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. № 12 Москва 2020

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.89:510.6

М.И. Забежайло

О некоторых оценках сложности вычислений при прогнозировании свойств новых объектов средствами характеристических функций*

Обсуждаются возможности и инструменты оценки качества результатов интеллектуального анализа данных в задачах диагностического типа. Надежность (неоспариваемость) эмпирических зависимостей, формируемых в процессе обучения (интерполяции-экстраполяции) на прецедентах, оценивается средствами специального логического инструментария – характеристическими функциями. Порождение таких функций по имеющейся выборке эмпирических данных базируется на анализе сходства описаний прецедентов, уточняемого как бинарная алгебраическая операция. Представлены некоторые оценки сложности вычислений, сопутствующих применению предлагаемой математической техники характеристических функций при прогнозировании (диагностике) свойств вновь изучаемых прецедентов.

* Статья частично содержит результаты проекта «Математические основы интеллектуального анализа больших данных», выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 15.08.2019 № 7/1251/2019

Ключевые слова: *интеллектуальный анализ данных, задачи диагностического типа, эмпирические зависимости, анализ причинности, оценки сложности вычислений*

DOI: 10.36535/0548-0027-2020-12-1

ВВЕДЕНИЕ

Все более пристальный интерес сегодня привлекает к себе новый актуальный класс объектов, требующих развития адекватных средств анализа (математических моделей, методов и алгоритмов, а также реализующих их компьютерных систем). Речь идет о последовательностях расширяющихся описаниями новых прецедентов коллекций эмпирических данных, используемых для решения задач диагностического типа. В части приложений это, например, задачи поддержки принятия врачебных решений, основанных на опыте ранее поставленных медицинских диагнозов; проблематика выявления мошенничества в финансовой сфере; задачи поддержании работоспособности сложных технических систем; организация противодействия компьютерным атакам при обеспечении информационной безопасности и др.

С процедурной точки зрения критически важной особенностью математических задач этого типа оказывается необходимость анализа динамически изменяемого во времени объекта – базы фактов (БФ), аккумулирующей текущий объем накапливаемых эмпирических данных. Цель подобного исследования – *интерполяция,* т. е. аналитическое описание всех имеющихся эмпирических фактов («точечных» событий) средствами эмпирических зависимостей (ЭЗ) того или иного класса, которые далее можно было бы использовать для «диагностирования» вновь изучаемых фактов (прецедентов) проверкой *экстраполируемости* на эти факты найденных ЭЗ (возможности *экстраполировать* какие-либо из найденных ЭЗ на вновь анализируемое «точечное» событие).

Решающее требование для реализации такого подхода – необходимость учета (идентификации и отслеживания в процессе пополнении БФ новыми фактами) *устойчивых,* т. е. *наследуемых* при расширении анализируемой базы фактов эмпирических зависимостей. Именно такие *устойчивые* эмпирические зависимости существенны для корректного, в том числе сохраняющего *наследуемость* ранее уже принятых *диагностических решений,* анализа новых прецедентов.

Наличие эмпирических зависимостей, наследуемых при расширении новыми фактами используемой БФ, может служить индикатором *репрезентативности* текущего состояния БФ (как своего рода обучающей выборки для построения соответствующих ЭЗ). Контроль такой *репрезентативности* текущего состояния БФ необходим (как уже было сказано выше) для обеспечения корректности прогноза свойств новых прецедентов, базирующегося на использовании уже сформированных на имеющейся БФ эмпирических зависимостей. Особое значение здесь имеет *множественный* характер формируемых ЭЗ, что требует разработки таких инструментов анализа данных, которые позволяли бы обозревать при принятии решений *все* множество порождаемых на конкретной базе фактов зависимостей.

ДИАГНОСТИКА ЦЕЛЕВОГО ЭФФЕКТА И КАУЗАЛЬНАЯ РЕПРЕЗЕНТАТИВНОСТЬ ИМЕЮЩИХСЯ ЭМПИРИЧЕСКИХ ДАННЫХ

Обсуждаемое представление о *репрезентативности* базы фактов, используемой для формирования ЭЗ, может быть естественным образом сопоставлено с содержательным (опирающимся на неформальные знания об исследуемой предметной области – медицинской диагностике и др.) представлением о природе *причинности* возникновения анализируемых эффектов. Действительно, если в используемом для формализованного описания анализируемых прецедентов языке представления знаний (ЯПЗ) выразимы необходимые характеристики *каузальности* («*причин*»[1] возникновения) изучаемого эффекта, то необходимо, чтобы формируемые причинные эмпирические зависимости были однозначным образом сопоставимы с уже имеющимися в БФ примерами (фактами наличия исследуемого эффекта) и контрпримерами (фактами отсутствия такого эффекта в соответствующих случаях\прецедентах). Здесь наличие причины (соответствующей комбинации факторов влияния) – обязательное условие для каждого примера в БФ, а ее отсутствие – обязательное условие для каждого контрпримера. Если же какой-либо из контрпримеров «содержит» какую-либо из причин (т. е. соответствующая причинная эмпирическая зависимость «проходит»[2] через данный контрпример), то налицо очевидное содержательное противоречие. Это является сигналом о невозможности каузально-корректно разделить примеры и контрпримеры данной БФ с помощью формируемых эмпирических зависимостей. Именно в такой ситуации текущую базу фактов и предлагается рассматривать как *каузально нерепрезентативную* (разумеется, с точностью до используемого в ее анализе языка представления знаний). Наоборот, легко видеть, что *каузально-репрезентативная* БФ представляет *непротиворечивое* основание для формирования соответствующих диагностических заключений: принимаемые решения (о наличии или отсутствии конкретного исследуемого эффекта) будут *наследоваться* вдоль всей последовательности сохраняющих *репрезентативность* расширений анализируемой базы фактов, так как в основе всех этих заключений будут лежать *наследуемые* представления о *причинности* (формализуемые в виде соответствующих эмпирических зависимостей).

[1]Взятие в кавычки здесь отражает возможную ситуацию, когда в используемом ЯПЗ выразима лишь *часть* факторов, характеризующих механизм причинности возникновения соответствующих эффектов, а для исчерпывающего описания причинности необходим более полный ЯПЗ.

[2]Выполняется на данном контрпримере.

ОБЩИЕ ПРЕДСТАВЛЕНИЯ О МАТЕМАТИЧЕСКОЙ ТЕХНИКЕ ХАРАКТЕРИСТИЧЕСКИХ ФУНКЦИЙ

В качестве «инструмента» для анализа обсуждаемых последовательностей расширяющихся баз фактов (**FB**) в работах [1-3] была предложена математическая техника так называемых характеристических функций (*ChF*), являющихся каузально-ориентированным подклассом семейства частичных функций, интерполирующих выборки прецедентов – примеров и контрпримеров диагностируемого явления. По своей «архитектуре» каждая *ChF* – это такое особым способом формируемое (см. далее) логическое условие (связывающее специальным образом выделяемые элементы описаний прецедентов текущей **FB**), которое принимает значение:

«*истина*» на всех фактах ϕ (примерах) текущей базы фактов **FB**, характеризуемых наличием анализируемого целевого свойства (диагностируемого явления);

«*ложь*» на всех фактах ϕ (контрпримерах) текущей базы фактов **FB**, характеризуемых отсутствием анализируемого целевого свойства:

$$\forall \phi \, [\, (\phi \in \mathbf{FB}) \supset (\, ChF(\mathbf{FB}) \models \phi)].$$

Процедура построения каждой *ChF* по имеющейся **FB** определяется (подробнее см. [1, 3]) следующей схемой.

i. Формируется формализованное описание прецедентов, например, в виде кортежа значений определенных признаков и др.

ii. На формализованных описаниях прецедентов определяется бинарная (идемпотентная, симметричная и ассоциативная [4]) операция сходства $\otimes$. (Для множеств значений признаков это может быть, в частности, операция $\cap$ пересечения множеств).

iii. Отношение сходства прецедентов определяется по непустому результату вычисления операции сходства прецедентов: два прецедента сходны, если результат применения операции $\otimes$ к их описаниям не является пустым объектом $\ulcorner\varnothing\urcorner$.

iv. Для каждого прецедента O из множества **FB** класс сходства $\mathbf{T}(O)$ всех сходных с ним прецедентов из **FB** формируется объединением в $\mathbf{T}(O)$ всех сходных с O (в смысле отношения сходства из п. iii) элементов множества **FB.**

v. Классы эквивалентности прецедентов определяются выделением из каждого из найденных ранее классов сходства $\mathbf{T}(O)$ всех подклассов, каждый из которых задаваем фиксированным непустым результатом вычисления сходства $\otimes$ исходных прецедентов. При этом по каждому отношению эквивалентности, задаваемому каким-либо непустым результатом V вычисления операции сходства $\otimes$, исходное множество прецедентов (соответствующего знака) **FB** разбивается на один большой класс эквивалентности $\mathbf{E}_V(\mathbf{FB})$, который сформирован всеми содержащими данное V прецедентами (соответствующего знака) из **FB**, а вместе с ним также соответствующим числом одноэлементных классов, сформированных всеми не вошедшими в $\mathbf{E}_V(\mathbf{FB})$ прецедентами (соответствующего знака) из **FB.**

vi. Сформированные классы сходства и классы эквивалентности распадаются на две части – на те, что построены на примерах, и те, что построены на контрпримерах.

vii. Специальное условие з*апрета на контрпримеры* (*ЗКП*) требует, чтобы ни один из примеров класса α (где $\alpha \in \{+,-\}$, т. е. для $\alpha = +$ это примеры, а для $\alpha =$ – это контрпримеры из исходного множества прецедентов) не вкладывался бы ни в один из классов эквивалентности, сформированных на контрпримерах противоположного знака $-\alpha$.

viii. Те классы сходства (построенные на примерах, где $\alpha = +$), для которых выполнено условие запрета на контрпримеры используются для формирования покрытия всего исходного множества прецедентов (примеров и контрпримеров), и если таковое существует, то по нему (взятием дизъюнкции по всем формирующим это покрытие классам сходства конъюнкций тех значений признаков, которые – как результат вычисления операции $\otimes$ – формируют каждый из таких классов сходства) и строится соответствующая *ChF*. Каждое полученное таким способом дизъюнктивно-конъюнктивное логическое условие собственно и представляет (см., например, [3]) характеристическую функцию *ChF*, сопоставляемую имеющейся коллекции прецедентов **FB** и принимающую истинностные значения:

– «*истина*» тогда и только тогда, когда данный факт характеризуется наличием анализируемого целевого свойства;

– «ложь» тогда и только тогда, когда данный факт характеризуется отсутствием анализируемого целевого свойства,

и наоборот (см. Утверждение 2 в работе [3]):

– на каждом факте ϕ из текущей **FB**, характеризуемом наличием анализируемого целевого свойства, характеристическая функция $ChF_i(\mathbf{FB})$ принимает значение «*истина*», а значение «ложь» тогда и только тогда, когда данный факт характеризуется отсутствием анализируемого целевого свойства.

Непустота множества $\boldsymbol{ChF}(\mathbf{FB})$ всех порождаемых на текущей базе фактов **FB** характеристических функций *ChF* может рассматриваться как своего рода «индикатор» *каузальной репрезентативности* анализируемой **FB**. Фактически, это – «сигнал» о достаточности накопленных в текущей **FB** описаний прецедентов для *каузально* корректного (обеспечиваемого выполнимостью условия запрета на контрпримеры) разделения примеров и контрпримеров и, как следствие, о возможности их использования (как обучающей выборки) для корректной (см. выше) «диагностики» новых (ранее еще не изученных) прецедентов. Таким образом, каждая из характеристических функций представляет формируемую на соответствующей **FB** эмпирическую зависимость, а каждая из таких зависимостей, отражая определенные механизмы *причинности* возникновения диагностируемого явления, может быть использована при прогнозировании целевого эффекта на вновь анализируемом прецеденте, т. е. является проверкой возможности экстраполировать эту зависимость на данный прецедент.

ЗАДАЧА О ПРОГНОЗИРОВАНИИ СВОЙСТВ ВНОВЬ АНАЛИЗИРУЕМОГО ПРЕЦЕДЕНТА СРЕДСТВАМИ ХАРАКТЕРИСТИЧЕСКИХ ФУНКЦИЙ

Итак, пусть база фактов **FB** представляет объединение двух наборов прецедентов (пополняемых в процессе накопления новых эмпирических данных) – множества $\Omega^+=\{O^+_1,O^+_2, \dots ,O^+_m+\}$ примеров O^+_i (прецедентов, которые характеризуются наличием анализируемого целевого эффекта) и множества $\Omega^- =\{O^-_1,O^-_2, \dots ,O^-_m\}$ контрпримеров O^-_j (прецедентов, которые характеризуются отсутствием этого целевого эффекта). Пусть также зафиксировано некоторое множество $\boldsymbol{U} = \{a_1, a_2, \dots , a_n\}$, образующих a_i, из которых формируются как примеры из множества $\Omega^+ = \{O^+_1, O^+_2, \dots , O^+_m+\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$, так и контрпримеры из множества $\Omega^- = \{O^-_1, O^-_2, \dots , O^-_m\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$. Пусть $\otimes$ есть бинарная алгебраическая операция (сходства прецедентов из **FB**), которая определена отдельно и на примерах из Ω^+, и контрпримерах из Ω^- так, что удовлетворяет трем стандартным условиям (для всех i,j и k одновременно из множества примеров Ω^+ или же одновременно из контрпримеров Ω^-):

1. $O_i \otimes O_i = O_i$
2. $O_i \otimes O_j = O_j \otimes O_i$
3. $O_i \otimes O_j \otimes O_k = (O_i \otimes O_j) \otimes O_k = O_i \otimes (O_j \otimes O_k)$.

Результатом применения операции $\otimes$ к паре объектов O_i и O_j будет их общий подобъект O_{ij}, а в ситуации, когда O_i и O_j не имеют общих частей, – пустой подобъект $\ulcorner\varnothing\urcorner$. Таким образом, в рассматриваемом здесь случае $\otimes$ есть операция пересечения множеств $\cap$, а O_{ij} – максимальное общее подмножество O_i и O_j .

Отношение $\sqsubseteq$ вложимости подобъектов в объекты из **FB** определяется естественным образом с помощью используемой операции $\otimes$ сходства прецедентов: вложение $O_{ij} \sqsubseteq O_i$ подобъекта O_{ij} в объект O_i имеет место тогда и только тогда, когда $O_i \otimes O_{ij} = O_{ij}$.

Задача ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХАРАКТЕРИСТИЧЕСКАЯ ФУНКЦИЯ (ХФ)*** о прогнозировании (диагностике) средствами характеристических функций наличия или же, наоборот, отсутствия целевого эффекта у вновь анализируемого прецедента формулируется следующим образом:

Дано:
Два множества прецедентов – примеров $\Omega^+ = \{O^+_1, O^+_2, \dots , O^+_m+\}$ и контрпримеров $\Omega^- = = \{O^-_1, O^-_2, \dots , O^-_m\}$, объединение которых $\Omega^+ \cup \Omega^-$ формирует текущее состояние базы фактов **FB**, а также новый (не содержащийся в текущей **FB**) прецедент O^τ .

Найти (проверить):
Существует ли в множестве ***ChF*(FB)** всех характеристических функций ChF_i, порождаемых на текущей базе фактов **FB**, такая функция ChF_0 , которая обращается в истину на этом новом прецеденте O^τ ?

В случае существования хотя бы одной подобной ChF_0 мы будем говорить о прогнозировании (формировании положительного диагноза для) *наличия* целевого эффекта на новом прецеденте O^τ , а при отсутствии таких функций – об *отсутствии* целевого эффекта на этом новом прецеденте. Если рассуждать неформально, то при обнаружении подобной характеристической функции есть все основания утверждать[3], что текущая база фактов позволяет сформировать *эмпирическую зависимость* каузального характера, успешно *экстраполируемую* на соответствующий новый прецедент, требующий диагностики наличия у него анализируемого целевого эффекта.

ЭФФЕКТИВНАЯ РАЗРЕШИМОСТЬ ЗАДАЧИ ОБ ОЦЕНКЕ КАУЗАЛЬНОЙ РЕПРЕЗЕНТАТИВНОСТИ КОНКРЕТНОЙ БАЗЫ ФАКТОВ

Легко видеть, что простейшим необходимым условием существования искомой ChF_0 с требуемыми (см. выше) свойствами является каузальная репрезентативность текущей базы фактов **FB** (т.е. непустота множества ***ChF*(FB)**). Соответствующая задача ***РЕПЕРЗЕНТАТИВНОСТЬ БАЗЫ ФАКТОВ*** формулируется следующим образом:

Дано:
Два множества прецедентов – примеров $\Omega^+ = \{O^+_1, O^+_2, \dots , O^+_m+\}$ и контрпримеров $\Omega^- = \{O^-_1, O^-_2, \dots , O^-_m\}$, объединение которых $\Omega^+ \cup \Omega^-$ формирует текущее состояние базы фактов **FB**.

Найти (проверить):
Существует ли в множестве ***ChF*(FB)** всех характеристических функций ChF_i , порождаемых на текущей базе фактов **FB**, хотя бы один элемент *ChF* ?

Эффективную (требующую объема вычислений, который полиномиально быстро растет[4] с линейным ростом размеров **FB**) разрешимость задачи ***РЕПРЕЗЕНТАТИВНОСТЬ БАЗЫ ФАКТОВ*** демонстрирует

Утверждение 1. Задача о непустоте множества ***ChF*(FB)** характеристических функций, порождаемых на текущей базе фактов **FB** принадлежит к классу *P* полиномиально быстро разрешимых комбинаторных проблем.

Доказательство. Рассмотрим множество ***Bin***(Ω^+) всех попарных сходств примеров из множества Ω^+. Выделим в нем подмножество ***ЗКП-Bin***(Ω^+) таких, на которых выполняется условие з*апрета на контпримеры* (*ЗКП* – см. ранее п.vii – в описании процедуры формирования характеристических функций). Построим множество $\Omega^+_{\textit{ЗКП-Bin}}$, представляющее объединение всех тех примеров из Ω^+, которые участвуют в формировании $\Omega^+_{\textit{ЗКП-Bin}}$. Несложно убедиться, что множество ***ChF*(FB)** непусто тогда и только тогда, когда $\Omega^+ = \Omega^+_{\textit{ЗКП-Bin}}$. Действительно, выполнимость условия *ЗКП* для любого конкретного элемента – парного сходства $Bin_i(\Omega^+)$ из множества ***Bin***(Ω^+) – означает, что у формируемого на $Bin_i(\Omega^+)$ подобъекта $Sim_i(\Omega^+)$ двух порождающих его объектов-примеров имеется такой соответствующий подобъект $\underline{Sim_j}(\Omega^+)$, формируемый по-парными, или по-троечными или … и т.д. – до $\mathbf{m}^+$-сходствами примеров из Ω^+, который не вкладывается (в смысле отношения $\sqsubseteq$) ни в один из контрпримеров из множества Ω^-. При этом его не-

[3] И это утверждение *невозможно* будет *оспорить* на *текущей* базе фактов!

[4] Мы будем называть такие задачи полиномиально быстро разрешимыми.

вложимость в любой из контрпримеров обеспечивает выполнение этого же свойства и для содержащего его подобъекта $Sim_i(\Omega^+)$. И наоборот: если такого (удовлетворяющего условию *ЗКП*) подобъекта $\underline{Sim_i}(\Omega^+)$ в $Sim_i(\Omega^+)$ не окажется, то порождающая $Bin_i(\Omega^+)$ пара объектов не может «участвовать» в формировании искомой характеристической функции *ChF*. Полиномиально-быстрый обзор множества $\boldsymbol{Bin}(\Omega^+)$ определяется его размерами, растущими (с линейным ростом размеров множества объектов-примеров Ω^+) не быстрее чем полином второй степени от числа элементов в Ω^+.

