop // IT Professional. – 2016. –Vol. 18(1). – P. 10-13. DOI: 10.1109/MITP.2016.14.
6. Buldakova T.I., Suyatinov S.I. Assessment of the state of production system components for digital twins technology // Cyber-Physical Systems: Industry 4.0 Challenges. Studies in Systems, Decision and Control. Vol. 259 / eds. A. Kravets, A. Bolshakov, M. Shcherbakov. – Cham: Springer, 2020. – P. 253-262. DOI: 10.1007/978-3-030-32579-4_20.
7. Обычайко Д.С., Шихин В.А. Методика формализации киберфизических систем в задачах анализа надёжности // Автоматизация. Современные технологии. – 2018. – Т. 72, № 9. – С. 414-421.
8. Булдакова Т.И., Суятинов С.И. Разработка адекватных моделей в технологии цифровых двойников // Автоматизация. Современные технологии. – 2019. – Т. 73, № 8. – С. 367-373.
9. Suyatinov S.I. Criteria and method for assessing the functional state of a human operator in a complex organizational and technical system // IEEE Global Smart Industry Conference –GloSIC (Chelyabinsk, Russia, 13-15 Nov. 2018). – P. 1-6. – URL: https://ieeexplore.ieee.org/document/8570088. DOI: 10.1109/GloSIC.2018.8570088.
10. Булдакова Т.И., Джалолов А.Ш. Анализ информационных процессов и выбор технологий обработки и защиты данных в ситуационных центрах // Научно-техническая информация. Сер. 1. – 2012. – № 6. – С. 16-22.
11. Buldakova T.I., Sokolova A.V. Network services for interaction of the telemedicine system users // IEEE Proceedings – 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency – SUMMA (Lipetsk, Russia, 20-22 Nov. 2019). – P. 387-391. – URL: https://ieeexplore.ieee.org/document/8947552. DOI: 10.1109/SUMMA48161.2019.8947552.
12. Булдакова Т.И., Миков Д.А. Анализ информационных процессов виртуального центра охраны здоровья // Научно-техническая информация. Сер. 2. – 2014. – № 2. – С. 10-20.
13. Bashi N., Karunanithi M., Fatehi F., Ding H., Walters D. Remote monitoring of patients with heart failure: an overview of systematic reviews // Journal of Medical Internet Research. – 2017. – Vol. 19(1). – P. 1-14.
14. Забежайло М.И., Трунин Ю.Ю. К проблеме доказательности медицинского диагноза: интеллектуальный анализ эмпирических данных о пациентах в выборках ограниченного размера // Научно-техническая информация. Сер. 2: Информационные процессы и системы. – 2019. – № 12. – С. 12-18;. Zabezhailo M.I., Trunin Yu.Yu. On the problem of medical diagnostic evidence: intelligent analysis of empirical data on patients in samples of limited size // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol. 53, №. 6. – P. 322-328.
15. Анищенко В.С., Булдакова Т.И., Довгалевский П.Я., Лифшиц В.Б., Гриднев В.И., Суятинов С.И. Концептуальная модель виртуального центра охраны здоровья населения // Информационные технологии. – 2009. – № 12. – С. 59-64.
16. Булдакова Т.И., Игнатьева Е.В., Ляпина Н.С., Суятинов С.И. Оценка состояния человека и выделение групп риска развития хронических заболеваний // Системный анализ и управление в биомедицинских системах. – 2011. – Т. 10, № 2. – С. 391–395.
17. Suyatinov S. Bernstein's theory of levels and its application for assessing the human operator state // Recent Research in Control Engineering and Decision Making. ICIT-2019. Studies in Systems, Decision and Control. Vol. 199 / eds. O. Dolinina at al. – Cham: Springer, 2019. – P. 298-312. DOI: 10.1007/978-3-030-12072-6_25.
18. Buldakova T.I., Suyatinov S.I. Biological principles of integration information at big data processing // Proceedings – 2019. International Russian Automation Conference –RusAutoCon (Sochi, Russia, 8-14 Sept. 2019). – P. 1-6. – URL: https://ieeexplore.ieee.org/document/8867710. DOI: 10.1109/RUSAUTOCON.2019.8867710.
19. Oweis R.J., I-Tabbaa B.O. QRS detection and heart rate variability analysis: A survey // Biomedical Science and Engineering. – 2014. – Vol. 2(1). –2014. – P. 13–34.
20. Fedotchev A., Kruk V., Oh S.J., Semikin G. Eliminating pain-induced risks of operator reliability via transcutaneous electroneurostimulation controlled by Patient's breathing // International Journal of Industrial Ergonomics. – 2018. – Vol. 68. – P. 256-259. DOI: 10.1016/j.ergon.2018.08.004.