ЭФФЕКТИВНАЯ РАЗРЕШИМОСТЬ ЗАДАЧИ ОБ НЕПУСТОТЕ МНОЖЕСТВА ХАРАКТЕРИСТИЧЕСКИХ ФУНКЦИЙ, ПРИНИМАЮЩИХ ЗНАЧЕНИЕ «ИСТИНА» НА ВНОВЬ ДИАГНОСТИРУЕМОМ ПРЕЦЕДЕНТЕ

Итак, задача об оценке каузальной репрезентативности конкретной базы фактов эффективно (полиномиально быстро) разрешима. Следующим шагом будет оценка непустоты множества характеристических функций, формируемых на текущей базе фактов и обращающихся в истину на заданном вновь диагностируемом прецеденте O^τ.

Утверждение 2. Задача ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХФ*** о непустоте множества ***ChF*(FB)** характеристических функций, порождаемых на текущей базе фактов **FB** и обращающихся в *истину* на заданном вновь диагностируемом прецеденте O^τ, принадлежит к классу *P* полиномиально быстро разрешимых комбинаторных проблем.

Доказательство. Рассмотрим множество $\boldsymbol{Bin}(\Omega^+)$ всех попарных сходств примеров из множества Ω^+. Выделим в нем подмножество $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$ таких, на которых выполняется условие *запрета на контрпримеры* (*ЗКП* – см. п vii – в описании процедуры формирования характеристических функций). Проверим совпадает ли множество всех задействованных при формировании этого подмножества $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$ примеров с исходным множеством примеров Ω^+. Если это не так, то для идентификации отсутствия искомой характеристической функции (и пустоты всего множества характеристических функций, формируемых на анализируемой базе фактов) понадобился объем вычислений, ограниченный сверху полиномом (см. сложность формирования $\boldsymbol{Bin}(\Omega^+)$ и проверку условия ЗКП при формировании множества $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$) от характерных размеров **FB**.

В противном случае для каждого V_{ij} из множества $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$ найдем его общую часть $V_{ij}\,|_{O^\tau}$ с описанием вновь диагностируемого прецедента O^τ:

$$V_{ij}\,|_{O^\tau} = V_{ij} \otimes O^\tau = V_{ij} \cap O^\tau.$$

Далее, сперва для каждой образующей a_k из исходного $\boldsymbol{U} = \{a_1, a_2, \dots, a_n\}$, содержащейся в рассматриваемом $V_{ij}\,|_{O^\tau}$, построим множество[5] $[\{a_k\}]_{\boldsymbol{U},\Omega^+}$ всех таких a_{kl} из $\boldsymbol{U}$, что они вкладываются в каждый из примеров из Ω^+ исключительно *вместе* с данной a_k. А затем выделим из множества $[\{a_k\}]_{\boldsymbol{U},\Omega^+}$ те a_{kl}, замыкание $[\ \_\ ]_{\boldsymbol{U},\Omega^+}$ (см. сноску [5]) которых выводит за пределы рассматриваемого $V_{ij}\,|_{O^\tau}$ (содержит образующие, не входящие в $V_{ij}\,|_{O^\tau}$). Удалив из $V_{ij}\,|_{O^\tau}$ все такие a_{kl}, получим множество $V^*_{ij}\,|_{O^\tau}$. Далее проверим, выполняется ли на $V^*_{ij}\,|_{O^\tau}$ условие з*апрета на контрпримеры*. В случае пустоты $V^*_{ij}\,|_{O^\tau}$ или же невыполнимости на нем условия *ЗКП*, перейдем к новому элементу множества $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$. В противном случае (непустота $V^*_{ij}\,|_{O^\tau}$ и выполнимость *ЗКП*) зарезервируем найденное $V^*_{ij}\,|_{O^\tau}$ для последующего формирования искомой характеристической функции.

Несложно убедиться, что $V^*_{ij}\,|_{O^\tau}$ замкнуто (т.е. $[V^*_{ij}\,|_{O^\tau}]_{\boldsymbol{U},\Omega^+} = V^*_{ij}\,|_{O^\tau}$) тогда и только тогда, когда оно получено удалением из $V_{ij}\,|_{O^\tau}$ всех a_{kl} с описанными выше характеристиками. Действительно, если $V^*_{ij}\,|_{O^\tau}$ замкнуто (т.е. $[V^*_{ij}\,|_{O^\tau}]_{\boldsymbol{U},\Omega^+} = V^*_{ij}\,|_{O^\tau}$), то в нем нет таких a_{kl}, каждая из которых входила бы в примеры из Ω^+ исключительно вместе с некоторой образующей $a_{kl\text{-}out}$, лежащей вне $V^*_{ij}\,|_{O^\tau}$ (т. е. такой, что $a_{kl\text{-}out} \notin V^*_{ij}\,|_{O^\tau}$). И наоборот, если в $V^*_{ij}\,|_{O^\tau}$ нет такой образующей, что в примеры из Ω^+ она входит только вместе с некоторой образующей $a_{kl\text{-}out}$, лежащей вне $V^*_{ij}\,|_{O^\tau}$, то данное $V^*_{ij}\,|_{O^\tau}$ – замкнуто (по построению из $\boldsymbol{Bin}(\Omega^+)$ по предложенному выше способу).

После проверки описанным выше способом всех парных сходств из множества $\boldsymbol{ЗКП\text{-}Bin}(\Omega^+)$, выясним, покрывают ли «задействованные» (для построения искомой характеристической функции) в них примеры все множество Ω^+. Если да, то искомая (принимающая значение *истина*) характеристическая функция над анализируемой базой фактов **FB** существует (и может быть сформирована, в частности, всеми найденными $V^*_{ij}\,|_{O^\tau}$). В противном случае такой характеристической функции (принимающей значение *истина* на прецеденте O^τ и формируемой на данной **FB**) не существует.

Полиномиальная сложность вычислений, реализующих представленную схему рассуждений, определяется:

- не более чем квадратичной сложностью анализа парных сходств $\boldsymbol{Bin}(\Omega^+)$ примеров из Ω^+;
- линейной сложностью проверки выполнимости условия запрета на контрпримеры;
- не более чем квадратичной сложностью вычисления каждого из замыканий $[\ \_\ ]_{\boldsymbol{U},\Omega^+}$;
- линейно сложностью проверки «покрываемости» всего множества Ω^+ порождающими соответствующие $V^*_{ij}\,|_{O^\tau}$ примерами.

[5] Можно показать, что это множество представляет замыкание Галуа (см. например [4]) для одноэлементного множества образующих $\{a_k\}$ относительно алфавита $\boldsymbol{U}$ и построенного на нем множества примеров Ω^+ при использовании операции $\cap$ пересечения множеств в качестве операции сходства $\otimes$.

ЁМКОСТЬ МНОЖЕСТВА ХАРАКТЕРИСТИЧЕСКИХ ФУНКЦИЙ, ПРИНИМАЮЩИХ ЗНАЧЕНИЕ «ИСТИНА» НА ВНОВЬ ДИАГНОСТИРУЕМОМ ПРЕЦЕДЕНТЕ

Ответ на естественный вопрос: «Сколь большим по числу элементов может быть множество ***ChF***(**FB**) характеристических функций, порождаемых на текущей базе фактов **FB** и обращающихся в *истину* на заданном вновь диагностируемом прецеденте O^{τ} », дает

Утверждение 3. Задача ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХФ*** о непустоте множества ***ChF***(**FB**) характеристических функций, порождаемых на текущей базе фактов **FB** и обращающихся в *истину* на заданном вновь диагностируемом прецеденте O^{τ} , принадлежит к классу *#PC* перечислительно полных комбинаторных проблем.

Доказательство. Для доказательства принадлежности задачи ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХФ*** классу *#PC* продемонстрируем сводимость к ней полиномиально-сложным алгоритмом хорошо известной переборной задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ*** о выполнимости монотонной булевской функции, представленной в виде образованной лишь двухэлементными дизъюнкциями конъюнктивной нормальной формы (так называемой 2-КНФ - см., например, [5-7] и др.):

Дано:

Монотонная булевская функция **Φ** в виде 2-КНФ

$$\mathbf{\Phi}(x_1,x_2, \dots ,x_n) = \mathbf{\Phi_1}\&\mathbf{\Phi_2}\& \dots \&\mathbf{\Phi_m} = (x_{11}\vee x_{12}) \& (x_{21}\vee x_{22}) \& \dots \& (x_{m1}\vee x_{m2})$$

Найти:

Число наборов значений переменных $x_1,x_2, \dots ,x_n$, выполняющих функцию **Φ**.

Схема сведения:

- по заданной функции **Φ** полиномиально быстро строятся:

– пара множеств **U(Φ)**,**Ω(Φ)**, такая, что классы эквивалентности примеров из **Ω(Φ)** над множеством **U(Φ)** «перечисляют» все нули функции **Φ** – сперва так называемые «примитивные», а затем и их комбинации – «композитные» нули. При этом «примитивным» нулем функции **Φ** называется такой набор истинностных значений переменных, что **Φ** обращается в 0 только в одном из формирующих эту функцию конъюнктов: значение 0 принимают обе входящие в этот конъюнкт переменные; «композитными» считаются по определению все остальные нули функции **Φ**); а также

– предлагаемый для проверки выполнимости искомой характеристической функции новый прецедент O^{τ};

- каждое несократимое покрытие исходного множества Ω^+ примеров из **Ω(Φ)** *ЗКП*-классами эквивалентности примеров из Ω^+ позволяет сформировать некоторую характеристическую функцию;

- размер множества единиц функции **Φ** есть размер дополнения множества всех нулей функции **Φ** до множества всех возможных наборов значений переменных $x_1,x_2, \dots ,x_n$.

Пусть монотонная булевская функция **Φ** задана в виде 2-КНФ:

$$\mathbf{\Phi}(x_1,x_2, \dots ,x_n) = \mathbf{\Phi_1}\&\mathbf{\Phi_2}\& \dots \&\mathbf{\Phi_m} = (x_{11}\vee x_{12}) \& (x_{21}\vee x_{22}) \& \dots \& (x_{m1}\vee x_{m2}).$$

Каждому из конъюнктов $\mathbf{\Phi_i}$ (где $i\in\{1,m\}$) функции **Φ** $(x_1,x_2, \dots ,x_n)$ сопоставим множество из четырех «объектов» O^+_{i1}, O^+_{i2} , O^+_{i3} и O^+_{i4}, являющихся соответствующими (см. далее) подмножествами некоторого исходного «алфавита» **U** (точное описание которого также представлено далее):

$O^+_{i1} = \{a_{i1}, a_{i2}, a_{i5}\}$,
$O^+_{i2} = \{a_{i2}, a_{i3}, a_{i6}\}$,
$O^+_{i3} = \{a_{i1}, a_{i4}, a_{i6}\}$,
$O^+_{i4} = \{a_{i3}, a_{i4}, a_{i5}\}$,

а также дополнительную пару «объектов» O^-_{i1} и O^-_{i2} вида:

$O^-_{i1} = \{a_{i5}, b_{i1}\}$,
$O^-_{i2} = \{a_{i6}, b_{i2}\}$.

При этом будем считать, что

$$\mathbf{U} = \{a_{11},a_{12}, \dots a_{16},\ a_{i21},a_{i22}, \dots, a_{26}, \dots,\ a_{i1},a_{i2}, \dots ,a_{i6}, \dots, a_{m1},a_{m2}, \dots ,a_{m6}, b_{11},b_{12}, b_{21},b_{22},\ \dots, b_{i1},b_{i2}, \dots, b_{m1},b_{m2}\},$$

т. е. в **U** имеется 8m попарно различных элементов («образующих»).

Множество прецедентов $\Omega(\mathbf{\Phi}) = \Omega^+(\mathbf{\Phi}) \cup \Omega^-(\mathbf{\Phi})$, где $\Omega^+(\mathbf{\Phi})$ мы будем называть множеством *примеров*, а $\Omega^-(\mathbf{\Phi})$ – множеством *контрпримеров*. $\Omega(\mathbf{\Phi})$ формируется по заданной функции **Φ** следующим образом:

– прежде всего $\Omega^-(\mathbf{\Phi}) = \bigcup\limits_i \{O^-_{i1},O^-_{i2})\}$;

– далее для каждого конъюнкта $\mathbf{\Phi_i} = (x_{i1}\vee x_{i2})$, такого, что переменная x_{ij} впервые встречается в 2-КНФ-представлении функции **Φ** именно в нем, каждое последующее вхождение переменной x_{ij} также в любой другой конъюнкт $\mathbf{\Phi_k}$ влечет объединение соответствующих множеств O^+_{i1} с O^+_{k1} и O^+_{i4} с O^+_{k4}, если x_{ij} входит в $\mathbf{\Phi_k}$ как x_{k1}. В противном случае (т. е. если x_{ij} входит в $\mathbf{\Phi_k}$ как x_{k2}) производятся объединения O^+_{i2} с O^+_{k2} и O^+_{i3} с O^+_{k3} ;

– в полученном после завершения объединения всех соответствующих повторным вхождениям переменных x_{ij} во все конъюнкты исходной функции **Φ** множестве примеров оставляются лишь те (полученные объединением по всем повторным вхождениям переменных) множества «образующих», которые сопоставлены первым вхождениям соответствующих переменных в 2-КНФ функции **Φ**. Именно это усеченное множество примеров и представляет искомое множество $\Omega^+(\mathbf{\Phi})$.

Несложно убедиться, что множества $\Omega(\mathbf{\Phi})$, $\Omega^+(\mathbf{\Phi})$ и $\Omega^-(\mathbf{\Phi})$ формируются по заданной функции **Φ** полиномиально (от параметров n и m) сложным алгоритмом: из 8m «образующих» исходного алфавита **U** формируются 6m «кандидатов» в *примеры* (в элементы финального $\Omega^+(\mathbf{\Phi})$), каждый из которых содержит всего 3 элемента, а также 2m контрпримеров, каждый из которых содержит 2 элемента. «Кодировка» повторных вхождений каждой из переменных функции **Φ** в ее конъюнкты $\mathbf{\Phi_i}$ требует вычисления не более

чем (nx2m) объединений множеств, каждое из которых содержит не более 8m элементов.

Легко видеть, что минимальным покрытием сходствами примеров всех 4-х объектов (примеров) из каждого «блока» номер i – $<O^+_{i1},O^+_{i2},O^+_{i3},O^+_{i4}>$ являются всего три пары сходств:

$O^+_{i1} \cap O^+_{i2}, = \{a_{i2}\}$ и $O^+_{i3}, \cap O^+_{i4} = \{a_{i4}\}$,
$O^+_{i1} \cap O^+_{i3}, = \{a_{i1}\}$ и $O^+_{i2}, \cap O^+_{i4} = \{a_{i3}\}$, и
$O^+_{i1} \cap O^+_{i4}, = \{a_{i5}\}$ и $O^+_{i2}, \cap O^+_{i3} = \{a_{i6}\}$.

Однако последняя пара сходств не удовлетворяет условию з*апрета на контрпримеры* (см. п. vii – в описании процедуры формирования характеристических функций): первое вкладывается в контрпример O^-_{i1}, а второе – в контрпример O^-_{i2}. Таким образом, сопоставив первую пару сходств такой ситуации, когда обе переменные в конъюнкте $\mathbf{\Phi_i}$ принимают значение 0, а вторую пару сходств всем остальным ситуациям (т.е. ситуациям, когда хотя бы одна из переменных обращает данный конъюнкт в 1), можно получить комбинаторный объект – комбинацию нулевых и ненулевых значений конъюнктов исходной функции $\mathbf{\Phi_i}$, позволяющий перечислить все нули (сперва – «примитивные», а затем и «композитные» – как все комбинации «примитивных» (подробнее см. [8]). Таким образом, каждое минимальное покрытие $Cov_j(\mathbf{\Phi})$ множества *примеров* $\Omega^+(\mathbf{\Phi})$ соответствующими *ЗКП*-сходствами[6] (за исключением того, которое в каждом из «блоков» $<O^+_{i1},O^+_{i2},O^+_{i3},O^+_{i4}>$ сформировано лишь теми сходствами, которые сопоставлены ненулевым значениям соответствующих переменных из $\mathbf{\Phi_i} = (x_{i1} \vee x_{i2})$) может быть взаимно-однозначным образом сопоставлено одному из нулей функции $\mathbf{\Phi}$. И, как следствие, при этом число N_Ω минимальных по вхождению покрытий всего множества *примеров* $\Omega^+(\mathbf{\Phi})$ соответствующими ЗКП-сходствами по числу $N_{\mathbf{\Phi}}$ наборов значений переменных, обеспечивающих выполнение рассматриваемой функции $\mathbf{\Phi}$, определяется следующим образом:

$$N_{\mathbf{\Phi}} = 2^n - (N_\Omega - 1).$$

Каждое $Cov_j(\mathbf{\Phi})$ – минимальное покрытие только что представленного типа– породит некоторую характеристическую функцию ChF_j, которая будет принимать значение *истина* на новом прецеденте O^τ следующего вида:

$$O^\tau = \{a_{11}, a_{12}, a_{13}, a_{14}, \dots, a_{i1}, a_{i2}, a_{i3}, a_{i4}, \dots, a_{m1}, a_{m2}, a_{m3}, a_{m4}\}.$$

Это будет происходить ввиду вложимости в представляющее этот прецедент множество $\{a_{11}, a_{12}, a_{13}, a_{14}, \dots, a_{i1}, a_{i2}, a_{i3}, a_{i4}, \dots, a_{m1}, a_{m2}, a_{m3}, a_{m4}\}$ подходящих комбинаций сходств $\{a_{i1}\}$ и $\{a_{i3}\}$, $\{a_{i2}\}$ и $\{a_{i4}\}$, формируемых на соответствующих примерах из «блоков» вида $\{O^+_{i1}, O^+_{i2}, O^+_{i3}$ и $O^+_{i4}\}$.

Итак, задача ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ*** полиномиально быстро сводима к задаче ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХФ***. Таким образом, с учетом полиномиальной разрешимости задачи ***ИСТИННАЯ НА НОВОМ ПРЕЦЕДЕНТЕ ХФ*** (см. Утверждение 2) доказательство ее перечислительной полноты завершено.

[6] Т.е. удовлетворяющими условию запрета на контрпримеры.

ЗАКЛЮЧЕНИЕ

В завершение обратим внимание на некоторые особенности использованных процедурных конструкций в части их применимости при обработке различных типов данных, востребованных в тех или иных прикладных задачах. Так, теоретико-множественный случай (ситуация, когда анализируемая база фактов представлена примерами и контрпримерами, сформированными как непустые подмножества некоторого исходного «алфавита» – конечного множества «атомарных» элементов – *образующих*) позволяет надежно формализовать класс исследовательских ситуаций, где накапливаемые эмпирические данные аккумулируются в виде таблиц значений признаков. При этом признаки, принимающие значения из конечного множества наименований, могут быть прямо представлены соответствующими образующими, а параметры с числовыми значениями могут быть разнесены в определенные интервалы, каждый из которых далее будет «поименован» соответствующей образующей. Этот теоретико-множественный язык представления данных и знаний позволяет адекватным образом работать с данными большого числа значимых прикладных предметных областей (медицинской и технической диагностики, защиты данных и информационной безопасности, задачами финансового мониторинга и др.).