21. Can Y.S., Chalabianloo N., Ekiz D., Ersoy C. Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study // Sensors. – 2019. – Vol. 19(8). – P. 1-21. DOI: 10.3390/s19081849.

Материал поступил в редакцию 08.07.2020

Сведения об авторах

БУЛДАКОВА Татьяна Ивановна – доктор технических наук, профессор кафедры «Компьютерные системы и сети» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: buldakova@bmstu.ru

СОКОЛОВА Аксинья Владимировна – аспирант кафедры «Компьютерные системы и сети» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: aksinya.sokolova@yandex.ru

ХАЛАЙДЖИ Алексей Константинович – аспирант кафедры «Компьютерные системы и сети» Московского государственного технического университета имени Н.Э. Баумана.
e-mail: aleksei_halaidzh@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2

В.А. Яцко

Методика использования конкорданса и табличного процессора с целью авторской атрибуции*

Предлагается оригинальная методика авторской атрибуции текстов на основе отклонений от распределения Ципфа и дается её описание с использованием статистических данных, полученных с помощью конкорданса и вычислений в табличном процессоре. Эта методика предусматривает нахождение расстояний между входными и эталонным текстами на основе анализа отклонений частотностей стоп-слов. Полученные результаты показали, что предлагаемая методика позволяет производить достоверную атрибуцию текстов и благодаря простоте вычислений может быть использована в образовательном процессе для развития у студентов навыков и компетенций, связанных с автоматической обработкой текстовых документов.

Ключевые слова: *методы авторской атрибуции, конкордансы, закон Ципфа, распределение стоп-слов, обучение информационным технологиям*

DOI: 10.36535/0548-0027-2020-10-4

ВВЕДЕНИЕ

Информационные технологии на сегодняшнем этапе развития общества проникают во все сферы его жизни, включая научные дисциплины. Последние десятилетия характеризуются выделением и интенсивным развитием в рамках различных научных дисциплин информационных направлений, таких как биоинформатика, медицинская информатика, химическая информатика, правовая информатика, историческая информатика. Основная задача этих направлений – это разработка информационных технологий с учётом специфики данных дисциплин для поддержки исследований, разработок и информационного обслуживания специалистов. Начало процесса разработки информационных технологий в языкознании можно отнести к 1964 г., когда в Брауновском университете США был создан первый электронный аннотированный корпус текстов, возможности использования которого были описаны в работе Г. Кучеры (H. Kucera) и У. Френсиса (W. Francis) [1], ознаменовавшей появление корпусной лингвистики как нового направления, предназначенного для поддержки лингвистических исследований. С тех пор во многих странах были созданы текстовые корпуса разных видов (национальные, исторические, тематические), а использование информационных технологий стало неотъемлемой частью лингвистических исследований.

Однако достаточно часто лингвисты сталкиваются с необходимостью провести анализ распределения лингвистических единиц в тексте/текстах, которые отсутствуют в существующих корпусах. Для этого можно использовать приложения, специально предназначенные для статистического анализа текста – конкордансы. К настоящему времени разработан целый ряд таких приложений, которые распространяются как бесплатно, так и платно и устанавливаются локально, не требуя подключения к Интернету.