Однако с обработкой более сложных типов данных (например, символьные последовательности\цепочки, графы общего вида и др.) ситуация оказывается сложнее. Так, в частности, если для проверки репрезентативности анализируемой базы фактов (как и для воспроизведения на более сложных типах данных используемой при доказательстве Утверждения 3 схемы сведения задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ***) можно обойтись лишь операцией сходства $\otimes$, то для полиномиальной разрешимости утверждения о существовании характеристической функции, которая принимает значение *истина* на новом прецеденте $O\tau$, этого уже оказывается недостаточно: одной операции сходства мало для «отделения» «самостоятельных» компонентов\фрагментов в описаниях прецедентов. Необходима также операция *разности объектов* (аналогичная *разности множеств*), позволяющая выделять дополнение подобъекта до объекта\прецедента. Причем эта операция должна носить нетривиальный характер (см., например, случаи анализа множественных вхождений подграфа в граф при вычислении соответствующих дополнений и т.п.). Говоря неформально, необходим формализованный «инструмент» для учета конкретного в каждом прецеденте «*контекста*» вхождения подобъекта в объект. А это, по-видимому, требует перехода от *бинарного* к *тернарному* отношению причинности, например:

(*причина*) & («*контекст*» ее наблюдения) => целевой *эффект*.

Тем не менее, даже в случае уточнения представлений о причинности средствами бинарного отношения математическая техника характеристических функций при оценке (*прогнозировании*) свойств новых прецедентов (например, при *диагностике* нового

пациента с учетом ранее накопленного в текущей базе фактов БФ опыта анализа подобных случаев) – инструмент, позволяющий оперировать сразу множеством *всех* формируемых на конкретной БФ эмпирических зависимостей. И даже в ситуации, когда таких зависимостей может оказаться *экспоненциально много*, собственно диагностическое заключение (НОРМА или ПАТОЛОГИЯ) может быть получено *полиномиально* быстро. Ведь, как было показано в настоящей статье, задача о диагностике нового прецедента эффективно разрешима.

СПИСОК ЛИТЕРАТУРЫ

1. Забежайло М.И. О емкости семейств характеристических функций, обеспечивающих корректное решение задач диагностического типа // Тезисы докладов 19-й Всероссийской конференции «Математические методы распознавания образов» – ММРО-2019 (25-28 октября 2019 г., Москва). – Москва: Российская Академия Наук, 2019.– С. 305-306.
2. Забежайло М.И., Трунин Ю.Ю. К проблеме доказательности медицинского диагноза: интеллектуальный анализ данных о пациентах в выборках ограниченного размера // Научно-техническая информация. Сер 2. – 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. – № 6. – P. 322-328.
3. Грушо А.А., Забежайло М.И., Тимонина Е.Е. О каузальной репрезентативности обучающих выборок прецедентов в задачах диагностического типа // Информатика и ее применения. – 2020. – № 1 – С. 80-86.
4. Кон П.М. Универсальная алгебра. – М.: Мир, 1968. – 359 с.
5. Simon J. On the difference between one and many // Lecture Notes in Computer Science**.** –1977. – Vol. 52. – P. 480-491.
6. Valiant L.G. The complexity of enumeration and reliability problems// SIAM Journal on Computing. – 1979. – Vol. 8, №1. – P. 410-421.
7. Valiant L.G. The complexity of computing the permanent // Theoretical Computer Science**.** – 1979. – Vol. 8, Iss.2. – P. 189-201.
8. Грушо А.А., Забежайло М.И., Зацаринный А.А., Тимонина Е.Е. О некоторых возможностях управления ресурсами при организации проактивного противодействия компьютерным атакам // Информатика и ее применения. – 2018. – Т. 12, №1. – С. 62-70.

Материал поступил в редакцию 23.08.20.

Сведения об авторе

ЗАБЕЖАЙЛО Михаил Иванович – доктор физико-математических наук, заведующий отделом Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: m.zabezhailo@yandex.ru

УДК 004.765

А.А. Рязанова

Концепция цифровых платформ как подход к интеграции научно-информационных процессов*

Рассматриваются базовые понятия цифровых платформ и свойства, отличающие их от информационных систем, а также особенности и возможности интеграции информационных процессов (в первую очередь – научно-информационных) в рамках этих платформ. Показано, что все свойства цифровой платформы интегративны, поскольку обеспечивают связь субъектов внутри платформы и взаимодействие платформ между собой. Анализируется актуальный уровень развития и соответствия базовым свойствам цифровых платформ некоторых важнейших областей общественного производства.

Ключевые слова: *цифровая платформа, интеграция цифровых платформ, свойства цифровых платформ, потенциал развития, информационная система*

DOI: 10.36535/0548-0027-2020-12-2

ВВЕДЕНИЕ

Термин «платформа», определяемый в геологии как крупное, монолитное, устойчивое к внешним тектоническим воздействиям основание, в наше время достаточно часто используется в различных сферах в отношении какого-либо объекта, созданного с целью объединения двух или нескольких систем (например, транспортной системы и городской инфраструктуры), либо размещенного для максимально эффективного выполнения объектом его функций (подобно морской нефтяной платформе). Уже первоначальные значения понятия платформы предполагают наличие свойств интегративности (способности объединять две или множество частей в единое органично функционирующее целое) и развития (способности позитивно влиять на формирование элементов системы в результате синергии взаимодействия).

В начале XXI в. наблюдались такие процессы, как активный переход коммуникаций в информационную плоскость и появление понятий коммуникационной и информационной платформ, при этом информационная платформа (в узком смысле) представляет собой совокупность аппаратного обеспечения и операционной системы, что необходимо для выполнения прикладных программ. Функциональная направленность субъектов (программ) в рамках аппаратно-программного комплекса определяется, например, обеспечением передачи данных (для транспортной и коммуникационной платформ), исполнением программного кода (для процессора) [1].

Настоящая работа посвящена описанию возможностей интеграции и потенциала развития цифровых платформ в широком смысле – как совокупности технологий и инструментов реализации процессов и приложений, необходимых для обеспечения работы сложных систем, объединяющих участников разнонаправленных информационных процессов, управляющих субъектами информационной системы (ИС) с целью эффективного достижения целей их функциональной деятельности. Ввиду большой потенциальной значимости цифровых платформ во всех сферах общественного производства и для ускорения научно-технического развития общества такая постановка задачи объясняет необходимость формулирования основных свойств цифровых платформ.

ДВИЖЕНИЕ ОТ БАЗОВЫХ ПОНЯТИЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ К КОНЦЕПЦИИ ЦИФРОВЫХ ПЛАТФОРМ

Для описания свойств цифровых платформ обратимся к основным понятиям информационной или компьютерной системы.

Согласно международному стандарту ISO/IEC 2382:2015 [2] информационная система включает систему обработки информации, связанные с ней ресурсы (человеческие и технические) и подсистемы (состоящие, например, из персональных ЭВМ, периферийных устройств и программного обеспечения для обработки данных). В учебном пособии [3] обозначено, что компьютерная система является подсистемой информационной системы.

В рамках теоретической субъектно-объектной модели компьютерных систем как информационная, так и компьютерная системы являются совокупностью (с точки зрения системного анализа – системной целостностью, состоящей из частей или компонентов) активных сущностей – субъектов и объектов, с которыми оперируют субъекты [4]. Фиксированная

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-37-90042.

декомпозиция стабильной (функционирующей) системы на компоненты определяется тем, что в текущий момент времени они не могут быть удалены или обновлены.

При этом мы не рассматриваем самоорганизующиеся и саморазвивающиеся системы, так как все изменения происходят под управлением пользователя (конструктора, эксплуатирующей организации и др.) и информация доводится до него только через субъекты (что является их важнейшим свойством), а процесс передачи информации с момента сообщения пользователем запроса компьютерной системе включает элементы разного уровня – от субъектов, обслуживающих периферийные устройства (инструменты управления), до программных модулей операционной системы.

Таким образом, субъекты порождают поток данных от одного объекта к другому и оказывают влияние друг на друга и на объекты. При этом в соответствии с принципом транзитивности в рамках платформы обеспечивается непрерывность потока между объектами на всех участках составного потока.

Понимание сущности информационного процесса как определенной последовательности порождения субъектов необходимо для реализации механизмов интеграции в рамках цифровых платформ. Передача информации через объекты при помощи потоков и взаимодействие субъектов через порождаемые ими объекты – это неотъемлемая составляющая интеграции информационных процессов, которая, однако, становится возможной при реализации интегрируемых объектов и процессов в рамках одной платформы. В этом и заключается значение цифровых платформ для интеграции процессов.

В отличие от информационных систем цифровая платформа (здесь и далее до описания свойств – в узком смысле, как аппаратно-программный комплекс) уже в соответствии с базовыми понятиями имеет четкую функциональную направленность и целевое назначение, от которых зависит тип аппаратного и программного обеспечения. Этот факт приводит к проблеме несовместимости платформ при реализации крупных проектов, вследствие чего приоритетным стало требование к их универсальности. Для обеспечения универсальности платформ используются соответствующие программные и аппаратные решения (специальные платы для переноса процессора и оперативной памяти на целевую платформу). К программным решениям относятся прежде всего программы-эмуляторы, воспроизводящие в операционной системе целевой платформы программы, разработанные для других операционных систем, и эмуляторы операционных систем, воспроизводящие на целевой платформе саму операционную систему, не совместимую с аппаратным обеспечением целевой платформы.

Критерии выбора платформы:

1) высокая производительность;

2) отказоустойчивость, обеспечиваемая возможностями автоматической реконфигурации при возникновении сбоя в работе;

3) надежность как свойство, прежде всего, аппаратной части платформы сохранять целостность данных;

4) масштабируемость, позволяющая увеличить производительность всей системы при замене одного компонента (например, процессора).

В настоящее время наблюдается переход от однородных сетей программно-совместимых компьютеров к неоднородным распределенным сетям, объединяющим компьютеры разных производителей, вследствие чего сети стали собой представлять распределенные многозадачные мультиресурсные системы. К подобным системам выдвигаются дополнительные требования, а именно – возможности:

- использования нового аппаратного и программного обеспечения в соответствии с новыми требованиями и решаемыми задачами на прежней аппаратной платформе;
- поддержки мобильности программного обеспечения (работы программных средств на разных платформах);
- применения одних и тех же интерфейсов на всех компонентах.

Описанные тенденции и эмпирический опыт свидетельствуют о том, что свойства универсальности, отказоустойчивости, надежности, а также масштабируемость (применительно к платформам в узком смысле) и высокая производительность могут расцениваться как необходимые условия корректного функционирования информационной системы.

ОСНОВНЫЕ СВОЙСТВА ЦИФРОВЫХ ПЛАТФОРМ

Для того чтобы определить основные закономерности процессов интеграции систем, необходимо сформулировать требования к информационной системе, позволяющие классифицировать ее как платформу в широком смысле слова.

С учетом основных свойств [5], информационная система становится цифровой платформой, когда имеет следующие свойства:

1) масштабируемость – свойство (или способность) информационной системы обрабатывать растущий объем задач, добавляя дополнительные ресурсы (увеличивать вычислительные возможности или включать функциональные элементы, в том числе нового поколения, выполняющие сходные задачи) [6]. Система является масштабируемой, если ее производительность с включением новых участников или увеличением объема обрабатываемых данных не падает, например, если она корректно функционирует при увеличении числа пользователей и количества запросов и индексируемых тем;

2) тиражируемость – система является тиражируемой, если ее можно адаптировать и внедрить в других условиях, например на другом предприятии без изменения ее структуры и состава субъектов. При этом должны сохраняться ее общие или типовые свойства, что создает условия для дальнейшей интеграции систем;

3) расширяемость – свойство, связанное с дополнением к информационной системе субъектов, реализующих новые функции;

4) развитие – свойство системы по меньшей мере сохранять свои качества и при возможности приобретать новые (наращивать потенциал) на всех этапах жизненного цикла, например, при переходе от количественных характеристик к качественным в системе

обработки Больших Данных. Одно из необходимых условий развития системы – это включенность в неё средств разработки информационного и программного обеспечения;

5) замкнутость в настоящий момент – фиксированное количество субъектов в конкретный момент времени;

6) целостность – свойство, когда система решает задачи, которые не могут быть решены отдельными ее компонентами при сохранении внутренней логики и структуры;

7) безопасность – свойство, связанное с целостностью и замкнутостью системы, к которому дополнены свойства конфиденциальности и доступности;

8) возможность связи платформ между собой, в первую очередь, за счет единых или стандартизированных интерфейсов.

Все эти свойства должны быть отнесены и к субъектам, реализованным в рамках платформы. В противном случае становится невозможной интеграция как информационных процессов в рамках платформы, так и платформ между собой.

Таким образом, все свойства цифровой платформы в основе интегративны, поскольку они обеспечивают связь субъектов внутри платформы и взаимодействие платформ между собой, однако интеграция достигается и полнотой процессов для достижения цели платформы. Следовательно, если информационная система не обладает указанными выше свойствами, реализованными на уровне всех её компонентов, то она не является платформой.

Далее мы рассмотрим предпосылки формирования цифровых платформ и их соответствие основным требованиям в наиболее важных областях общественного производства.

КОММУНИКАЦИОННЫЕ ЦИФРОВЫЕ ПЛАТФОРМЫ

В связи с ощутимыми преимуществами новых видов коммуникации в реальном времени, таких как видео-связь, мессенджеры и соцсети, голосовая и электронная почта, стали появляться первые платформы с основной функцией обеспечения коммуникаций [7].

На сегодняшний день создано множество коммуникационных платформ, самые известные из которых – dropbox, google drive, share point, однако ни одна из них не обладает полным набором функций, необходимых для эффективного выполнения всех задач организации в области коммуникаций. В связи с этим возникает острая необходимость в установке специального программного обеспечения, либо использовании платформ, интегрируемых в существующие платформы и ПО.

На существующих платформах функции связи в реальном времени уже заданы, например, в установленном приложении клиент может позвонить в банк, но не может установить видеосвязь с менеджером.

Гибкая и масштабируемая коммуникационная платформа как услуга (platform as a service) не только объединяет несколько продуктов, фактически освобождая пользователя (индивидуального, корпоративного) от высоких расходов на услуги телефонии и Интернета, приобретение программного и аппаратного обеспечения и его дорогостоящее обслуживание, но и позволяет интегрировать коммуникации в бизнес-процессы (и приложения) через интерфейсы прикладного программирования [8].

При этом предоставляется полноценное техническое сопровождение и техническая документация, а в некоторых случаях – комплекты для разработки программного обеспечения и библиотеки для интеграции приложений как на персональных компьютерах, так и на мобильных устройствах.

ЦИФРОВЫЕ ПЛАТФОРМЫ В ПРОИЗВОДСТВЕ

Современный этап четвертой промышленной революции означает для экономики, что умные устройства и роботы, организованные в единую информационно-коммуникационную систему нового поколения, самостоятельно управляют производственным процессом и логистикой, выполняя за человека самые тяжелые или рутинные операции. Это позволяет оптимизировать все стадии жизненного цикла продукта. За пределами «умных заводов» [9] идет углубление экономических связей за счет включения в производственную или логистическую цепочку новых участников, необязательно принадлежащих одной отрасли. При этом возможности интеграции процессов в рамках цифровых платформ позволяют реализовать на практике новые способы производства и создавать новые бизнес-модели, в частности делать сектора производства (стадии) на предприятии пригодными для изготовления модифицированных версий продукта. Цифровые платформы в производстве являются катализатором процессов четвертой промышленной революции и механизмом объединения усилий производителей, ассоциаций, науки и техники, политики.

На базе интегрированных платформ успешно могут разрабатываться методы решения и конкретные предложения в области стандартизации, безопасности информационных систем в промышленности, правовых основ производства, бизнес-моделей и т.д. Например, в Германии насчитывается около 350 платформенных решений для предприятий и городской инфраструктуры.

Для промышленной платформы «индустрия 4.0» [10] характерны общие коммуникационные структуры – коммуникационная сеть и протоколы передачи данных, действуют единые стандарты информационной безопасности и защиты персональных данных, общий язык представления данных (общий алфавит, символы, словарь, заданы структуры предложений, семантические инструменты). Эти инструменты позволяют расширять спектр функций продукта, объединять участников экономических отношений, организовывать информационные потоки между ними, а также гибкое, эффективное и ресурсосберегающее управление производственными процессами.

Модель платформы «индустрия 4.0» охватывает три направления:

1) все стадии жизненного цикла продукта – от разработки прототипа продукта до его переработки и утилизации;

2) организация всех внутренних бизнес-процессов на предприятии, всех информационных процессов и доступа к информации. Интеграция как объединение

физических и цифровых функций и свойств продукта, а также в целом информационного пространства и любых объектов, относящихся к производству (деталей, запчастей, технической документации, договоров и т.д.). Цифровой образ объекта представляет собой интерфейс его связи с другими объектами и субъектами сети (предметами, устройствами, участниками). Это главное условие для функционирования производственной платформы;

3) в условиях традиционной бизнес-иерархии выполнение функций привязано к техническим средствам, а продукт не встроен в структуру бизнес-процессов. Современная платформа «индустрия 4.0» позволяет распределить в сети функции продукта, который становится частью этой сети, при этом сама сеть охватывает и другие предприятия. Участники могут коммуницировать друг с другом вне зависимости от бизнес-иерархии.

Из-за отсутствия указанных выше интегративных базовых свойств цифровых платформ долгое время было практически невозможно создание единой корпоративной информационной системы, функционирующей в интересах всех структурных подразделений предприятия и выполняющей все задачи, связанные с производственными процессами, поэтому отдельные группы задач по управлению производством, финансовой деятельностью предприятия и др. традиционно решались в рамках отдельных информационных систем. Напомним, что в промышленности в общем информационная система рассматривается как совокупность программ и данных, реализующих какую-то часть тактических и стратегических задач предприятия и его бизнес-процессов.

Цифровая платформа «индустрия 4.0», обладающая всеми описанными здесь свойствами, представляет собой действенный инструмент интеграции всех производственных и коммерческих процессов, которые могут быть каким-либо образом отражены в информационном поле как предприятия, так и конкретной отрасли, в то время как информационная система в промышленности таковым инструментом не является.

ЦИФРОВЫЕ ПЛАТФОРМЫ В ОБРАЗОВАНИИ

Появление цифровых платформ в образовании в значительной степени обусловлено потребностью сохранить высокое качество предоставляемых услуг при создании дополнительных возможностей удаленного доступа к образовательным программам.

Образовательные платформы, как правило, имеют в своей основе программное обеспечение как услугу (software as a service) и предназначены для взаимодействия с другими образовательными и социальными приложениями (социальная функция), обеспечивая соответствие образовательного контента потребностям обучающихся. В отдельных случаях они включают встроенную аналитику, основанную на объединении сведений об учащихся по курсам, учебным заведениям и за их пределами, и позволяют находить необходимый учебный и пользовательский контент.

Наиболее известными являются образовательные платформы массовых открытых онлайн-курсов, как Coursera и edX, которые позволяют выбирать и изучать предметы известных ВУЗов без отрыва от основной деятельности или в дополнение к главной образовательной программе. Среди отечественных образовательных платформ можно выделить библиотеку видео-уроков школьной программы Interneturok.ru, Национальный открытый университет Интуит, систему онлайн-обучения «Argus-M» и соответствующую базу ответов на тестовые вопросы, просветительский проект и базу видео-лекций «Лекториум».

Существующие образовательные платформы практически представляют собой ресурсы, на которых сотни и тысячи лекций собраны в тематические кластеры. Есть возможность, с помощью удобного поиска, находить подходящие видеоматериалы, а затем оценивать курс и его влияние на дальнейший карьерный рост. Однако в рамках этих образовательных проектов невозможно интерактивное взаимодействие преподавателя с обучающимися и обучающихся между собой, что является неотъемлемой частью наиболее эффективных образовательных методик.