Цель настоящей статьи – показать возможности использования конкорданса AntConc [2] (наиболее функционального бесплатного приложения данного типа) для решения задачи авторской атрибуции текстов, а также для обучения студентов лингвистическим технологиям. Это приложение разрабатывается профессором Университета Васэда (Waseda University, Япония) Лоренсом Энтони (Laurence Anthony) с 2014 г., регулярно обновляется и ни чем не уступает по функциональности платным конкордансам. Наш опыт показал, что его можно успешно использовать на занятиях по дисциплине "Информационные технологии в лингвистике", которая изучается в рамках направления подготовки 45.03.02 Лингвистика (уровень бакалавриата). Федеральный государственный стан-

* Исследование выполнено при поддержке гранта РФФИ 20-07-00124

дарт, утверждённый в 2014 г., предусматривает в качестве требований к обучающимся по этому направлению «способность работать с основными информационно-поисковыми и экспертными системами, системами представления знаний, синтаксического и морфологического анализа, автоматического синтеза и распознавания речи, обработки лексикографической информации и автоматизированного перевода, автоматизированными системами идентификации и верификации личности (ПК-19); владением основными математико-статистическими методами обработки лингвистической информации с учетом элементов программирования и автоматической обработки лингвистических корпусов»[1]. Нами будут показаны возможности развития указанных компетенций на основе математико-статистических методов обработки исходных данных, полученных с помощью конкорданса AntConc. Вначале приведено описание функциональности современных конкордансов, далее – методика применения указанного конкорданса для получения статистических данных о распределении единиц текстов, а затем – предложенный ранее метод автоматической классификации текстов на основе отклонений от распределения Ципфа [3]. Теоретическую основу исследования составляют положения: о дистантном и словарном подходах к автоматической классификации текстовых документов [4]; о возможности использования стоп слов как одного из классификационных параметров [5]; о необходимости нормализации размеров текстовых документов [6]. Теоретическое значение исследования заключается в актуальности разработки новых методов эффективной автоматической классификации текстов.

КОНКОРДАНСЫ: ОСНОВНЫЕ ФУНКЦИИ

Конкордансы – это приложения, выполняющие три основные функции:

1) получение данных о распределении лингвистических единиц текста. К ним обычно относятся слова и словосочетания (н-грамы). Наряду с их частотностью в обязательном порядке генерируется информация об общем количестве токенов и уникальных слов;

2) получение данных о коллокациях – статистическом распределении терминов слева и справа от ключевого слова, заданного пользователем. Пользователю выдаётся информация о частотностях терминов, находящихся на определённой позиции (первое/второе/третье и т.д. слово справа/слева от ключевого слова);

3) просмотр контекста использования ключевого слова, заданного пользователем (*key word in context – KWIC*). Размер контекста задаётся пользователем (в количестве слов или символов справа/слева от ключевого слова). Эта функция и называется конкордансом, по ней и получил название весь этот вид лингвистического программного обеспечения.

Одним из конкордансов, популярных в начале 2000-х гг., являлся ConcApp (*Concordancer Application*), разработанный Крисом Гривсом (Chris Greaves), он позволял реализовывать все три указанные функции [7]. В 2000-2003 гг. были выпущены четыре бесплатных версии приложения с поддержкой английского, японского и китайского языков; в 2008 г. – платная версия, поддерживавшая также французский, русский и тайский языки. Дальнейшая работа над конкордансом прекратилась, хотя некоторые версии приложения до сих пор доступны в Интернете и могут быть использованы для обучения компьютерной обработке текстов на начальном уровне [8]. Распространяемые в настоящее время конкордансы предлагают намного более расширенную функциональность, включающую:

- распознавание ключевых слов на основе определённых метрик (TF*IDF, log-likelihood). Эта функция требует загрузки корпуса текстов, с которым сопоставляется входной текст;
- логические операции объединения и вычитания текстов. Объединение реализуется в виде пакетной обработки текстовых файлов и возможности получать статистические данные по всем файлам, включённым в пакет. Благодаря этому можно получить данные о специфике какого-то типа или жанра текста, что необходимо для решения задач автоматической классификации. Вычитание текстов предусматривает фильтрацию определённых слов, представленных в виде списка. Обычно имеются ввиду стоп-слова – служебные слова, встречающиеся с равномерно высокой частотностью в текстах различных жанров. К ним относятся артикли, местоимения, предлоги, союзы, частицы;
- средства визуализации распределения единиц текста, позволяющие наглядно представить распределение единиц текста в виде графиков, диаграмм, схем.

Наряду с этими функциями может предлагаться и дополнительная функциональность. Приложение *WordSmith* [9] предусматривает возможность получения статистики по отдельным символам и их сочетаниям, позволяет находить дубликаты текстовых файлов, объединять текстовые файлы в один файл, разбивать большие текстовые файлы на несколько отдельных файлов. В приложении *WordStat* [10] есть возможность предварительной обработки текста с помощью встроенных стеммеров и лемматайзеров (недоступно для русского языка). Соответственно, пользователи могут получать статистические данные и о распределении основ слов и частей речи. В этом же приложении в экспертном режиме с помощью встроенного словаря можно выполнять и анализ мнений, получая для входного текста коэффициенты положительной или отрицательной оценки.

WordStat и *WordSmith* – платные приложения. Минимальная цена первого – 400 долларов (лицензия на год для одного пользователя); *WordSmith* 8.0 стоит 50 фунтов стерлингов (лицензия для одного пользователя), при этом старая версия *WordSmith* 4.0 распространяется бесплатно. Платное распространение и достаточно сложный интерфейс, связанный с дополнительной функциональностью, ограничивают возможность использования этих приложений.

[1] Об утверждении федерального государственного образовательного стандарта высшего образования по направлению подготовки 45.03.02 Лингвистика (уровень бакалавриата): приказ Минобрнауки России от 07.08.2014 N 940. – URL: http://fgosvo.ru/uploadfiles/fgosvob/450302_Lingvistika.pdf (дата обращения: 25.06.2020).

МЕТОДИКА ПОЛУЧЕНИЯ ИСХОДНЫХ ДАННЫХ И ВЫЧИСЛЕНИЙ

Закон Ципфа устанавливает зависимость между рангом слова и его частотностью в тексте [4], которую можно выразить формулой:

$$F_r \propto \frac{1}{r^a}\ , \qquad (1)$$

где: F – частотность данного слова, r – его ранг в ранжированном списке, а экспонента a примерно равна единице. Закон Ципфа имеет предсказательную силу: зная частотность и ранг определенного слова, можно определить частотности и ранги всех остальных слов в данном тексте. Если, допустим, десятое по рангу слово повторяется в тексте 36 раз, то частотность пятого по рангу слова будет равна 36*10/5=72. Закон Ципфа задаёт идеальное распределение слов по частотностям. В конкретных текстах частотности слов, как правило, отклоняются от этого распределения. В настоящей статье мы продемонстрируем методику применения отклонений от распределения Ципфа с целью авторской атрибуции текстов.

1. Отбор тестовых документов. Были отобраны произведения: *The Man of Property*, *Indian Summer of a Forsyte*, *Oliver Twist*, а для обозначения текстовых файлов приняты соответствующие условные обозначения: *G*1, *G*2, *D*. Первые два произведения, написанные Дж. Голсуорси (J. Galsworthy), относятся к одному циклу – «Сага о Форсайтах» (The Forsyte Saga); третье – написано Ч. Диккенсом (Ch. Dickens). Отбирая эти тексты, мы исходили из гипотезы, что расстояние (Ds) между текстами Дж. Голсуорси должно быть существенно меньше, чем расстояние между текстами Дж. Голсуорси и Ч. Дикенса. Это обуславливается не только принадлежностью двух первых текстов одному автору, но и тому, что они относятся к разным временным периодам. Роман Ч. Дикенса был написан в середине XIX в., а произведения Дж. Голсуорси – в начале XX в., т. е. Ds(G2, G1) < Ds (D, G1). В соответствии с принятой нами методикой, сначала сопоставлялось распределение терминов (в нашем случае – стоп-слов) в G2 и G1, затем – в D и G1. Таким образом, файл G1 был принят как эталонный, а два остальных – как входные тексты.