Для обеспечения вовлеченности слушателей, непрерывности и эффективности учебного процесса образовательная платформа должна выполнять следующие функции [11]:

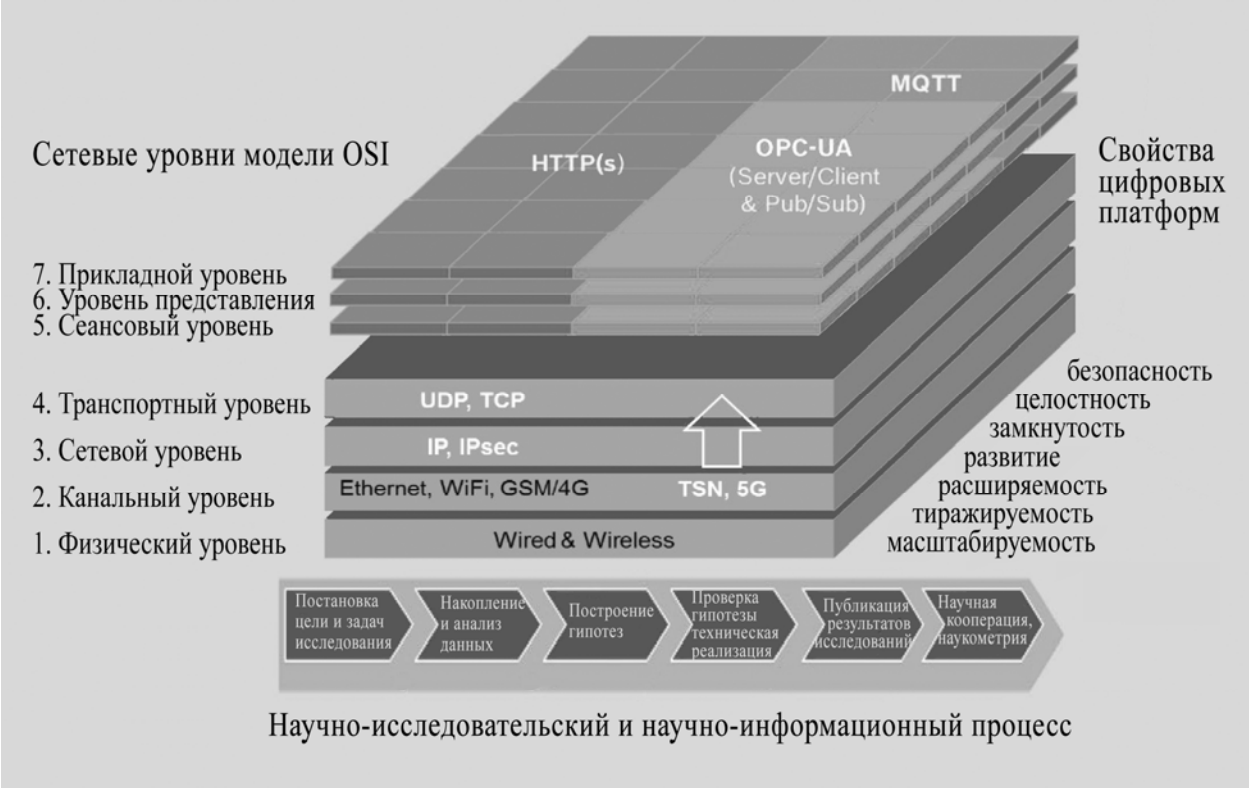
- проведение дистанционных лекций, конференций и презентаций с использованием дополнительных инструментов работы с материалом;
- надежное, интуитивно понятное, структурированное хранение учебных материалов (облачные решения) с возможностью передачи заданий и обратной связи;
- безопасный обмен сообщениями в специализированном мессенджере, разработанном с учетом защиты персональных данных.

Следующим неотъемлемым этапом развития образовательной платформы станет интеграция этих функций с помощью единого доступа.

ЦИФРОВЫЕ ПЛАТФОРМЫ В НАУКЕ

В современном научном дискурсе все чаще обсуждается проблема рассредоточенности данных, касающихся различных аспектов научной деятельности, прежде всего исследовательских и научно-технических проектов, наукометрических данных, в различных источниках (базах данных министерств, фондов, институтов), и отсутствия единой платформы обработки и обмена данными между участниками научной деятельности и заинтересованными в получении и применении научных результатов сторонами.

В связи с этим возникает необходимость создания платформенных решений для достижения эффективного внедрения результатов научных исследований в практическую деятельность. В частности, в рамках национального проекта «Наука» планируется создание цифровой платформы, которая в соответствии с Концепцией создания Единой цифровой платформы науки и высшего образования Минобрнауки России не только обеспечит для промышленности и наукоемкой экономики доступ к результатам интеллектуальной деятельности (в том числе результатам фундаментальных исследований), но и позволит сконструировать инструменты обработки больших данных с использованием технологий искусственного интеллекта и анализа больших данных [12].



Связь интеграции научно-исследовательских и научно-информационных процессов
с семиуровневой моделью Open System Interconnection (OSI)

Сегодня мы видим процесс интеграции как с помощью его реализации на общих технических и технологических принципах, например, с использованием распределенных реестров, так и с позиции обеспечения общих интерфейсов взаимодействия («наука-государство» «наука-образование», «наука-общество»). Поэтому целесообразно ввести и рассматривать понятие наукоцентричной платформы (НЦП), имея в виду, что центральной частью платформы для процессов, программных комплексов и обрабатываемых данных являются процессы и данные, относящиеся к обеспечению научных исследований и интерфейсов «наука-наука», «наука-промышленность», «наука-образование» и «наука-государство». Приведем некоторые направления интеграции научно-информационных процессов в рамках НЦП: по направлению «наука-наука» – виртуальные лаборатории и центры коллективного пользования, научная аналитика, научно-справочная деятельность, научные социальные сети; по направлению «наука-государство» – доведение госзадания до научных организаций, отчетность, наукометрия; в области взаимодействия науки и производителей материальных благ – доступ бизнеса к информации о результатах интеллектуальной деятельности, участие научных организаций в коммерческих проектах; в области взаимодействия науки и образования – коррекция образовательных программ [13].

Проиллюстрируем связь описанных базовых свойств цифровых платформ с семиуровневой моделью взаимодействия открытых систем в соответствии с этапами научно-исследовательского и научно-информационного процесса [14] (OSI – Open System Interconnection) в науке (рисунок).

Как следует из представленных нами процессов формирования платформ в основных сферах общественного производства, максимальный набор свойств, присущих цифровым платформам, проявляется в современных коммуникационных и промышленных платформах, в то время как научные и образовательные платформы, как правило, таковыми не являются, представляя собой только ресурсы хранения данных.

Интегративные свойства платформ позволяют существенно увеличивать охват эффективно решаемых задач, включая электронную обработку данных, автоматизацию функций управления (использование компьютера для комплексного решения отдельных функциональных задач), поддержку принятия решений (с использованием математических моделей и методов), экспертную поддержку и функционирование искусственных интеллект-помощников.

Дополнительно можно рассматривать следующие характеристики цифровых платформ:

- **сфера применения –** определяет, в какой сфере общественного производства используется цифровая платформа;
- **автономность** – определяет независимость платформы от других информационных систем;
- **первичность** – платформа не является «наследуемой» от существующих платформ;
- **базовые свойства объектов** – определяют назначение платформы;

Дополнительные характеристики отраслевых платформ

Характеристика (параметр)	Наукоцентричная	Образовательная	Производственная	Коммуникационная
Первичность	Да	Да	Да	Нет
Тип связности	Сложные связи (и линейные)	Линейные	Сложные связи (и линейные)	Линейные
Взаимодействие со смежными платформами	Частное	Частное	Частное	Универсальное
Управление оборудованием	Есть	Нет	Есть	Нет
Структура платформы	Иерархичная (изменяются уровни субъектов и объектов)	Линейная (уровни не меняются)	Иерархичная (изменяются уровни субъектов и объектов)	Линейная (уровни не меняются)

- **базовые свойства субъектов** – определяют свойства платформы;
- **взаимодействие со смежными платформами** – определяет «встроенность» платформы в общий информационный процесс;
- **взаимодействие с внешней средой** – отношение информационной технологии с объектами управления, предприятиями и системами, наукой, промышленностью программных и технических средств автоматизации;
- **структура платформы** включает функциональные компоненты (отвечают за процессы циркуляции и переработки информации) и их взаимосвязи, образующие внутреннюю организацию платформы и объединенные в опорную технологию и базу знаний;
- **предпосылки к интеграции** задают возможности взаимодействия с другими платформами;
- **модифицируемость** – возможность динамичного развития и изменения структуры.

Некоторые дополнительные характеристики отраслевых цифровых платформ приведены в таблице.

ЗАКЛЮЧЕНИЕ

Основным свойством цифровых платформ является потенциал их развития, следующий из назначения платформы – как основы для построения целостной системы информационных процессов (что и обеспечивает интеграцию) в области создания, хранения, передачи и обработки информации во всех сферах общественной жизни, производства и производственных отношений. В развитии информационных систем в сторону платформ проявляется второй закон диалектики – переход количественных изменений в качественные – наращивание возможностей, свойств, инструментов интеграции – позволяет в полной мере обеспечить процесс развития.

Свойства интегративности должны быть отнесены и к субъектам, реализованным в рамках платформы. Иначе становится невозможной интеграция как информационных процессов в рамках платформы, так и платформ между собой.

Таким образом, все свойства цифровой платформы в своей основе интегративны, поскольку они обеспечивают связь субъектов внутри платформы и взаимодействие платформ между собой, однако интеграция достигается и полнотой процессов для достижения цели платформы и ее свойств. Следовательно, если информационная система не обладает указанными нами свойствами, реализованными на уровне всех компонентов, то она не является платформой.

Интеграция информационных процессов с помощью цифровых платформ – это один из важнейших аспектов формирования доверенной и корректной цифровой среды во всех сферах общественного производства. Смыслом и целью интеграции являются консолидация как научного потенциала, так и уровня производства, создание среды конкурентоспособности всех отраслей в мировом контексте, снижение затрат на их инфраструктуру, установление объективного контроля уполномоченных государственных органов в области важнейших научных, производственных, социально-экономических процессов. В решении этих задач – высока значимость создания цифровых платформ, включая наукоцентричные.

Интеграция информационных процессов заключается в обеспечении универсального взаимодействия субъектов научной, производственной и хозяйственной деятельности (например, через единые интерфейсы передачи данных между субъектами) и обеспечении единого или взаимопреобразуемого формата объектов, используемых субъектами, объединенными в информационный процесс.

Интеграция цифровых платформ (включая наукоцентричные) в полной мере может быть достигнута в процессе развития субъектно-объектной модели компьютерной системы с точки зрения формулирования и доказательства достаточных условий интеграции информационных процессов и обеспечения их "здоровья", непрерывности и доверенности жизненного цикла корпоративных и общедоступных платформ, использующих современные информационные технологии.

СПИСОК ЛИТЕРАТУРЫ

1. Таненбаум Э., Остин Т. Архитектура компьютера. 6-е изд. – СПб: Питер, 2013. – 816 с. ISBN 978-5-496-00337-7
2. ISO/IEC 2382:2015(en) Information technology – Vocabulary. – URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en
3. Щербаков А.Ю. Современная компьютерная безопасность. Теоретические основы. Практические аспекты // Учебное пособие для студентов высших учебных заведений. – Сер. "Высшая школа". – Москва: Книжный мир, 2009. – 352 с.
4. Биктимиров М.Р., Щербаков А.Ю. Избранные главы компьютерной безопасности. – Казань: Изд-во казанского матем. общества, 2004. – 372 с.
5. Ракитов А.И., Бондяев Д.А., Романов И.Б., Егерев С.В., Щербаков А.Ю. Системный анализ и аналитические исследования: руководство для профессиональных аналитиков. – Москва: Типография «Возрождение», 2009. – 448 с.
6. Bondi André B. Characteristics of scalability and their impact on performance // WOSP '00: Proceedings of the second international workshop on Software and performance. – Ottawa, Ontario, Canada: Publisher Association for Computing Machinery, 2000. – P. 195-203. DOI:10.1145/350391.350432. ISBN 158113195X.
7. Why Should You Invest in a Communication Platform? – URL: https://www.cequens.com/story-hub/why-should-you-invest-in-a-communication-platform#:~:text=The%20Importance%20of%20Communication%20Platforms,services%2C%20platforms%20are%20access%20gateways
8. What is CPaaS? Communications Platform as a Service Explained. – URL: https://www.onsip.com/voip-resources/voip-fundamentals/what-is-cpaas-communications-platform-as-a-service-explained
9. Digitale Transformation in der Industrie. – URL: https://www.bmwi.de/Redaktion/DE/Dossier/Industrie-40.html
10. RAMI 4.0 – Ein Orientierungsrahmen für die Digitalisierung. – URL: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/rami40-einfuehrung-2018.html
11. Eine Plattform die verbindet. – URL: https://bildungsplattform.org/
12. Концепция создания Единой цифровой платформы науки и высшего образования Минобрнауки России. – URL: https://minobrnauki.gov.ru/common/upload/library/2019/07/20190705_Kontseptsiya_ETSP_1.4.9.pdf
13. Рязанова А.А. Виртуальные научные коммуникации как перспективный инструмент осуществления научной деятельности // Технические науки: научные приоритеты ученых. Выпуск 1. Сборник научных трудов по итогам международной научно-практической конференции (25 ноября 2016 г.). – С. 96-100.
14. Рязанова А.А. Значение технологии распределенных реестров для улучшения качества получения нового научного знания // Вестник современных цифровых технологий. – 2020. – № 2. – С. 21-29.

Материал поступил в редакцию 15.10.20.

Сведения об авторе

РЯЗАНОВА Алина Александровна – научный сотрудник, аспирант, заместитель начальника Центра развития криптовалют и цифровых финансовых активов по международной деятельности ВИНИТИ РАН, Москва
e-mail: a.ryazanova@c3da.org

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.89 : 81’374

Е.В. Котельников, Е.В. Разова, А.В. Котельникова, С.В. Вычегжанин

Современные словари оценочной лексики для анализа мнений на русском и английском языках* (аналитический обзор)

Рассматриваются способы создания словарей оценочной лексики на русском и английском языках с указанием их достоинств и недостатков. Анализируются 13 русскоязычных и 19 англоязычных словарей – приводятся их количественные характеристики и способы создания, вычисляются объединения и пересечения, определяется общая лексика, исследуется распределение по частям речи, указывается доля словосочетаний. Представлены современные области и методы применения словарей оценочной лексики.

Ключевые слова: *оценочная лексика, словари оценочной лексики, анализ тональности, анализ мнений*

DOI: 10.36535/0548-0027-2020-12-3

ВВЕДЕНИЕ

Анализ мнений (или *анализ тональности, sentiment analysis, opinion mining*) – это область компьютерной лингвистики, в которой исследуются мнения и оценки людей по отношению к различным объектам, таким как продукты, услуги, организации, персоны, события [1]. Под *тональностью* (или *полярностью*) понимают выраженную в тексте субъективность как позитивное или негативное отношение к некоторому объекту [2]. Тональность представляется в виде значения на определенной шкале, которая может быть бинарной (позитивное/негативное отношение), тернарной (добавляется нейтральное или противоречивое), *n*-арной или вещественной (например, [–1, 1]).

Анализ мнений в текстах весьма востребован в настоящее время: существует широкий диапазон приложений, например, учет общественного мнения при принятии решений в государственном управлении, организация обратной связи в образовательном процессе, прогнозирование результатов выборов, построение рекомендательных систем, планирование ценообразования и др. [3].

Существуют три основных подхода к анализу мнений в текстах – на основе машинного обучения, на основе словарей и гибридный [2, 4]. В подходе на основе машинного обучения требуются качественно размеченные обучающие данные и тратится значительное время на процедуру обучения; подход на основе словарей лишен указанных недостатков, но точность анализа часто оказывается недостаточно высокой; в гибридных системах комбинируются два рассмотренных выше подхода.

Ключевым элементом в последних двух подходах являются словари оценочной лексики. Точность анализа тональности в этом случае будет определяться качеством таких словарей. Существует много работ, посвященных созданию словарей оценочной лексики для анализа мнений в текстах [5–7], но проблеме исследования существующих словарей уделяется недостаточно внимания.

Цель настоящей статьи – представить аналитический обзор исследований в области создания, анализа и применения словарей русскоязычной и англоязычной оценочной лексики:

1) проанализировать все основные доступные на текущий момент словари оценочной лексики для двух языков – 13 русскоязычных и 19 англоязычных;

2) определить состав общей лексики для множества словарей – так называемое ядро оценочной лексики;

3) сделать выводы о сходстве и различии русскоязычной и англоязычной оценочной лексики;

4) предложить вариант описания автоматических методов создания словарей на основе пространства поиска;

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-17-50117

5) сформулировать рекомендации по созданию словарей.

В первом разделе статьи приводятся определение и классификация словарей оценочной лексики; во втором – рассматриваются способы создания таких словарей; в третьем – описываются существующие русскоязычные и англоязычные словари; четвертый посвящен их совместному анализу; в пятом – перечислены области и методы применения словарей для анализа тональности текстов. Заключение содержит основные выводы и рекомендации.

ОПРЕДЕЛЕНИЕ И КЛАССИФИКАЦИЯ СЛОВАРЕЙ ОЦЕНОЧНОЙ ЛЕКСИКИ

Словарь оценочной лексики (тональный словарь, sentiment lexicon, opinion lexicon) – это список оценочных слов и словосочетаний [8, с. 90]. Под *оценочными* понимаются слова и словосочетания, которые передают в тексте позитивное или негативное отношение к каким-либо объектам, например, *хороший*, *прекрасный*, *плохой*, *ужасный*. Понятие «словосочетание» в компьютерной лингвистике достаточно неопределенно [9, 10]. В настоящей работе под *словосочетанием* мы понимаем лингвистическую единицу, которая встречается, когда два и более слов используются совместно для выражения некоторого значения традиционным способом [11, с. 151]. Для краткости изложения, если специально не оговорено иное, под словами понимаем как отдельные слова, так и словосочетания.

Словари оценочной лексики можно классифицировать различными способами:

1) по составу – словари включают только отдельные слова или, также, словосочетания;

2) по шкале тональности – каждому элементу словаря приписывается только знак тональности (позитивная/негативная) или тональность представлена более детальной шкалой (например, действительные числа в диапазоне [–1, +1]);

3) по предметной области – словари могут быть универсальными или предметно-ориентированными;

4) по количеству языков – словари включают оценочную лексику для одного языка или являются многоязычными.

СПОСОБЫ СОЗДАНИЯ СЛОВАРЕЙ

Существуют три основных способа построения словарей оценочной лексики (рис. 1):

1) ручной – словарь создается, в основном, с помощью разметки слов усилиями людей (аннотаторов);

2) автоматический – словарь строится преимущественно на основе применения различных методов машинной разметки;

3) гибридный – существенную роль при создании словаря играют как ручные методы, так и автоматические.

Это относится, главным образом, к процессу разметки слов – например, при ручном способе создание списка слов-кандидатов, как правило, осуществляется автоматически, а при автоматическом – начальное множество слов для расширения часто формируется разработчиками.