Выбор английских текстов был обусловлен тем, что стоп-слова в английском языке характеризуются отсутствием флексий, что позволяет не применять дополнительные алгоритмы предварительной обработки текстов (стемминг, лемматизацию). Произведения художественной литературы были выбраны по следующим критериям: а) они характеризуются большим разнообразием лексики (в том числе и стоп-слов), своеобразием стилевых особенностей, что отражается на распределении стоп-слов; б) для произведений XIX – начала XX вв. характерен большой объём текста, позволяющий получать достоверные результаты; в) тексты были отобраны с сайта проекта Гутенберг (Project Gutenberg) [11], на котором размещаются произведения с истекшим сроком действия авторских прав. Эти произведения редактируются и вычитываются. Нами было проведено дополнительное редактирование, в результате которого удалена информация о самом проекте Гутенберг, которая занимает достаточно много места в конце и начале каждого произведения, размещаемого на сайте.

2. Для входного файла D было выполнено выравнивание по нижнему пределу (*undersampling*). В работе [12] было показано, что при выравнивании текстов более эффективно именно выравнивание по нижнему пределу, а выравнивание по верхнему пределу (*oversampling*) для текстовых документов применять нецелесообразно. В нашем случае выравнивание по нижнему пределу заключалось в удалении части большего по размеру входного текста (файл D), для того чтобы его размер, определяемый в количестве токенов, сравнялся с размером другого входного файла (G2). Для этого с помощью AntConc (функция Wordlist) вначале были получены данные о количестве уникальных слов и токенов для каждого из двух текстов. Затем, с помощью той же функции определялось количество токенов в оставшемся после удаления тексте. Удаление проводилось с конца текста, а эталонный текст выравниванию не подвергался. В результате выравнивания совпало количество токенов, в то время как число уникальных слов оставалось различным, что и являлось признаком стилевых различий.

Заметим, что применённый способ выравнивания – наиболее простой и достаточно грубый, так как в результате удаления нарушается структура и смысл текста. Такой способ применим, если используется анализ текста на основе подхода *bag of words* (списка слов), в соответствии с которым текст рассматривается как набор слов – без учёта контекста их использования и семантико-синтаксических связей. Если анализ предусматривает рассмотрение связей между лингвистическими единицами, то применяются более тонкие методы сглаживания разницы в размерах текстов. Наиболее общепринято использование косинусных величин [6]; нами для нейтрализации разницы в размерах текстов был предложен метод логарифмического выравнивания текстов [13].

В табл. 1 приводятся статистические данные по трём текстам.

3. Для каждого из текстов с помощью приложения AntConc были получены списки стоп-слов. В качестве основы фильтрации использовался список Фокса [14], который считается эталонным для английского языка. Файл со словами из этого списка был загружен в приложение с помощью опции *'add words from file'* в настройках (Preferences) функции '*Wordlist*'. Затем с помощью этой функции были получены для каждого из текстов списки знаменательных слов. Далее список знаменательных слов загружался в приложение, и они вычитались из текста. Таким образом, было выполнено последовательное двойное вычитание: сначала вычитались стоп слова и определялись знаменательные слова, затем вычитались знаменательные слова и находились стоп слова. В табл. 2 показаны первые пять стоп-слов из каждого из текстов с их частотностями.

Таблица 1

Статистические данные исходных текстов

Текст	Автор	Количество			
		токенов		уникальных слов	
		Оригинал	Выровненный	Оригинал	Выровненный
G1	J. Galsworthy	113149	-	9083	-
G2	J. Galsworthy	111891	-	8844	-
D	Ch. Dickens	163849	111891	10451	8633

Таблица 2

Статистические данные, полученные с помощью AntConc

Текст		Количество		Первые 5 стоп-слов	Частотность стоп-слов
		уникальных стоп-слов	токенов		
Файл	Название текста				
G1	The Man of Property	387	72299	the	5644
				of	3461
				and	2997
				he	2980
				to	2874
G2	Indian Summer of a Forsyte	379	71800	the	4696
				and	3354
				he	3138
				to	2873
				of	2722
D	Oliver Twist	382	68917	the	6844
				and	3663
				a	2750
				of	2707
				to	2618