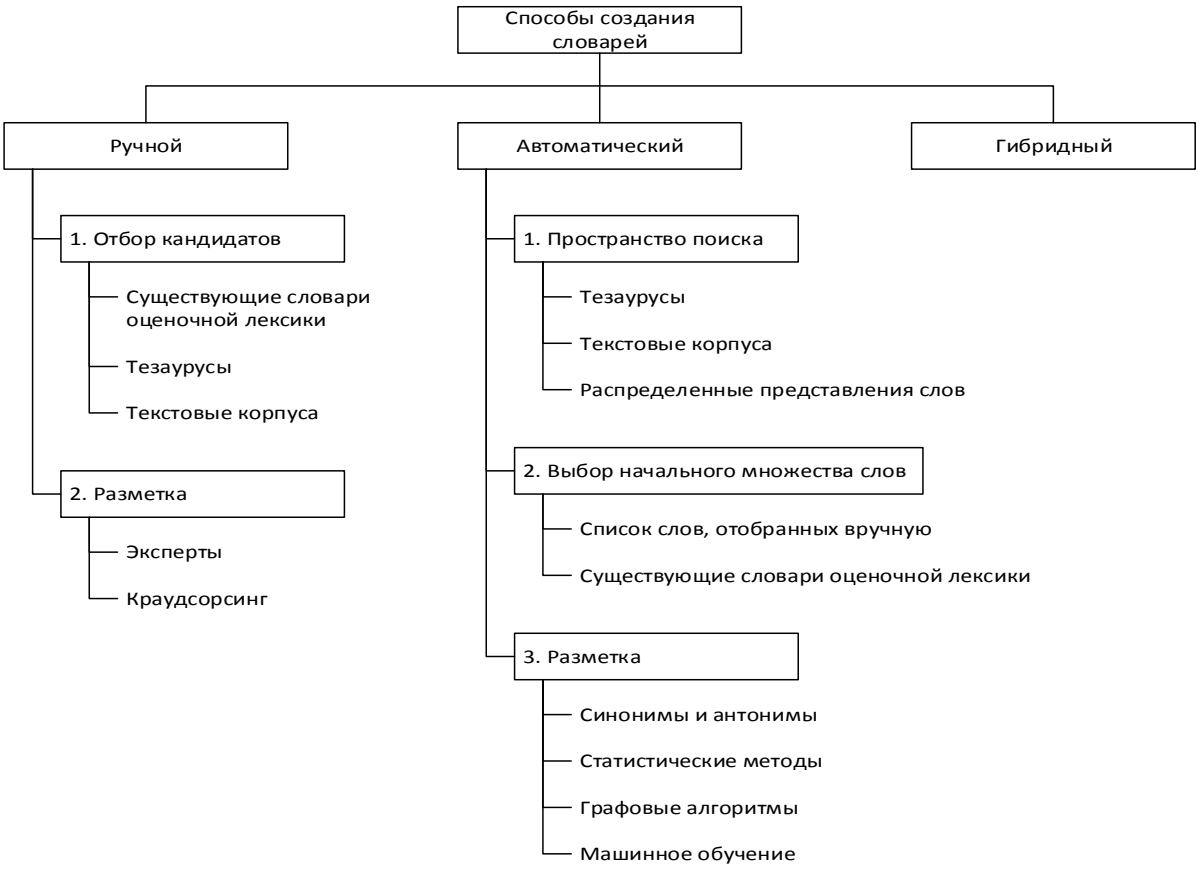


Рис. 1. Способы создания словарей оценочной лексики

Ручной способ

Построение словаря при *ручном способе*, как правило, выполняется в два этапа:

первый – создание списка слов-кандидатов на вхождение в словарь оценочной лексики;

второй – разметка слов-кандидатов в соответствии с выбранной шкалой тональности.

На первом этапе формируется список слов и словосочетаний, которые, вероятно, являются оценочными. Как правило, такой список строится автоматически, с использованием трех основных методов: 1) на основе существующих словарей оценочной лексики; 2) на основе тезаурусов; 3) с использованием текстовых корпусов.

В первом из указанных методов в качестве кандидатов используются слова из существующих словарей оценочной лексики. Такой метод применялся при создании словаря VADER [12]. Может быть использован перевод существующих словарей на требуемый язык [13, 14]. Во втором методе список слов-кандидатов строится на основе тезаурусов, таких как WordNet для английского языка [15] и РуТез для русского [16]. В третьем методе из текстовых корпусов слова-кандидаты извлекаются на основе различных синтаксических и статистических методов. Данный метод использовался при формировании словарей SemEval-2015 English Twitter Lexicon [17] и SentiRusColl [18].

Наиболее часто применяется комбинация двух или всех трех методов. Например, существующие словари оценочной лексики и тезаурусы использовались при создании словарей MPQA [19]; существующие словари и корпуса применялись для формирования словарей SCL-OPP [20], SCL-NMA [21] и SO-CAL [22]. Все три метода использовались при подготовке словарей EmoLex [23], RuSentiLex [24] и LinisCrowd [14].

Второй этап формирования словарей при ручном способе – разметка отобранных слов-кандидатов. Существуют два основных метода выполнения процедуры разметки: 1) экспертная разметка; 2) разметка с использованием краудсорсинга. В первом методе одному или нескольким экспертам в предметной области (например, лингвистам) предлагается назначить метки всем словам-кандидатам в соответствии с заданной шкалой тональности. Такой метод применялся для создания словарей MPQA [19], SO-CAL [22], RuSentiLex [24] и SentiRusColl [18]. Во втором методе разметка слов осуществляется на основе краудсорсинга (crowdsourcing) – задействования для решения задачи большого количества людей с использованием специальных интернет-платформ, таких как Amazon Mechanical Turk или CrowdFlower [25]. Данный метод использовался при формировании таких словарей как VADER [12], EmoLex [23], SemEval-2015 English Twitter Lexicon [17] и LinisCrowd [14].

Особым приемом в рамках краудсорсинга является геймификация процедуры разметки, которая позволяет повысить заинтересованность и вовлеченность аннотаторов [26, 27].

При ручном способе создания словарей важен контроль качества разметки. В общем случае качество построения словаря (при любом способе формирования) можно оценить на основе его применения для анализа тональности на заранее размеченных текстах в сравнении с существующими словарями [22, 28]. Также можно оценивать корреляцию разметки между независимыми группами аннотаторов [20]. При экспертной разметке, как правило, оценивается степень согласия между аннотаторами в соответствии с такими статистическими мерами как каппа Коэна (Cohen's κ), каппа Флейса (Fleiss's κ), пи Скотта (Scott's Π) [23]. В случае краудсорсинга контроль включает предварительную оценку знаний аннотаторами языка, оценку их разметки на небольшом контрольном наборе данных, анализ полноты разметки, определение выбросов в разметке [12, 23].

Автоматический способ

Процесс формирования словарей при автоматическом способе включает, как правило, три этапа (см. рис. 1):

первый – построение пространства поиска, в котором будет осуществляться разметка слов по тональности;

второй – выбор начального множества слов с известной тональностью;

третий – разметка слов в пространстве поиска на основе выбранного начального множества слов (бутстреппинг).

На первом этапе формируется (или используется существующее) пространство поиска, содержащее слова, требующие разметки. В таком пространстве определена метрика расстояния между словами, например, если пространство является векторным, то метрикой может служить косинусное или евклидово расстояние. Если пространство представляет собой связный граф, то расстояние определяется числом ребер в кратчайшем пути между вершинами, а в случае взвешенного графа – суммой весов ребер.

Существуют три основных метода при построении нового или использовании существующего пространства поиска: 1) на основе тезаурусов; 2) на основе корпусов; 3) с использованием распределенных представлений.

В первом из указанных методов формируется семантический граф понятий на основе существующих тезаурусов, таких как WordNet для английского языка [29, 30], РуТез для русского [24] и Wiktionary для множества языков [31, 32]. При этом может быть использован машинный перевод (как правило, с английского) для языков с недостаточной обеспеченностью такими лингвистическими ресурсами [32].

Во втором методе неявное пространство поиска образуется на основе корпуса текстов, снабженных разметкой по тональности. Информация о тональности текстов позволяет использовать такие корпуса для разметки слов на основе статистических методов или машинного обучения [33].

В третьем методе создаются или используются существующие распределенные представления слов, такие как word2vec [34] и GloVe [35]. Распределенные представления слов (word embeddings) – модели, в которых слова представлены в виде вещественных векторов фиксированной, обычно небольшой (несколько сотен), размерности [36, 37]. Строятся такие векторы на основе машинного обучения с использованием статистики совместной встречаемости слов в

неразмеченных текстовых корпусах. Цель обучения заключается в построении таких векторов, близость которых в пространстве распределенных представлений соответствует их семантической близости. В последнее время широкое распространение получили контекстные распределенные представления, такие как ELMo и BERT, в которых вектор слова зависит от текущего контекста [38]. Распределенные представления слов позволяют применять машинное обучение или просто функции расстояния для построения словарей оценочной лексики [5, 39].

На втором этапе выбирается начальное множество оценочных слов с известной тональностью. В качестве такого множества используется либо один из существующих словарей оценочной лексики [32, 40], либо небольшой список слов, отобранных вручную [29, 30, 33].

На третьем этапе осуществляется разметка слов в пространстве поиска с использованием начального множества оценочных слов. Для этого применяются следующие методы:

- расширение начального множества за счет синонимов, антонимов и других семантически связанных слов. Такой метод применим в случае использования пространства поиска на основе тезаурусов [40, 41];
- статистические методы, например, поточечная взаимная информация (pointwise mutual information, PMI), когда оценивается степень совместной встречаемости анализируемого слова и слов из начального множества [33];
- графовые алгоритмы распространения разметки, такие как graph propagation [42], label propagation [43] и random walk [29, 44]. Такие алгоритмы использовались при построении словарей в работах [5, 32];
- машинное обучение – на основе начального множества обучается классификатор, который затем применяется для определения тональности слов в пространстве поиска [29, 30, 45].

Гибридный способ

В этом способе совместно используются ручные и автоматические методы разметки. Например, при создании словаря Stanford Sentiment Treebank на первом этапе осуществлялась разметка фраз при помощи краудсорсинга, а на втором этапе полученная разметка использовалась для автоматической разметки слов на основе обучения рекурсивных нейронных сетей и с применением распределенных представлений слов [46]. Для формирования словаря ProductSentiRus сначала два аннотатора разметили оценочные слова в предметной области отзывов о фильмах, затем был обучен классификатор, который применялся для автоматической разметки слов в предметных областях отзывов о книгах, играх, фотокамерах и телефонах.

Достоинства и недостатки способов создания словарей

При использовании *ручного способа* можно выделить следующие достоинства:

- высокая точность – слова, вошедшие в словарь, с высокой вероятностью соответствуют указанной тональности [45];
- при качественном формировании словаря и мощном методе классификации, учитывающем лингвистические особенности, ручной способ может обеспечивать высокое качество анализа тональности [22].

Недостатки ручного способа:

- высокая трудоемкость работы по разметке [8];
- зависимость результатов от квалификации, мотивации и качества работы аннотаторов [23];
- низкая полнота [47], в частности, недостаточное количество слов, используемых в социальных медиа [12].

Автоматический способ обладает следующими достоинствами:

- высокая полнота (покрытие) – в рамках этого способа можно обеспечить попадание в словарь большого количества оценочных слов языка [45];
- низкая трудоемкость – минимальная ручная обработка [8].

Недостатки автоматического способа:

- низкая точность по сравнению с ручным способом – высока вероятность попадания в словарь слов, не являющихся оценочными;
- сложность создания универсального словаря в случае использования методов на основе корпусов;
- сложность создания предметно-ориентированного словаря в случае использования методов на основе тезаурусов;
- зависимость качества словаря от качества аннотированного корпуса при использовании методов на основе корпусов.

В целом, при создании словарей оценочной лексики на основе разных способов существует противоречие между точностью и полнотой (покрытием) [45]: ручной способ обеспечивает высокую точность определения оценочной лексики, но низкую полноту, в то время как автоматический – высокую полноту при низкой точности.

СЛОВАРИ ОЦЕНОЧНОЙ ЛЕКСИКИ ДЛЯ РУССКОГО И АНГЛИЙСКОГО ЯЗЫКОВ

В настоящее время известно множество словарей оценочной лексики, б*о*льшая их часть разработана для английского языка, но имеет переводные версии (в том числе, на русский). В настоящем разделе рассматриваются 18 англоязычных словарей, 8 русскоязычных и 3 словаря, имеющих версии как для английского, так и для русского языка.

Англоязычные словари

1. General Inquirer (URL: http://www.wjh.harvard.edu/ ~inquirer) – один из первых словарей оценочной лексики, созданный в Гарвардском университете в 60-х гг. XX в. [48], содержит более десяти тысяч слов, размеченных вручную в соответствии со множеством синтаксических и семантических категорий, включая тональность. Свободного доступа к словарю нет.

2. LIWC (Linguistic Inquiry and Word Counts – URL: http://liwc.wpengine.com) является в настоящее время частью одноименной коммерческой системы анализа текстов [49]. Разметка слов в LIWC сильно

коррелирована с General Inquirer. Свободного доступа к словарю нет.

3. ANEW (Affective Norms for English Words) создан в 1999 г. [50] и содержит разметку по категориям тональности (affective valence), активности и контроля (dominance). Слова размечены вручную студентами-психологами.

4. Словарь Бинга Лью (Bing Liu's Opinion Lexicon или Hu&Liu's Lexicon – URL: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html) – результат многолетней работы, начавшейся еще в 2004 г. [41]. Его исходная версия была создана на основе расширения начального списка из 30 прилагательных синонимами и антонимами из тезауруса WordNet. В дальнейшем словарь расширялся, в том числе за счет анализа текстов социальных медиа, поэтому присутствуют слова с ошибками.

5. MPQA (Multi-Perspective Question Answering – Subjectivity Lexicon – URL: http://mpqa.cs.pitt.edu/ lexicons/subj_lexicon) [19] является частью системы анализа мнений OpinionFinder (URL: http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2). Каждое слово в словаре MPQA имеет указание тональности (позитивная, негативная или нейтральная), а также степень тональности (сильная или слабая). При построении MPQA существующий словарь оценочной лексики [51] был расширен за счет тезауруса и словаря General Inquirer, а затем доразмечен вручную.

6. SO-CAL (Semantic Orientation CALculator – URL: https://github.com/sfu-discourse-lab/SO-CAL) – это программный инструмент, определяющий тональность текстов [22]. Словарь, используемый в этом инструменте, в рамках настоящей статьи также обозначается SO-CAL. Он был получен путем экспертной разметки словкандидатов, собранных из корпусов отзывов, а также из словаря General Inquirer.

7. SentiWordNet (URL: https://github.com/aesuli/sentiwordnet) [29] создан в рамках автоматического способа на основе разметки понятий тезауруса WordNet с использованием машинного обучения и алгоритма случайного блуждания (random walk).

8. AFINN (URL: https://www2.imm.dtu.dk/pubdb/pubs/6010-full.html) – словарь (назван по имени разработчика) создавался автором с 2009 г. [52]. Словарь был дополнен нецензурными и сленговыми выражениями с целью получения лучшего результата при автоматическом анализе сообщений в социальных медиа. В настоящей статье используется версия AFINN-111.

9. Sentiment140-Lexicon (URL: https://github.com/felipebravom/StaticTwitterSent/tree/master/extra/Sentiment140-Lexicon-v0.1) [33] создан на основе одноименного корпуса, включающего 1,6 млн твитов с позитивными и негативными хештегами. Слова размечались с использованием метода поточечной взаимной информации (PMI).

10. Stanford Sentiment Treebank (URL: https://nlp.stanford.edu/sentiment) – это словарь оценочной лексики, сформированный на основе одноименного корпуса, содержащего предложения, извлеченные из отзывов о фильмах. Для каждого предложения построено частичное дерево синтаксического разбора (partial parse trees) [46]. Предложения были разбиты на фразы, которые размечались по тональности при помощи краудсорсинга. Слова размечались на основе рекурсивных нейронных сетей с использованием деревьев синтаксического разбора и распределенных представлений слов.

11. ML-SentiCon (URL: https://github.com/mauropelucchi/catalan-referendum/tree/master/ML-SentiCon) – это автоматически созданные многоязычные (английский, испанский, каталонский, баскский и галисийский) многоуровневые оценочные словари [30]. Каждый словарь содержит 8 уровней, причем каждый вышележащий уровень включает все предыдущие, а также новые элементы. Многоуровневость словарей позволяет выбирать между количеством доступных слов и точностью оценок. Словари строились на основе машинного обучения и алгоритма PolarityRank с использованием WordNet.

12. VADER (Valence Aware Dictionary and sEntiment Reasoner – URL: https://github.com/cjhutto/vader Sentiment) – это название словаря и инструмента анализа мнений для социальных медиа на основе правил [12]. Исходный список оценочных слов из существующих словарей оценочной лексики (General Inquirer, LIWC и ANEW) был расширен смайликами, связанными с настроением, акронимами и часто используемым оценочным сленгом. Для разметки словкандидатов применялся краудсорсинг.

13. SCL-NMA (Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs – URL: http://saifmohammad.com/WebPages/SCL.html#NMA) [21] – представляет собой список слов и словосочетаний, включающих отрицания, модальные слова и наречия меры и степени. При создании словаря сначала были отобраны слова-кандидаты из General Inquirer, а также высокочастотные фразы из Британского национального корпуса, включающие слова из General Inquirer в комбинации с отрицаниями, модальными словами и наречиями меры и степени, которые затем были размечены при помощи краудсорсинга.

14. SCL-OPP (Sentiment Composition Lexicon for Opposing Polarity Phrases – URL: http://saifmohammad.com/WebPages/SCL.html#OPP) представляет собой список отдельных слов и фраз, включающих, по крайней мере, по одному позитивному и негативному слову, например, *счастливый инцидент* [20]. Слова-кандидаты отбирались из корпуса твитов (поэтому словарь содержит хештеги и слова с ошибками) с использованием словарей Бинга Лью, NRC Emotion lexicon (EmoLex), MPQA, ETSL и размечались с помощью краудсорсинга.

15. ETSL (SemEval-2015 English Twitter Sentiment Lexicon – URL: https://saifmohammad.com/WebPages/SCL.html#ETSL) – это список униграмм и биграмм с отрицанием [17], который использовался в качестве тестового множества на семинаре SemEval-2015 (задача 10, подзадача E) [53]. В качестве словкандидатов были отобраны высокочастотные термины (в том числе с ошибками в написании) из словаря хештегов и словаря Sentiment140-Lexicon, а разметка осуществлялась на основе краудсорсинга.

16. SocialSent (URL: https://nlp.stanford.edu/projects/socialsent) – это программный код и наборы данных, в том числе словари оценочной лексики, для проведе-

ния анализа мнений по конкретным предметным областям [5]. Алгоритм создания словарей SentProp включает два этапа: сначала формируется лексический граф на основе распределенных представлений слов, построенных с использованием предметно-ориентированных корпусов. Затем слова в лексическом графе размечаются на основе алгоритма случайного блуждания (random walk) и начального небольшого множества слов, специфичных для предметной области.

В SocialSent содержатся исторические словари оценочной лексики на английском языке. Для каждого десятилетия с 1850 по 2000 гг. построены пара словарей – один включает высокочастотные слова, второй – высокочастотные прилагательные. В настоящей статье использовалась пара словарей за последнее десятилетие.

17. SentiWords (URL: https://hlt-nlp.fbk.eu/technologies/sentiwords) создан в процессе исследования, каким образом на основе словаря SentiWordNet можно получить априорную оценку тональности слова, т.е. оценку, которая не зависит от различных семантических значений данного слова [45]. С этой целью было использовано машинное обучение и различные признаки, выводимые из характеристик слов SentiWordNet.

18. WordStat (URL: https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries) создан путем объединения отрицательных и положительных слов из словарей Harvard-IV, Regressive Imagery Dictionary и LIWC [54]. Затем список был автоматически расширен синонимами и связанными словами, а также различными морфологическими формами входящих в него слов.