Таблица 3

Числовые ряды и параметры текстов

Текст	Слово	Ранг	FR	FZ	DevZ	σ(DevZ)
G1	the	1	5644	5644	0	225,89
	of	2	3461	2822	639	
	and	3	2997	1881,33	1115,67	
G2	the	1	4696	4696	0	251,39
	and	2	3354	2348	1006	
	he	3	3138	1565,33	1572,67	
D	the	1	6844	6844	0	145,95
	and	2	3663	3422	241	
	a	3	2750	2281	469	

4. Для каждого стоп-слова в текстах был подсчитан коэффициент Ципфа и отклонение от него. Коэффициент Ципфа подсчитывался по формуле:

$$Z\left(w_{ij}\right) = F\left(w1_j\right) / R\left(w_{ij}\right) \ , \qquad (2)$$

где $F\left(w1_j\right)$ – частотность первого по рангу слова в некотором *j*-ом тексте, а *R* – номер ранга слова. Если $F\left(w1_j\right)$ = 4696, то распределение Ципфа $Z\left(w2_j\right)$ = 2348. Отклонение от распределения Ципфа подсчитывалось как разница по модулю:

$$\mathrm{Dev}(w_{ij}) = |Z(w_{ij}) - F(w_{ij})| \ . \qquad (3)$$

При этом, в соответствии с формулой (1), $F(w1j)=Z(w1j)$, соответственно, отклонение для первого по рангу слова равно нулю.

В приведённом примере: если частотность второго по рангу слова $F(w2)=3354$, то

$$\mathrm{Dev}(w2_j)=|2822-3354|=639 .$$

Таким образом было создано три числовых ряда: 1) с реальными частотностями слов (FR), полученными с помощью AntConc; 2) с коэффициентами Ципфа (FZ); 3) с разницей по модулю между двумя показателями DevZ.

5. Для последнего числового ряда, в котором указывались разницы между реальными частотностями и коэффициентами Ципфа, было вычислено среднее квадратичное отклонение по формуле:

$$\sigma\left(DevZ_j\right)=\sqrt{Var\left(DevZ_j\right)} \ , \qquad (4)$$

где *Var* – дисперсия, а *DevZ* – указанный числовой ряд. Полученное отклонение стало параметром, на основе которого вычислялись расстояния между входными и эталонным текстом. Вычисления производились в табличном процессоре MS Excel 2013, дроби округлялись до двух десятичных знаков.

В табл. 3 приводятся данные о числовых рядах (первые три слова) и параметрах текстов.

6. Вычисление расстояний между текстами. В соответствии с принятой методикой, вычисляются два расстояния: Dis(G1,G2) и Dis(G1, D). Если первое расстояние существенно меньше, чем второе, то можно считать, что G1, G2 написаны одним автором и предложенная методика работает адекватно. Для того, чтобы сделать такой вывод необходимо, чтобы разница в расстояниях превышала статистическую погрешность, которая обычно оценивается в 5% (если исходить из диапазона от 0 до 100%). Расстояния вычислялись по формулам:

$$\mathrm{Dis}(G1, G2)=|\sigma\left(DevZ_{G1}\right)-\sigma\left(DevZ_{G2}\right)|= \\ =|225,89\text{-}251,39|=25,50 \qquad (5)$$

$$\mathrm{Dis}(G1, D)=|\sigma\left(DevZ_{G1}\right)-\sigma\left(DevZ_{D}\right)|= \\ =|225,89\text{-}79,94|=79,94 \ . \qquad (6)$$

Разница между двумя величинами составляет 68,1%, т. е. расстояние между текстами, написанными одним автором (Дж. Голсуорси) на 68,1% меньше, чем расстояние между текстами Дж. Голсуорси и Ч. Диккенса, что подтверждает эффективность нашей методики.