Русскоязычные словари

1. ProductSentiRus (URL: http://panchenko.me/data/snalp/sentiment/ProductSentiRus.txt) разработали Илья Четвёркин и Наталья Лукашевич для предметной области товаров (фильмы, книги, игры, цифровые фотокамеры, мобильные телефоны) [55]. Сначала было использовано машинное обучение для построения модели на основе множества вручную размеченных слов из отзывов о фильмах и набора статистических и лингвистических признаков слов. Затем построенная модель применялась для классификации оценочной лексики в других областях. В результате общий словарь оценочной лексики для всех пяти областей включает 5 000 слов. Слова в словаре отсортированы по мере убывания вероятности их оценочности, но не разделены на позитивные и негативные.

2. Словарь Блинова (URL: https://github.com/kotelnikov-ev/BlinovSentimentLexicon). Павел Блинов с соавторами [31] сформировали вручную список из 969 наиболее позитивных и 1 138 наиболее негативных слов из словаря ProductSentiRus, а затем автоматически расширили список синонимами и антонимами из русского Викисловаря (https://ru.wiktionary.org/wiki).

3. LinisCrowd (URL: http://www.linis-crowd.org). Олеся Кольцова с соавторами [14] создавали свой словарь при помощи краудсорсинга. Сначала они отобрали 7 546 слов на основе списка высокочастотных прилагательных, словаря ProductSentiRus, толкового словаря и перевода англоязычного словаря оценочной лексики SentiStrength [56]. Затем не менее трех аннотаторов каждому слову присваивали оценки от –2 до +2. В настоящем исследовании позитивными и негативными считаются такие слова, которые получили большинство оценок соответствующей тональности.

4. Словарь Котельникова (URL: https://github.com/kotelnikov-ev/KotelnikovSentimentLexicons). Евгений Котельников с соавторами [57] сначала автоматически отобрали по 10 000 слов-кандидатов для каждой из пяти предметных областей (отзывы о ресторанах, автомобилях, фильмах, книгах и камерах), четыре аннотатора оценили каждое слово как позитивное, негативное, нейтральное или противоречивое, затем было создано два объединенных по предметным областям словаря: в первый вошли слова, относительно тональности которых были согласны три аннотатора из четырех (*Kotelnikov_large*), во второй – слова, относительно тональности которых согласны все аннотаторы (*Kotelnikov_small*).

5. Словарь Тутубалиной (URL: https://www.ispras.ru/dcouncil/docs/diss/2016/tutubalina/dissertacija-tutubalina.pdf). Елена Тутубалина в своей диссертации [58] создала вручную словарь на основе строго позитивных и негативных отзывов пользователей об автомобилях (в отзывах учитывались только разделы преимуществ и недостатков). Словарь был расширен за счет добавления синонимов.

6. RuSentiLex (URL: https://www.labinform.ru/pub/rusentilex). Наталья Лукашевич и Анатолий Левчик [24] создали словарь RuSentiLex, в котором для каждого слова указывается тональность (позитивная, негативная, нейтральная) и источник (мнение, факт, чувство). Сначала были сгенерированы списки оценочных слов на основе тезауруса РуТез, существующих словарей оценочной лексики, новостных статей и Twitter, затем лингвисты анализировали полученные списки для формирования итогового словаря. Словарь содержит более десяти тысяч слов и выражений, имеющих оценку тональности.

В нашей работе использовалась версия словаря 2017 г. и только слова и сочетания с позитивной или негативной тональностью. Исследовались две версии – *RuSentiLex_large*, включающая все позитивные и негативные элементы, и *RuSentiLex_small*, включающая позитивные и негативные элементы, для которых источником является мнение.

7. SentiRusColl (URL: https://github.com/kotelnikov-ev/SentiRusColl) [18] – это оценочный словарь словосочетаний. Для его создания использовался корпус отзывов по десяти предметным областям (книги, фильмы, музыка, автомобили, компьютеры, бытовая техника, телефоны, банки, отели, рестораны), из него автоматически отбирались словосочетания-кандидаты, которые далее размечались тремя аннотаторами. В словарь внесены словосочетания, получившие большинство голосов. Для нашего исследования использовался вариант словаря SentiRusColl со стоп-словами, содержащий предлоги и союзы.

8. Карта слов (https://kartaslov.ru) – это онлайн-карта слов и выражений русского языка [59]. Оценочный словарь, разработанный в рамках этого проекта, содержит слова русского языка, снабжённые

метками тональности (позитивная, негативная или нейтральная) и силы эмоционально-оценочного заряда. При создании словаря использовался краудсорсинг: в процессе разметки пользователю предлагалось оценить то или иное слово как нейтральное, позитивное, негативное или ответить «не знаю». В настоящем исследовании используется версия данного словаря от ноября 2019 г. (https://github.com/dkulagin/kartaslov/tree/master/ dataset/emo_dict), которая содержит только отдельные слова.

Многоязычные словари

1. Словарь Чена-Скиены (Chen-Skiena's lexicon – URL: https://sites.google.com/site/ datascienceslab/projects/multilingualsentiment). Яньцин Чен (Yanqing Chen) и Стивен Скиена (Steven S. Skiena) [32] автоматически построили словари оценочной лексики для 136 языков (включая русский и английский). На основе тезаурусов WordNet и Wiktionary и с помощью машинного перевода (Google Translate) был построен многоязычный семантический граф, связывающий слова на разных языках. Затем, отталкиваясь от англоязычного словаря Бинга Лью, были сформированы оценочные словари для других языков на основе алгоритма распространения разметки. Варианты для английского и русского языков далее называются *Chen-Skiena_en* и *Chen-Skiena_ru*.

2. EmoLex (NRC Emotion Lexicon – URL: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm) – словарь составлен Саифом Мохаммадом (Saif Mohammad) и Питером Тёрни (Peter Turney) с помощью краудсорсинга [23]. Для формирования множества слов-кандидатов использовались тезаурус Macquarie, а также словари General Inquirer и WordNet Affect. Словарь EmoLex содержит слова, соотнесенные с позитивной и негативной тональностью, а также с эмоциями «гнев», «предвкушение», «отвращение», «страх», «радость», «грусть», «удивление» и «доверие». Для исследования были отобраны слова из версии словаря NRC-Emotion-Lexicon-Wordlevel-v0.92, имеющие позитивную или негативную тональность, в том числе словосочетания. Вариант для английского языка далее называется *Emolex_en*.

В ноябре 2017 г. словарь был переведен на более чем 100 языков (в том числе, на русский) с помощью Google Translate. Для русского языка были отобраны слова и словосочетания, имеющие позитивную или негативную тональность. Вариант для русского языка далее называется *Emolex_ru*.

3. SenticNet (https://sentic.net/downloads) – это проект по анализу мнений на уровне понятий, начатый в 2009 г. в MIT Media Laboratory. Для построения словаря оценочной лексики сначала был сформирован граф понятий с использованием нейронных сетей долгой краткосрочной памяти (Long short-term memory, LSTM), обучавшихся на основе распределенных представлений слов [39]. Затем для разметки понятий по тональности применялось специальное векторное пространство AffectiveSpace [60]. В нашей статье используется версия словаря SenticNet 5 (далее в статье этот словарь называется *SenticNet_en*). В рамках SenticNet имеется проект BabelSenticNet [61], содержащий словари для 40 языков, в том числе русского. Далее в нашей статье словарь для русского языка обозначается *SenticNet_ru*.

Характеристики словарей

С учетом отсутствия открытого доступа к словарям General Inquirer и LIWC, а также двух версий словарей RuSentiLex и Котельникова, в нашей работе далее исследовались 19 словарей для английского языка и 13 словарей для русского языка. В табл. 1 и 2 приведены характеристики рассмотренных словарей после следующей предобработки: все элементы были преобразованы к нижнему регистру, в словарях были оставлены только элементы, состоящие из букв (латинского алфавита для английских словарей и кириллицы для русских), знака дефиса и пробела. В этих таблицах содержится информация о размерах множеств позитивных и негативных слов; размере объединений и пересечений этих множеств; способе, использованном для создания словаря; шкале тональности (указывается, каким образом шкала делилась для получения позитивных и негативных значений тональности); диапазоне количества слов в элементах словаря; годе создания или последней модификации.

Объем словарей для английского языка варьируется сильнее, чем для русского: от 765 (AFINN) до 523 092 (Sentiment140-Lexicon) слов. Для русскоязычных словарей объем меняется от 1 115 (Котельников_small) до 24 765 (SenticNet_ru) слов.

Для десяти словарей английского языка и семи словарей русского языка множества позитивных и негативных слов имеют непустое пересечение. Видимо, в словарях Блинова и Тутубалиной это произошло из-за автоматического расширения исходных списков слов за счет синонимов, а в словаре EmoLex_ru – вследствие автоматического перевода. В словаре Котельникова одно и то же слово может быть позитивным для одной области и негативным для другой, например, *непредсказуемый сюжет – непредсказуемые отказы*. В словаре RuSentiLex одинаковые слова могут иметь разный смысл и тональность, что указывается в ссылке на статью тезауруса РуТез, например, *легкий (покладистый) – легкий (поверхностный)*. В английских словарях пересечение связано с многозначностью слов или автоматическим режимом формирования словарей.

Для русскоязычных словарей среднее количество позитивных оценочных слов составляет 44%, негативных – 56%: негативная лексика более разнообразна. Для англоязычных словарей среднее количество позитивных оценочных слов составляет 58%, негативных – 42%: позитивная лексика более разнообразна. Однако если из рассмотрения исключить три самых больших англоязычных словаря, сформированных автоматически (SenticNet_en, Sentiment140-Lexicon и Sentiment Treebank), то соотношение позитивной и негативной лексики в английских словарях становится в точности таким же, как и в русских словарях (44% – позитивная и 56% – негативная).

Интересно, что русскоязычные словари, как правило, строятся на основе ручного способа (9 из 13), а для англоязычных словарей способы распределились поровну.

Русскоязычные словари предпочитают бинарную шкалу тональности (10 из 13), в то время как в англоязычных бинарная шкала используется только в 5 из 19 словарей.

Шесть англоязычных словарей и восемь русскоязычных содержат только отдельные слова, остальные словари включают словосочетания. Наиболее длинные словосочетания в английских словарях (до 38 слов) имеются в словаре Sentiment Treebank, в русских словарях (до 8 слов) – в словаре EmoLex_ru.

Из дальнейшего рассмотрения были исключены три самых больших англоязычных словаря, сформированных автоматически (SenticNet_en, Sentiment140-Lexicon и Sentiment Treebank), а также самый большой русскоязычный словарь SenticNet_ru, поскольку они содержат слишком много некорректных элементов вследствие их автоматического формирования. Например, в Sentiment140-Lexicon входят такие словосочетания, как *they landed*, *describe what*, а в SenticNet_ru – *амортизационная стойка*, *адвокатское сословие полотенца*. Кроме того, в дальнейшем при рассмотрении пересечений и объединений словарей мы не анализируем словарь ProductSentiRus, поскольку для него невозможно выполнить разделение на позитивные и негативные слова. Таким образом, далее рассматривается 16 англоязычных словарей и 11 русскоязычных.

Таблица 1

Характеристики англоязычных словарей оценочной лексики

№ п/п	Словарь	Поз. эл.	Нег. эл.	Объед.	Перес.	Способ	Шкала	Кол-во слов в элементах	Год	Ссылка
1	ANEW	419	346	765	0	Ручной	Непрерывная шкала [1; 9]: [1, 4] – нег., [6, 9] – поз.	1	1999	[50]
2	Словарь Бинга Лью	2 005	4 774	6 776	3	Автом.	Бинарная шкала {–1, +1}	1	2004	[41]
3	MPQA	2 304	4 152	6 450	6	Ручной	Бинарная шкала {–1, +1}	1	2005	[19]
4	SO-CAL	2 446	3 566	6 004	8	Ручной	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–3	2007	[22]
5	SentiWordNet	16 436	18 244	32 902	1778	Автом.	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–9	2010	[29]
6	EmoLex_en	2 312	3 324	5 555	81	Ручной	Бинарная шкала {–1, +1}	1	2011	[23]
7	AFINN	878	1 596	2 474	0	Ручной	Дискретная шкала [–5, 5]: [–5, –1] – нег., [1, 5] – поз.	1–3	2012	[52]
8	Sentiment140-Lexicon	328 188	194 904	523 092	0	Автом.	Непрерывная шкала [–6, 8]: [–6, 0) – нег., (0, 8] – поз.	1–2	2013	[33]
9	Sentiment Treebank	33 201	25 790	58 980	11	Гибрид.	Непрерывная шкала [0, 1]: [0; 0,4] – нег., (0,6; 1] – поз.	1–38	2013	[46]
10	ML-SentiCon	12 774	12 134	24 818	90	Автом.	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–8	2014	[30]
11	VADER	3 187	4 034	7 221	0	Ручной	Непрерывная шкала [–4, 4]: [–4, 0) – нег., (0, 4] – поз.	1–2	2014	[12]
12	Chen-Skiena_en	1 421	2 955	4 376	0	Автом.	Бинарная шкала {–1, +1}	1	2014	[32]
13	SCL-NMA	1 607	1 575	3 182	0	Ручной	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–4	2016	[21]
14	SCL-OPP	513	640	1 153	0	Ручной	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–3	2016	[20]
15	ETSL	654	496	1 150	0	Ручной	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–2	2016	[17]
16	SocialSent	1 255	1 213	2 463	5	Автом.	Непрерывная шкала [–3,9; 2,76]: [–3,9; –0.5] – нег., [0,5; 2,76] – поз.	1	2016	[5]
17	SenticNet_en	55 311	44 689	100 000	0	Автом.	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–5	2018	[39]
18	SentiWords	18 280	21 599	39 663	216	Автом.	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–9	2018	[45]
19	WordStat	5 492	10 486	15 955	23	Автом.	Бинарная шкала {–1, +1}	1–4	2018	[54]
	В среднем	**25 720**	**18 764**	**44 367**	**117**					

Таблица 2

Характеристики русскоязычных словарей оценочной лексики

№ п/п	Словарь	Поз. эл.	Нег. эл.	Объед.	Перес.	Способ	Шкала тональности	Кол-во слов в элементах	Год	Ссылка
1	ProductSentiRus			5 000	0	Гибрид.	Нет разделения	1	2012	[55]
2	Словарь Блинова	1 864	2 145	3 839	170	Автом.	Бинарная шкала {–1, +1}	1	2013	[31]
3	Chen-Skiena_ru	1 246	1 630	2 876	0	Автом.	Бинарная шкала {–1, +1}	1	2014	[32]
4	LinisCrowd	566	1 940	2 506	0	Ручной	Дискретная шкала [–2, 2]: [–2, –1] – нег., [1, 2] – поз.	1	2016	[14]
5	Котельников_large	1 046	2 211	3 247	10	Ручной	Бинарная шкала {–1, +1}	1	2016	[57]
6	Котельников_small	389	727	1 115	1	Ручной	Бинарная шкала {–1, +1}	1	2016	[57]
7	Словарь Тутубалиной	1 077	1 458	2 508	27	Ручной	Бинарная шкала {–1, +1}	1	2016	[58]
8	EmoLex_ru	2 085	2 805	4 750	140	Ручной	Бинарная шкала {–1, +1}	1–8	2017	[23]
9	RuSentiLex_large	3 433	9 485	12 784	134	Ручной	Бинарная шкала {–1, +1}	1–6	2017	[24]
10	RuSentiLex_small	2 686	5 719	8 331	74	Ручной	Бинарная шкала {–1, +1}	1–6	2017	[24]
11	SenticNet_ru	14 650	10 115	24 765	0	Автом.	Непрерывная шкала [–1, 1]: [–1, 0) – нег., (0, 1] – поз.	1–6	2018	[39]
12	SentiRusColl	4 008	2 569	6 577	0	Ручной	Бинарная шкала {–1, +1}	2–7	2019	[18]
13	Карта слов	4 550	6 774	11 324	0	Ручной	Бинарная шкала {–1, +1}	1	2019	[59]
	В среднем	**3 133**	**3 965**	**6 894**	**43**					

АНАЛИЗ СЛОВАРЕЙ

Объединения и пересечения словарей

Англоязычные словари. Объединение всех 16 английских словарей включает 64 597 элементов: 27 980 позитивных (43,3%)[1], 34 271 негативных (53,1%), а также 2 346 элементов, у которых не удалось однозначно определить тональность (3,6%). Пересечение всех словарей (обозначим его *En_Intersection16*) содержит 4 слова: 1 позитивное (*pretty*) и 3 негативных (*hell, hurt, sick*). В множество общих слов для всех словарей входит по одному прилагательному, существительному, глаголу и наречию. В это множество не попали слова *good* и *bad*[2].

Множество слов, которые встречаются хотя бы в 15 словарях из 16 – это объединение всех пересечений из 15 словарей (*En_Intersection15*) (табл. 3). Это множество включает 41 слово: 15 позитивных и 26 негативных (22 прилагательных, 11 существительных, 5 глаголов 2 наречия и 1 междометие).

En_Intersection14 включает уже 122 слова: 51 позитивное и 71 негативное (72 прилагательных, 37 существительных, 9 глаголов, 2 междометия и 2 наречия).

Постепенное уменьшение N при объединении всех пересечений из N словарей позволяет формировать так называемое *ядро оценочной лексики*, т.е. такое множество слов, относительно которых согласны все или почти все словари[3].

На рис. 2 в каждой ячейке приведено отношение мощности пересечения словаря, указанного в строке, и словаря, указанного в столбце, к объему словаря в строке (в процентах) в виде тепловой карты. Например, для пары словарей (ANEW, Словарь Бинга Лью) значение 49,2% указывает на то, что такая доля лексики словаря ANEW присутствует в словаре Бинга Лью.

Словарь SentiWords базировался на словаре SentiWordNet, поэтому SentiWordNet полностью входит в SentiWords, а в словаре Chen-Skiena_en оценочные слова были взяты из словаря Бинга Лью, поэтому Chen-Skiena_en полностью входит в данный словарь.

Для словарей SentiWords и SentiWordNet средняя доля вхождений в них других словарей является наивысшей (64,2% и 54,5% соответственно). Это объясняется тем, что данные словари имеют самый большой объем (39 663 и 32 902 слова). Минимальную долю вхождений показывают словари наименьшего размера ANEW, SCL-OPP и ETSL.

Словари SCL-NMA, SCL-OPP, ETSL и VADER имеют наименьшее вхождение в самый большой словарь SentiWords (от 37,0% до 45,9%), так как указанные словари ориентированы на анализ социальных медиа и включают соответствующую специфическую лексику.

В целом совпадение лексики между английскими словарями оказывается относительно невысоким (30,0%).

[1] Тональность слова определялась голосованием по большинству словарей, в которые оно входит.

[2] Слова *good* и *bad* отсутствуют в словаре SocialSent, слово *bad* отсутствует в ANEW. Кроме того, слово *bad* входит в словарь ML-SentiCon как позитивное.

[3] Другой подход к формированию ядра предложен в работе [62], в которой ядро определялось на основе анализа областей концентрации оценочной лексики в пространстве распределенных представлений слов.