ЗАКЛЮЧЕНИЕ

Предложена методика авторской атрибуции текстовых документов на основе статистических данных, полученных с помощью конкорданса и вычислений, выполненных в табличном процессоре MS Excel. В настоящее время для автоматического анализа текстовых документов активно применяются пакеты библиотек и инструментов на платформе различных языков программирования. Наиболее популярны *Natural language toolkit* для языка *Python* и пакет *Text mining* для языка *R*, позволяющие как получать исходные статистические данные, так и проводить соответствующие вычисления в одной среде. Однако изучение языков программирования представляет существенную трудность для студентов, обучающихся по гуманитарным специальностям. Как показывает наш опыт, предложенный вариант с использованием конкорданса и табличного процессора осваивается студентами гораздо легче и позволяет получать знание и навыки, необходимые в дальнейшем для изучения языков программирования и создания лингвистического программного обеспечения, к которым относятся:

понятие токенизации как процесса лексической декомпозиции текста и понятие токена как результата процесса токенизации;

понятие ранжированного списка;

понятие о различии между стоп-словами и знаменательными словами;

понятие вычитания и пересечения множеств, включающих лингвистические единицы текста;

понятие отклонения от закона Ципфа и умение вычислять отклонения с использованием табличного процессора;

представление об основных подходах к автоматической классификации текстовых документов.

Исходя из дидактического принципа последовательности, нами был представлен наиболее простой вариант авторской атрибуции, предусматривающий сопоставление распределения стоп-слов в двух входных текстах и эталонном. Далее можно повышать трудность задачи, увеличивая количество входных текстов, а также сопоставляя тексты друг с другом без использования эталонного текста.

Предложенная нами методика автоматической классификации на основе отклонений от распределения Ципфа и среднего квадратичного отклонения даёт адекватные результаты, не требуя сложных вычислений и применения алгоритмов машинного обучения. Эта методика может быть применена не только с целью авторской атрибуции, но и для решения других классификационных задач, таких как распознавание плагиата, тематическая категоризация, распознавание жанра произведений.

СПИСОК ЛИТЕРАТУРЫ

1. Francis W.N., Kucera H. Computational analysis of present day American English. Providence. – R.I.: Brown University Press, 1967. – 424 p.
2. Anthony L. AntConc 3.5.8. – 2019. – Tokyo, Japan: Waseda University. – URL: https://www.laurenceanthony.net/software (дата обращения: 25.06.2020)
3. Яцко В.А. Закон Ципфа как показатель эталонного распределения данных // Роль и место информационных технологий в современной науке. – Омск, 2016. – С. 48-50. – URL: https://os-russia.com/SBORNIKI/KON-129.pdf#page=48 (дата обращения: 25.06.2020)
4. Яцко В.А. Метод автоматической классификации текстов, основанный на законе Ципфа // Научно-техническая информация. Сер 2. – 2015. – № 5. – С. 19-24.

5. Amarasinghe K., Manic M., Hruska R. Optimal stop word selection for text mining in critical infrastructure domain // Resilience Week (RWS). – Philadelphia, PA, 2015. – P. 1-6. DOI: 10.1109/RWEEK.2015.7287440. – URL: https://www.researchgate.net/publication/281377695_Optimal_Stop_Word_Selection_for_Text_Mining_in_Critical_Infrastructure_Domain#fullTextFileContent.
6. Singhal A., Buckley C., Mitra M. Pivoted document length -normalization // SIGIR Forum. – 2017. – Vol. 51, № 2. – P. 176–184. DOI: 10.1145/3130348.3130365. – URL: http://singhal.info/ pivoted-dln.pdf (дата обращения: 25.06.2020)
7. Sinclair J. Reading concordances. – London: Longman, 2003. – 180 p. – URL: http://www.twc.it/rc/readings.htm (дата обращения: 25.06.2020)
8. Concapp.rar. – URL: https://docs.zoho.com/file/1hhltd2e9dd94a00d4aec88094394 b1d42255 (дата обращения: 25.06.2020).
9. Scott M. WordSmith Tools version 8. – 2020. – Stroud: Lexical Analysis Software. – URL: https://lexically.net/wordsmith/?gclid=EAIaIQobChMIpLbtuSV6gIVkpIYCh208guuEAAYASAAEgKAAvD_BwE (дата обращения: 25.06.2020).
10. WordStat / Provalis Research. – 2020. – URL: https://provalisresearch.com/products/content-analysis-software/ (дата обращения: 25.06.2020).
11. Free eBooks – Project Gutenberg. – 2020. – URL: https://www.gutenberg.org/ (дата обращения: 25.06.2020).
12. Dendamrongvit S., Vateekul P., Kubat M. Irrelevant attributes and imbalanced classes in multi-label text-categorization domains // Intelligent data analysis. – 2011. – Vol. 15, № 6. – P. 843-859. – URL: https://content.iospress.com/articles/intelligent-data-analysis/ida00499 (дата обращения: 25.06.2020).
13. Yatsko V. Zonal text processing // Digital scholarship in the humanities. – 2016. – Vol. 31, Issue 4. – P. 773–781. DOI: https://doi.org/10.1093/llc/fqv022
14. Fox C. A stop list for general text // SIGIR Forum year. – 1989. – Vol. 24, № 1-2. – P. 19–21. DOI: 10.1145/378881.378888. – URL: https://dl.acm.org/doi/pdf/10.1145/378881.378888 (дата обращения: 25.06.2020).