Таблица 3

Пересечение 15 из 16 английских словарей (*En_Intersection15*)

Позитивные слова	Негативные слова
beautiful, beauty, cute, elegant, good, happy, love, lucky, nice, pretty, respect, safe, smile, sweet, wonderful	*abuse, badly, bastard, blind, bloody, bother, cancer, damn, danger, death, dirty, disaster, dreadful, hate, hell, hopeless, hurt, lonely, lost, mad, pain, sad, sick, stupid, ugly, wrong*

Таблица 4

Пересечение 10 из 11 русских словарей (*Ru_Intersection10*)

Позитивные слова	Негативные слова
благоприятный, великолепный, волшебный, достойный, замечательный, красивый, красота, легендарный, превосходный, преимущество, прекрасный, привлекательный, приятный, роскошный, удобный, ценный, чудесный, энергичный, эффективный, яркий	*бессмысленный, глупый, грязный, неприятный, неудачный, опасный, оскорбительный, трудный, тупой, ужасный*

Словари	ANEW	Словарь Бинга Лью	MPQA	SO-CAL	SentiWordNet	EmoLex_en	AFINN	ML-SentiCon	VADER	Chen-Skiena en	SCL-NMA	SCL-OPP	ETSL	SocialSent	SentiWords	WordStat
ANEW	100	49,2	51,5	48,9	66,1	59,5	36,5	47,2	46,3	47,8	31,2	14,5	15,4	38,2	74,0	60,8
Словарь Бинга Лью	5,5	100	79,8	50,5	58,1	35,1	19,4	44,5	32,7	64,6	15,3	3,7	4,3	12,2	70,6	73,1
MPQA	6,1	83,8	100	57,2	65,9	39,5	19,0	49,1	32,3	55,9	19,9	4,4	4,6	13,9	79,6	74,8
SO-CAL	6,2	57,0	61,5	100	63,8	38,0	17,6	47,6	29,5	44,2	19,5	3,9	4,6	14,8	76,2	64,9
SentiWordNet	1,5	12,0	12,9	11,6	100	10,6	4,1	37,9	7,9	8,7	3,8	1,2	1,4	4,1	100	20,0
EmoLex_en	8,2	42,8	45,8	41,0	62,6	100	17,7	44,6	26,2	39,6	17,6	5,4	4,7	14,6	73,9	61,6
AFINN	11,3	53,1	49,5	42,6	54,5	39,7	100	43,1	98,0	48,7	23,2	8,1	10,3	22,7	64,3	68,8
ML-SentiCon	1,5	12,1	12,8	11,5	50,2	10,0	4,3	100	8,4	8,7	3,5	1,0	1,1	3,6	63,2	20,3
VADER	4,9	30,6	28,9	24,5	35,9	20,1	33,6	29,0	100	24,9	10,7	3,5	4,5	10,2	43,7	43,4
Chen-Skiena_en	8,4	100	82,3	60,6	65,2	50,2	27,6	49,2	41,1	100	22,3	5,7	6,4	18,7	77,8	82,2
SCL-NMA	7,5	32,7	40,4	36,7	39,7	30,8	18,0	27,6	24,2	30,7	100	6,3	6,4	16,8	44,7	43,8
SCL-OPP	9,6	21,9	24,8	20,5	34,6	26,1	17,4	21,0	22,2	21,6	17,3	100	16,0	21,9	37,0	28,1
ETSL	10,3	25,6	25,7	23,8	41,5	22,9	22,2	24,6	28,4	24,4	17,7	16,1	100	24,6	45,9	33,1
SocialSent	11,9	33,6	36,4	36,0	54,9	33,0	22,8	35,8	29,8	33,3	21,6	10,2	11,5	100	62,0	43,3
SentiWords	1,4	12,1	13,0	11,5	83,0	10,3	4,0	39,5	8,0	8,6	3,6	1,1	1,3	3,9	100	20,4
WordStat	2,9	31,0	30,3	24,4	41,3	21,5	10,7	31,6	19,6	22,6	8,7	2,0	2,4	6,7	50,6	100

Рис. 2. Отношение попарных пересечений словарей к размеру словарей в строке (%)

Словари	Словарь Блинова	Chen-Skiena_ru	LinisCrowd	Котельников_large	Котельников_small	Словарь Тутубалиной	EmoLex_ru	RuSentiLex_large	RuSentiLex_small	SentiRusColl	Карта слов
Словарь Блинова	100	17,3	21,5	32,3	14,8	29,1	21,0	43,5	36,2	0,0	38,3
Chen-Skiena_ru	23,1	100	16,5	15,8	6,7	11,7	32,9	35,3	24,3	0,0	39,4
LinisCrowd	33,0	19,0	100	31,3	15,6	26,1	31,6	74,3	56,2	0,0	60,2
Котельников_large	38,2	14,0	24,2	100	34,3	23,4	19,8	48,5	39,9	0,0	46,8
Котельников_small	50,9	17,2	35,2	100	100	34,4	26,0	59,6	50,9	0,0	53,4
Словарь Тутубалиной	44,5	13,4	26,1	30,3	15,3	100	24,1	52,7	48,0	0,0	41,6
EmoLex_ru	17,0	19,9	16,7	13,5	6,1	12,7	100	34,9	24,8	0,0	35,5
RuSentiLex_large	13,1	7,9	14,6	12,3	5,2	10,3	13,0	100	65,2	0,1	36,5
RuSentiLex_small	16,7	8,4	16,9	15,5	6,8	14,5	14,2	100	100	0,1	38,6
SentiRusColl	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	100	0,0
Карта слов	13,0	10,0	13,3	13,4	5,3	9,2	14,9	41,2	28,4	0,0	100

Рис 3. Отношение попарных пересечений словарей к размеру словарей в строке (%)

Русскоязычные словари. Объединение всех 11 русских словарей включает 32 902 элемента: 13 240 позитивных (40,3%)[4], 19 323 негативных (58,7%), а также 339 элементов, у которых не удалось однозначно определить тональность (1,0%). Пересечение всех 11 словарей (обозначим его *Ru_Intersection11*) пусто**.** Множество слов, которые встречаются хотя бы в 10 словарях из 11, – это объединение всех пересечений из 10 словарей (*Ru_Intersection10*) (табл. 4). Это множество включает 30 слов: 20 позитивных (18 прилагательных и 2 существительных) и 10 негативных (10 прилагательных). Интересно, что в это множество не попали слова *хороший* и *плохой*[5]**.**

Ru_Intersection9, т. е. множество слов, которые встречаются хотя бы в 9 словарях из 11, – это объединение всех пересечений из 9 словарей, включает уже 134 слова: 71 позитивное (65 прилагательных и 6 существительных) и 63 негативных (51 прилагательное, 10 существительных, 1 наречие, 1 глагол). Следует отметить преобладание прилагательных для обоих языков, но в разной степени: например, в *En_Intersection15* прилагательных 53,7%, в то время как в *Ru_Intersection10* – 93,3%.

На рис. 3 в каждой ячейке приведено отношение мощности пересечения словаря, указанного в строке, и словаря, указанного в столбце, к размеру словаря в строке (в процентах) в виде тепловой карты.

Словари RuSentiLex_small и Котельников_small являются частью, соответственно, словарей RuSentiLex_large и Котельников_large.

Так же, как для английского языка, для самых крупных русскоязычных словарей RuSentiLex_large и Карта слов средняя доля вхождений в них других словарей является наивысшей (49,0% и 39,0% соответственно). Наименьшую долю вхождений показывает SentiRusColl, так как он содержит только словосочетания.

Словари Chen-Skiena_ru и EmoLex_ru имеют наименьшее вхождение в самый большой словарь RuSentiLex, что объясняется формированием этих словарей при помощи машинного перевода.

[4] Тональность слова определялась голосованием по большинству словарей, в которые оно входит.

[5] *Хороший* отсутствует в словаре EmoLex_ru (слово *good* было переведено как *хорошо*) и в словаре SentiRusColl (этот словарь содержит лишь словосочетания), *плохой* – в словаре Котельников_small (один из аннотаторов отнес *плохой* к нейтральным словам для нескольких областей), EmoLex_ru (слово *bad* было переведено как *плохо*) и в словаре SentiRusColl (этот словарь содержит лишь словосочетания). Кроме того, слово *плохой* входит в словарь Блинова как позитивное.

В целом, совпадения лексики между русскими словарями (без учета словарей-подмножеств RuSentiLex и словаря Котельникова) оказывается в среднем ниже (20,5%), чем между английскими словарями (30,0%).

Части речи

Англоязычные словари. Части речи для английских словарей были получены при помощи библиотеки Stanza от Stanford NLP Group [63].

В табл. 5 показано распределение частей речи в отдельных словарях и в объединенном словаре. Во втором столбце приведен размер словаря, в третьем и четвертом – количество элементов словаря, являющихся отдельными словами, и их доля в словаре, далее доля существительных, глаголов, прилагательных и наречий среди элементов словаря, состоящих из одного слова.

В 14 из 16 словарей наблюдается преобладание существительных, лишь в SO-CAL и Chen-Skiena_en количество прилагательных хоть и незначительно, но превышает количество существительных. В объединенном словаре 58,0% слов – это существительные. В большинстве словарей, в том числе и в объединенном словаре, количество прилагательных значительно больше количества глаголов, исключение составляют словари AFINN и ETSL. Интересно отметить довольно высокую долю наречий в MPQA, SO-CAL и словаре Бинга Лью.

По сравнению со словарями *En_Intersection16*, *En_Intersection15* и *En_Intersection14*, значительную часть которых составляют прилагательные, их доля значительно уменьшилась. Таким образом, ядро оценочной лексики составляют прилагательные, но по мере расширения словаря существительные и глаголы начинают преобладать.

Словосочетания составляют 23,0% (14 862) от объема объединенного словаря. Из них 81,9% являются биграммами, 15,4% – триграммами. Наиболее частотными комбинациями по сочетанию частей речи являются существительное + существительное (30,6%, например, *air alert*, *animal disease*, *food poisoning*) и прилагательное + существительное (27,4%, например, *cool stuff*, *right direction*, *accidental injury*). Доли других комбинаций не превышают 4%.

Русскоязычные словари. В табл. 6 приведено распределение частей речи в отдельных словарях и в объединенном словаре; также показано соотношение отдельных слов и словосочетаний. Части речи были получены при помощи морфологического парсера pymorphy2 [64]. Заметим, что в табл. 6 отсутствует информация о словаре SentiRusColl, поскольку он содержит элементы, включающие не менее двух слов.

Некоторые словари предпочитают одни части речи, например, в словаре Тутубалиной, словаре Блинова и в словаре Котельникова много прилагательных; в переводных словарях (EmoLex_ru и словаре Chen-Skiena_ru), а также в RuSentiLex высока доля существительных. В словаре LinisCrowd относительно сбалансировано соотношение существительных и прилагательных, а в словаре Карта слов – соотношение существительных и глаголов. Можно отметить высокую долю наречий в словаре Котельникова.

В объединенном словаре преобладают существительные (41,2%); их доля значительно ниже, чем в англоязычных словарях (58,0%). Интересно, что доля прилагательных почти одинакова – 22,8% в русском и 23,0% в английском словарях. Глаголов в русскоязычных словарях значительно больше, чем в англоязычных – 29,0% против 11,8%.

Таблица 5

Распределение частей речи (%)

Словари	Объем словаря	Отдельные слова		Части речи отдельных слов				
		Кол-во	%	Сущ.	Глаг.	Прил.	Нареч.	Другие
ANEW	765	765	100,0%	66,5%	10,1%	22,6%	0,5%	0,3%
Словарь Бинга Лью	6 776	6 776	100,0%	37,9%	15,9%	34,0%	11,7%	0,5%
MPQA	6 450	6 450	100,0%	38,0%	13,5%	34,7%	13,2%	0,6%
SO-CAL	6 004	5 920	98,6%	34,4%	11,6%	39,4%	14,0%	0,6%
SentiWordNet	32 902	25 178	76,5%	54,7%	9,1%	26,5%	9,1%	0,6%
EmoLex_en	5 555	5 555	100,0%	56,2%	14,9%	26,8%	1,7%	0,4%
AFINN	2 474	2 460	99,4%	35,1%	31,2%	30,7%	1,8%	1,2%
ML-SentiCon	24 818	18 874	76,0%	53,7%	8,9%	32,1%	4,7%	0,6%
VADER	7 221	7 217	99,9%	53,6%	16,6%	21,1%	7,4%	1,3%
Chen-Skiena_en	4 376	4 376	100,0%	37,6%	17,2%	38,5%	6,2%	0,5%
SCL-NMA	3 182	1 611	50,6%	43,8%	20,9%	33,3%	1,3%	0,7%
SCL-OPP	1 153	589	51,1%	46,9%	20,2%	24,6%	6,3%	2,0%
ETSL	1 150	902	78,4%	41,1%	24,1%	21,8%	3,7%	9,3%
SocialSent	2 463	2 463	100,0%	47,6%	13,1%	35,3%	2,5%	1,5%
SentiWords	39 663	30 784	77,6%	55,2%	8,5%	27,9%	8,0%	0,4%
WordStat	15 955	15 702	98,4%	51,2%	21,1%	21,4%	5,7%	0,6%
Объединенный словарь	64 597	49 735	77,0%	58,0%	11,8%	23,0%	6,4%	0,8%

Таблица 6

Распределение частей речи (%)

Словари	Объем словаря	Отдельные слова		Части речи отдельных слов				
		Кол-во	%	Сущ.	Глаг.	Прил.	Нареч.	Другие
Словарь Блинова	3 839	3 839	100,0%	18,5%	19,5%	47,0%	8,2%	6,8%
Chen-Skiena_ru	2 876	2 876	100,0%	48,8%	20,6%	19,6%	6,3%	4,7%
LinisCrowd	2 506	2 506	100,0%	38,2%	18,1%	42,8%	0,4%	0,5%
Котельников_ large	3 247	3 247	100,0%	26,4%	25,0%	35,0%	10,7%	2,9%
Котельников_ small	1 115	1 115	100,0%	26,4%	11,8%	44,8%	13,5%	3,5%
Словарь Тутубалиной	2 508	2 508	100,0%	9,3%	2,8%	78,5%	1,6%	7,8%
EmoLex_ru	4 750	4 486	94,4%	53,1%	13,0%	25,9%	1,7%	6,3%
RuSentiLex _large	12 784	10 543	82,5%	47,9%	24,1%	26,3%	0,4%	1,3%
RuSentiLex _small	8 331	7 151	85,8%	48,1%	18,3%	31,6%	0,5%	1,5%
Карта слов	11 324	11 324	100,0%	40,9%	39,7%	17,3%	1,7%	0,4%
Объединенный словарь	32 902	23 836	72,4%	41,2%	29,0%	22,8%	3,5%	3,5%

Так же, как в англоязычных словарях, по мере расширения ядра оценочной лексики существительные начинают преобладать над прилагательными.

Словосочетания составляют 27,6% (9 066) от размера объединенного словаря (близко к англоязычным словарям – 23,0%). Из них 56,2% являются биграммами, 30,5% – триграммами, 10,3% – квадрограммами (доля триграмм и квадрограмм значительно выше, чем в английском языке). Наиболее частотными комбинациями по сочетанию частей речи являются прилагательное + существительное (20,4%, например, *канительное дело*, *огромный респект*, *экономный расход*) и глагол + существительное (7,2%, например, *испытать судьбу*, *подкладывать свинью*, *пожалеть время*).

ПРИМЕНЕНИЕ СЛОВАРЕЙ ОЦЕНОЧНОЙ ЛЕКСИКИ

Области применения

Словари оценочной лексики применяются для анализа тональности текстов в различных предметных областях – в онлайн-торговле, в рекомендательных системах, в экономике, в политических исследованиях, в медицине и образовании.

В онлайн-торговле анализ тональности используется для изучения мнений потребителей с целью выявления преимуществ и недостатков предлагаемых на рынке товаров и услуг в интересах производителей и для помощи в выборе другим потребителям, для объяснения и предсказания продаж. Например, с помощью словарей оценочной лексики анализируются мнения относительно мобильных телефонов [65], фотокамер [66], электроники [67], книг [68], кухонных приборов [28, 67], услуг провайдеров облачных сервисов [69], услуг энергетической компании [70].

Анализ тональности применяется в рекомендательных системах для выработки рекомендаций, соответствующих интересам пользователей. Анализируются мнения пользователей о фильмах, банках, ресторанах, отелях [18, 28, 71].

В экономике, анализируя тональность текстовых сообщений с использованием словарей оценочной лексики, предсказывают направление движения стоимости акций на фондовом рынке [72]. В политических исследованиях анализ тональности используется для определения отношения избирателей к политическим партиям или кандидатам на выборах в президенты и законодательное собрание с целью предсказания результатов голосования [73]. В медицине с помощью словарей оценочной лексики анализируются мнения пациентов о врачах и лекарствах [74], о качестве обслуживания в медицинских учреждениях [75]. В сфере образования сообщения, полученные в качестве обратной связи от студентов, анализируются с целью оценки образовательных курсов и выступлений лекторов в учреждениях высшего образования [76].

Методы анализа мнений с использованием словарей

Как указывалось во Введении, существуют три основных подхода к анализу мнений в текстах – на основе машинного обучения, на основе словарей и гибридный. Словари оценочной лексики используются в последних двух подходах.

Методы на основе словарей. В работе [73] словарный подход используется для анализа тональности твитов, посвященных выборам в законодательное собрание в Индии. Применяются словари оценочной лексики AFINN, Бинга Лью, EmoLex, Syuzhet, SentiWordNet, SenticNet, VADER. В качестве признаков рассматриваются N-граммы и смайлики. В результате экспериментов наименьшее значение средней абсолютной ошибки было получено с использованием словаря VADER.

В статье [76] предлагается система анализа полученных в качестве обратной связи от студентов сообщений OMFeedback, оценивающая выступления лекторов в учреждениях высшего образования. Для анализа тональности сообщений используется метод на основе словаря VADER.

Способ формирования словаря для определенной предметной области, разработанный в [75], применяется для решения задачи классификации по тональности отзывов пациентов о качестве обслуживания в медицинских учреждениях. Сформированный словарь позволяет получить более высокое качество классификации, чем словари VADER и AFINN.

В работе [70] анализируются твиты потребителей услуг энергетической компании. Классификация текстов по тональности осуществляется с использованием инструмента Sentimentr и словаря Бинга Лью. Авторы отмечают, что каждый из инструментов наиболее хорошо выделяет тексты определенных классов тональности. В предлагаемом методе сначала используется Sentimentr для выделения негативных твитов, а затем оставшиеся твиты классифицируются с помощью словаря Бинга Лью на позитивные и нейтральные. Такой подход позволяет повысить качество классификации текстов, принадлежащих специфической предметной области, по сравнению с использованием каждого словаря по отдельности.

Авторы работы [74] применяют словарь испаноязычной оценочной лексики iSOL, полученный машинным переводом словаря Бинга Лью, для классификации по тональности мнений о лекарствах из корпуса DOS и мнений пациентов о врачах.

В работе [66] предложен метод вычисления весов оценочных слов на основе генетического алгоритма и совместной кластеризации слов и документов. Для оценки качества анализа тональности используются текстовые коллекции семинара РОМИП-2011, содержащие отзывы о фильмах, книгах и фотокамерах. По результатам экспериментов предложенный метод позволил получить более высокие значения F1-меры по сравнению с SVM и словарным методом.

В статье [18] разработан способ создания универсального словаря оценочной лексики SentiRusColl из словосочетаний. Сформированный словарь показывает в среднем более высокое качество классификации текстов по тональности применительно к десяти предметным областям по сравнению со словарем RuSentiLex.

В целом, на основе анализа предыдущих работ, сложно сделать вывод о преимуществе того или иного англоязычного словаря оценочной лексики: отсутствуют исследования, в которых сравнивались бы основные существующие словари на базе единого подхода для одних и тех же текстовых корпусов. Результаты, приводимые в статьях, сильно зависят от используемого метода анализа тональности и предметной области. Тем не менее, на основе обзора публикаций, проведенного в рамках настоящего исследования, можно сделать следующие предварительные выводы относительно англоязычных словарей:

- практически во всех статьях используется словарь SentiWordNet;
- SentiWordNet показывает результаты, сопоставимые со словарем MPQA;
- словари Бинга Лью и VADER демонстрируют преимущество по точности над словарями SentiWordNet и MPQA.