Материал поступил в редакцию 28.06.20.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

http://www.udcc.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ВИНИТИ РАН, как единственный в России владелец лицензии Консорциума УДК, предлагает издания УДК полного четвертого издания на русском языке в печатном и электронном виде:

1. Таблицы УДК

УДК. Том I Общая методика применения УДК. Вспомогательные таблицы. Основные таблицы. Общий отдел. Алфавитно-предметный указатель к Общему отделу

УДК. Том II **1/3** Философия. Психология. Религия. Богословие. Общественные науки (только электронное издание)

УДК. Том III **5/54** Математика. Естественные науки (только электронное издание)

УДК. Том IV **55/59** Геологические и биологические науки (только электронное издание)

УДК. Том V **6/61** Медицинские науки (только электронное издание)

УДК. Том VI (часть 1) 6/621 Прикладные науки. Технология. Инженерное дело (только электронное издание)

УДК. Том VI (часть 2) 622/629 Техника. Инженерное дело (только электронное издание)

УДК. Алфавитно-предметный указатель к т. VI (1 и 2 части) (только электронное издание)

УДК. Том VII **63/65** Сельское хозяйство. Домоводство. Управление предприятием (только электронное издание)

УДК. Том VIII **66** Химическая технология. Химическая промышленность. Пищевая промышленность. Металлургия. Родственные отрасли (только электронное издание)

УДК. Том IX **67/69** Различные отрасли промышленности и ремесел. Строительство (только электронное издание)

УДК. Том X **7/9** Искусство. Спорт. Филология. География. История.

УДК. АПУ (с в о д н ы й) к полному 4-му изданию

УДК. Изменения и дополнения. Выпуск 2 (к т.т. 1–3) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 3 (к т.т. 1–6) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 4 (к т.т. 1–7) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 5 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 6 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 7 (к т.т. 1–10), 2017 г. (только электронное издание)

Для подписки необходимо направить заявку по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499-155-42-85, 499-151-78-61

E-mail: feo@viniti.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!
УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ (УДК)

НОВОЕ ИЗДАНИЕ
УДК. ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ.
Выпуск 7

Содержание выпуска:

В настоящем электронном издании помещены **изменения и дополнения**, опубликованные Консорциумом УДК в выпусках 32 и 33 «Extentions and corrections to the UDC»:

ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ К ТАБЛИЦАМ ОБЩИХ ОПРЕДЕЛИТЕЛЕЙ

– Опубликованы изменения к **Таблице IG.** Общие определители времени

ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ К ОСНОВНЫМ ТАБЛИЦАМ УДК

Опубликованы изменения к классам:

– **2 Религия. Богословие**
– **33 Экономика. Народное хозяйство. Экономические науки**
– **582 Систематика растений**
– **551.7 Историческая геология.**

Для удобства пользователей издание открывает Общая методика применения Универсальной десятичной классификации.

Для подписки необходимо направить заявку по адресу:
125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 499-155-42-52, 499-155-42-85, 499-151-78-61
E-mail: typo@viniti.ru, feo@viniti.ru
http://www.udcc.ru