Повторим, что эти выводы не окончательные и требуют уточнения на базе масштабного сравнительного анализа.

Для русскоязычных словарей проводилось значительно меньше исследований. Только в 2018 г. был выполнен сравнительный анализ существующих словарей оценочной лексики на основе применения машинного обучения для разных предметных областей [28], который показал преимущество словаря ProductSentiRus. Однако в сравнении не участвовал словарь Карта слов, а используемый метод машинного обучения мог не выявить всех преимуществ словарей при анализе тональности.

Гибридные методы. Авторы статьи [65] предложили метод анализа тональности, основанный на использовании словаря SentiWordNet и машинном обучении – методе k ближайших соседей, наивном байесовском классификаторе и случайном лесе. Для тестирования метода используется корпус отзывов о мобильных телефонах с сайта Amazon. Наилучшее значение точности было получено с использованием случайного леса.

В работе [69] разработан гибридный метод, основанный на словарном подходе и методах нечеткой логики. Эксперименты проводятся с использованием корпуса, содержащего мнения пользователей об услугах поставщиков облачных сервисов.

В статье [77] предложен метод расширения словаря оценочной лексики для аспектно-ориентированного анализа тональности. Авторы сравнили два словаря, составленных вручную, и словарь, созданный при помощи предложенного метода. В качестве классификатора применялся метод максимальной энтропии. Эксперименты проводились с использованием корпуса отзывов о ресторанах с семинара SentiRuEval-2015 и продемонстрировали преимущество предложенного метода.

В статье [78] предлагается метод анализа тональности текстов TextJSM, основанный на ДСМ-методе интеллектуального анализа данных и использующий словарь оценочной лексики из работы [57]. Предложенный метод показывает преимущество над традиционными методами машинного обучения.

В целом, в работах, посвященных гибридным методам анализа тональности, такие методы показывают более высокие результаты, чем словарный подход или машинное обучение по отдельности.

ЗАКЛЮЧЕНИЕ

Проведенный анализ русскоязычных и англоязычных словарей оценочной лексики позволяет сделать следующие выводы.

1. Спектр англоязычных словарей шире, их размеры больше (даже после исключения автоматически построенных словарей с большим количеством ошибок объединенный англоязычный словарь в два раза превышает русскоязычный), в них чаще встре-

чаются словосочетания, а используемые шкалы тональности детальнее, чем у русскоязычных словарей.

2. Словари для обоих языков имеют ряд совпадающих характеристик: часто встречается ситуация, когда одно и то же слово рассматривается и как позитивное, и как негативное; негативная лексика более разнообразна, причем соотношение количества позитивных и негативных слов совпадает (44% и 56%); ядро оценочной лексики составляют прилагательные, а по мере расширения словаря начинают преобладать существительные; в словарях в среднем мало совпадающей лексики (30,0% для англоязычных и 20,5% для русскоязычных).

3. Невозможно однозначно говорить о преимуществе того или иного словаря оценочной лексики: отсутствуют масштабные исследования, в которых сравнивались бы основные существующие словари на базе единого подхода для одних и тех же текстовых корпусов.

Настоящий обзор позволяет сформулировать рекомендации для исследователей, имеющих потребность в словаре оценочной лексики. Такая потребность может быть удовлетворена либо за счет использования существующих словарей, либо с помощью разработки нового словаря. В первом случае рекомендуется использовать композиционный способ, т. е. формировать новый словарь на основе голосования существующих. При этом возможно контролировать баланс точности и полноты оценочной лексики – для высокой точности следует увеличивать порог количества словарей, проголосовавших за включение слова в новый словарь, а при снижении этого порога повышается полнота.

Если существующие словари не удовлетворяют исследователя (например, требуется словарь на другом языке или предметно-ориентированный словарь), то способ создания словаря будет зависеть от цели и имеющихся ресурсов:

- предметно-ориентированный словарь рекомендуется создавать с применением автоматического метода на основе корпусов;
- при наличии достаточных временных и/или финансовых ресурсов рекомендуется осуществить разметку словаря вручную;
- машинный перевод существующих словарей на другой язык, как правило, недостаточен для обеспечения высокой точности; требуется разметка или проверка вручную.

Таким образом, процесс исследования и разработки словарей оценочной лексики, в частности, русскоязычных, представляется незавершенным и требует дальнейших усилий научного сообщества. Одним из наиболее актуальных направлений является проведение масштабного исследования существующих словарей оценочной лексики на базе единого подхода, учитывающего преимущества словарных методов, в том числе с использованием композиции словарей.

СПИСОК ЛИТЕРАТУРЫ

1. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. – Cambridge: Cambridge University Press, 2015.
2. Taboada M. Sentiment Analysis: An Overview from Linguistics // Annual Review of Linguistics. – 2016. – Vol. 2. – P. 325–347.
3. Yue L., Chen W., Li X., Zuo W., Yin M. A survey of sentiment analysis in social media // Knowledge and Information Systems. – 2018. – P. 1–47.
4. Poria S., Hazarika D., Majumder N., Mihalcea R. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research // Computing Research Repository. – 2020. – arXiv: 2005.00357.
5. Hamilton W.L., Clark K., Leskovec J., Jurafsky D. Inducing domain-specific sentiment lexicons from unlabeled corpora // Proceedings of Conference on Empirical Methods in Natural Language Processing. – 2016. – P. 595–605.
6. Vo D.T., Zhang Y. Don't count, predict! An automatic approach to learning sentiment lexicons for short text // Proceedings of 54th Annual Meeting of the Association for Computational Linguistics. – 2016. – P. 219–224.
7. Wang L., Xia R. Sentiment Lexicon Construction with Representation Learning Based on Hierarchical Sentiment Supervision // Proceedings of Conference on Empirical Methods in Natural Language Processing. – 2017. – P. 502–510.
8. Liu B. Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. – 2012. – Vol. 5(1). – P. 1–167.
9. Боярский К.К., Каневский Е.А. Семантика устойчивых словосочетаний с глаголами // Научно-техническая информация. Сер. 2. – 2019. – № 11. – С. 23–31.
10. Multiword Units in Machine Translation and Translation Technology / eds. R. Mitkov, J. Monti, G.C. Pastor, V. Seretan. – Amsterdam: John Benjamins Publishing Company, 2018.
11. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. – The MIT Press, 1999. – 620 p.
12. Hutto C.J., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text // Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. – Palo Alto: The AAAI Press, 2014.
13. Abdaoui A., Azé J., Bringay S., Poncelet P. FEEL: a French Expanded Emotion Lexicon // Language Resources & Evaluation. – 2017. – Vol. 51(3). – P. 833–855.
14. Koltsova O.Yu., Alexeeva S.V., Kolcov S.N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2016". –2016. – Vol. 15(22). – P. 277–287.
15. WordNet. An electronic lexical database / ed. C. Fellbaum. – Cambridge, MA: MIT Press; 1998.
16. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во МГУ, 2011.

17. Kiritchenko S., Zhu X., Mohammad S. Sentiment Analysis of Short Informal Texts // Journal of Artificial Intelligence Research. – 2014. – Vol. 50. – P. 723–762.
18. Kotelnikova A.V., Kotelnikov E.V. SentiRusColl: Russian Collocation Lexicon for Sentiment Analysis // Artificial Intelligence and Natural Language Conference (AINL). Communications in Computer and Information Science (November 20–22, 2019. Tartu, Estonia). –Cham: Springer, 2019. – Vol. 1119. – P. 18–32.
19. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis // Proceedings of the 2005 Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-05). – 2005. – P. 347–354.
20. Kiritchenko S., Mohammad S.M. Happy Accident: A Sentiment Composition Lexicon for Opposing Polarities Phrases // Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC). – Portorož, Slovenia, 2016. – P. 1157–1164.
21. Kiritchenko S., Mohammad S.M. The Effect of Negators, Modals, and Degree Adverbs on Sentiment Composition // Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). – San Diego, California, 2016. – P. 43–52.
22. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis // Computational Linguistics. – 2011. – Vol. 37(2). – P. 267–307.
23. Mohammad S.M., Turney D.P. Crowdsourcing a word-emotion association lexicon // Computational Intelligence. – 2013. – Vol. 29(3). – P. 436–465.
24. Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon // Proceedings of Language Resources and Evaluation Conference LREC-2016. – 2016. – P. 1171–1176.
25. Bhatti S.S., Gao X., Chen G. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey // The Journal of Systems and Software. – 2020. – Vol. 167.
26. Hong Y., Kwak H., Baek Y. Tower of babel: a crowdsourcing game building sentiment lexicons for resource-scarce languages // Proceedings of the WWW 2013 companion. – Rio de Janeiro, Brazil, 13–17 May 2013. – New York: Association for Computing Machinery, 2013. – P. 549–556.
27. Thisone C.C., Ghasemi A., Faltings B. Sentiment analysis using a novel human computation game // Proceedings of the 3rd workshop on the people's web meets NLP, Jeju Island, Republic of Korea, 8–14 July 2012. – P. 1–9.
28. Kotelnikov E.V., Peskisheva T.A., Kotelnikova A.V., Razova E.V. A comparative study of publicly available Russian sentiment lexicons // 7th conference on Artificial Intelligence and Natural Language (AINL-2018). Communications in Computer and Information Science. – Cham: Springer, 2018. – Vol. 930. – P. 139–151.
29. Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining // Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC10). – 2010. – P. 2200–2204.
30. Cruz F.L., Troyano J.A., Pontes B., Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels // Expert Systems with Applications. – 2014. – Vol. 41. – P. 5984–5994.
31. Blinov P.D., Klekovkina M.V., Kotelnikov E.V., Pestov O.A. Research of lexical approach and machine learning methods for sentiment analysis // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2013". –2013. – Vol. 12(19). – P. 51–61.
32. Chen Y., Skiena S. Building Sentiment Lexicons for All Major Languages // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. – Baltimore, 2014. – P. 383–389.
33. Mohammad S.M., Kiritchenko S., Zhu X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets // Proceedings of the seventh international workshop on Semantic Evaluation – SemEval-2013 (June 2013, Atlanta, USA). – Madison: Omnipress, Inc., 2013. – P. 321–327.
34. Mikolov T., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // Proceedings of Conference on Neural Information Processing Systems. – 2013. – P. 3111–3119.
35. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation // Proceedings of Conference on Empirical Methods in Natural Language Processing. – 2014. – P. 1532–1543.
36. Almeida F., Xexeo G. Word Embeddings: A Survey // Computing Research Repository. – 2019. – arXiv:1901.09069.
37. Çano E., Morisio M. Word Embeddings for Sentiment Analysis: A Comprehensive Empirical Survey // Computing Research Repository. – 2019. – arXiv: 1902.00753.
38. Liu Q., Kusner M.J., Blunsom P. A Survey on Contextual Embeddings // Computing Research Repository. – 2020. – arXiv :2003.07278v.
39. Cambria E., Poria S., Hazarika D., Kwok K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). – 2018. – P. 1795–1802.
40. Loughran T., McDonald B. When is a liability not a liability? Textual Analysis, Dictionaries and 10-Ks // The Journal of Finance. – 2011. – Vol. 66(1). – P. 35–66.
41. Hu M., Liu B. Mining and Summarizing Customer Reviews // Proceedings of the ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining – KDD-2004 (Aug 22-25, 2004, Seattle, Washington, USA). – New York: Association for Computing Machinery, 2004. – P. 168–177.

42. Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R. The viability of web-derived polarity lexicons // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. – 2010. – P. 777–785.
43. Zhu X., Ghahramani Z. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMUCALD-02-107. – Carnegie Mellon University, 2002.
44. Hassan A., Radev D.R. Identifying Text Polarity Using Random Walks // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. – 2010. – P. 395–403.
45. Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Transactions on Affective Computing. – 2016. – Vol. 7(4). – P. 409–421.
46. Socher R., Perelygin A., Wu J., Chuang J., Manning C., Ng A., Potts C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). – 2013. – P. 1631–1642.
47. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2002. – P. 79–86.
48. Stone P.J., Dunphry D.C., Smith M.S., Ogilvie D.M. The General Inquirer: A Computer Approach to Content Analysis. – Cambridge, MA: MIT Press, 1966.
49. Pennebaker J.W., Boyd R.L., Jordan K., Blackburn K. The development and psychometric properties of LIWC2015. – Austin, TX: University of Texas at Austin, 2015.
50. Bradley M.M., Lang P.J. Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings (Tech. Report C-1). – Gainesville: University of Florida, Center for Research in Psychophysiology, 1999.
51. Riloff E., Wiebe J. Learning Extraction Patterns for Subjective Expressions // Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03). – Stroudsburg: Association for Computational Linguistics, 2003. – P. 105–112.
52. Nielsen F. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs // Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages, Heraklion. – 2012. – P. 93–98.
53. Rosenthal S., Nakov P., Kiritchenko S., Mohammad S., Ritter A., Stoyanov V. SemEval-2015 Task 10: Sentiment Analysis in Twitter // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). – 2015. – P. 451–463.
54. WordStat: content analysis and text mining software. – URL: https://provalisresearch.com/products/content-analysis-software/worldstat-dictionary/sentiment-dictionaries (дата обращения: 01.08.2020).
55. Chetviorkin I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // Proceedings of COLING 2012. – Mumbai, 2012. – P. 593–610.
56. Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A.A. Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. – 2010. – Vol. 61(12). – P. 2544–2558.
57. Kotelnikov E., Bushmeleva N., Razova E., Peskisheva T., Pletneva M. Manually Created Sentiment Lexicons: Research and Development // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2016”. – 2016. – Vol. 15(22). – P. 300–314.
58. Тутубалина Е.В. Методы извлечения и резюмирования критических отзывов пользователей о продукции: дис. … канд. физ.-мат.. наук. – М.: ИСП РАН, 2016. – 145 с.
59. Кулагин Д.И. Карта слов: переосмысление подхода к составлению онлайн-словарей в постмобильную эру // Международная конференция «Диалог 2017» – Компьютерная лингвистика и интеллектуальные технологии (Москва, 31 мая – 3 июня 2017 г.). – URL: http://www.dialog-21.ru/media/3974/kulagindi.pdf (дата обращения: 01.08.2020).
60. Cambria E., Fu J., Bisio F., Poria S. AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. – 2015. – P. 508–514.
61. Vilares D., Peng H., Satapathy R., Cambria E. BabelSenticNet: A Commonsense Reasoning Framework for Multilingual Sentiment Analysis // Proceedings of IEEE Symposium Series on Computational Intelligence. – 2018. – P. 1292–1298.
62. Razova E.V., Kotelnikov E.V. Concentration Areas of Sentiment Lexica in the Word Embedding Space // International Journal of Cognitive Informatics and Natural Intelligence. – 2019. – Vol. 13(2). – P. 48–62.
63. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. – Stroudsburg: Association for Computational Linguistics, 2020.
64. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Proceedings of 3rd Conference on Analysis of Images, Social Networks and Texts (AIST). – 2015. – P. 320–332.
65. Hosel C., Roschke C., Thomanek R., Ritter M. Lexicon-Based Sentiment Analysis of Online Customer Ratings as a Quinary Classification Problem // Communications in Computer and Information Science. – 2019. – Vol. 1034. – P. 75–80.
66. Kotelnikov E.V., Pletneva M.V. Text Sentiment Classification based on Genetic Algorithm and Word and Document Co-clustering // Journal of

Computer and Systems Sciences International. – 2016. – Vol. 55(1). – P. 106–114.
67. Han H., Zhang Y., Zhang J., Yang J., Zou X. Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias // PLOS ONE. – 2018. – Vol. 13(8). – P. 1–11.
68. Khatun F., Chowdhury S., Tumpa Z., Rabby S., Hossain S., Abujar S. Sentiment Analysis of Amazon Book Review Data Using Lexicon Based Analysis // Advances in Intelligent Systems and Computing. – 2019. – Vol. 1108. – P.1303–1309.
69. Alharbi J.R., Alhalabi W.S. Hybrid Approach for Sentiment Analysis of Twitter Posts Using a Dictionary-based Approach and Fuzzy Logic Methods: Study Case on Cloud Service Providers // International Journal on Semantic Web and Information Systems. – 2020. – Vol. 16(1). – P. 116–145.
70. Ikoro V., Sharmina M., Malik K., Batista-Navarro R. Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers // 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). – 2018. – P. 95–98.
71. Iqbal F., Maqbool J., Fung B., Batool R., Khattak A., Aleem S., Hung P. A Hybrid Framework for Sentiment Analysis using Genetic Algorithm based Feature Reduction // IEEE Access. – 2019. – Vol. 7. – P. 14637–14652.
72. Vo D.T., Zhang Y. Don't count, predict! An automatic approach to learning sentiment lexicons for short text // Proceedings of 54th Annual Meeting of the Association for Computational Linguistics. – 2016. – P. 219–224.
73. Bansal B., Srivastava S. Lexicon-based Twitter sentiment analysis for vote share prediction using emoji and N-gram features // International Journal of Web Based Communities. –2019. – Vol. 15(1). – P. 85–99.
74. Jiménez-Zafra S.M., Martín-Valdivia M.T., Molina-González M.D., Ureña-López L.A. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain // Artificial Intelligence in Medicine. – 2019. – Vol. 93. – P. 50–57.
75. Kumar C.S.P., Babu L.D.D. Evolving dictionary based sentiment scoring framework for patient authored text // Evolutionary Intelligence. – 2020.
76. Wook M., Razali N., Ramli S., Wahab N., Hasbullah N., Zainudin N., Talib M. Opinion mining technique for developing student feedback analysis system using lexicon-based approach (OMFeedback) // Education and Information Technologies. – 2020. – Vol. 25. – P. 2549–2560.
77. Tutubalina E., Nikolenko S. Constructing Aspect-Based Sentiment Lexicons with Topic Modeling // Proceedings of 5th Conference on Analysis of Images, Social Networks and Text. –2017. – P. 208–220.
78. Котельников Е.В. Метод анализа тональности текстов TextJSM // Научно-техническая информация. Сер. 2. – 2018. – № 2. – С. 8–20.

Материал поступил в редакцию 07.08.20.

Сведения об авторах

КОТЕЛЬНИКОВ Евгений Вячеславович – доктор технических наук, доцент, профессор кафедры прикладной математики и информатики, Вятский государственный университет.
e-mail: kotelnikov.ev@gmail.com

РАЗОВА Елена Владимировна – кандидат педагогических наук, доцент, доцент кафедры прикладной математики и информатики, Вятский государственный университет.
e-mail: razova.ev@gmail.com

КОТЕЛЬНИКОВА Анастасия Валерьевна – кандидат педагогических наук, доцент кафедры прикладной математики и информатики, Вятский государственный университет.
e-mail: kotelnikova.av@gmail.com

ВЫЧЕГЖАНИН Сергей Владимирович – инженер кафедры прикладной математики и информатики, Вятский государственный университет.
e-mail: vychegzhaninsv@gmail.com

Указатель статей, опубликованных в сборнике «Научно-техническая информация», и Авторский указатель за 2020 год

Указатель статей

ОБЩИЙ РАЗДЕЛ

ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

* 1 – означает номер сборника, (1) – серию, 1 - страницу

ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

ИНФОРМАЦИОННЫЙ АНАЛИЗ

ИНФОРМАЦИОННЫЕ ЯЗЫКИ

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

НАМ ПИШУТ

Авторский указатель

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

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

Телефон 499-155-42-85 499-151-78-61

E-mail: Contact@viniti.ru, Feo@viniti.ru