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. **№ 8** **Москва 2018**

ОБЩИЙ РАЗДЕЛ

УДК 519.2:303.71

В.А. Яцко

Теория вероятностей и методы обработки информации. Простые вероятностные величины*

На основе анализа государственных образовательных стандартов показывается значение теории вероятностей для различных предметных областей. Даётся общая характеристика теории вероятностей, рассматриваются свойства простых вероятностных величин и методы их вычислений. Особое внимание уделяется использованию простых вероятностных величин с целью сглаживания разниц в размерах статистических выборок.

Ключевые слова: *теория вероятностей, простые вероятностные величины, свойства и способы вычислений, методы выравнивания диспропорциональных статистических выборок*

ВВЕДЕНИЕ

Последние десятилетия характеризуются дифференциацией компьютерной науки (*computer science*) и информационной науки (*information science*), которая проявляется в формировании и быстром развитии информационных направлений научных дисциплин, таких как биоинформатика, медицинская информатика, химическая информатика, правовая информатика, историческая информатика, лингвистическая информатика [1]. Основная задача этих направлений – разработка информационных технологий с учётом специфики конкретных дисциплин, предметных областей, сфер деятельности. Особое значение для разработки предметно-ориентированных технологий имеют вероятностно-статистические методы, кото-

* Исследование поддержано грантом РФФИ № 16-07-00014

рые применяются во всех предметных областях и составляют математическую основу современной информационной науки.

Значение вероятностно статистических методов для решения теоретических и прикладных проблем разработки современных информационных технологий нашло отражение в содержании государственных образовательных стандартов, а также учебных планов и программ. В Российской Федерации изучение теории вероятностей и математической статистики предусмотрено государственными стандартами по всем математическим, естественно-научным и техническим направлениям обучения в высшей школе. В табл. 1 представлены данные, которые показывают значение теории вероятностей для направлений подготовки по математическим и информационным дисциплинам высшей школы (бакалавриат) в соответствии с требованиями государственных стандартов [2].

Теория вероятностей как обязательный компонент обучения указана в образовательных стандартах и по ряду гуманитарных дисциплин. Студенты, обучающиеся по специальности 0358000 *Фундаментальная и прикладная лингвистика* должны знать основы теории вероятности, математической статистики. Изучение математической статистики, методов математического моделирования является обязательным для студентов, обучающихся по специальности 030600 *История*. В естественно-научный цикл ряда гуманитарных специальностей включено изучение дисциплин, представляющих информационные направления, например, *Информационные технологии в лингвистике*, *Информационные технологии в юридической деятельности*.

Теория вероятностей присутствует и в программе обучения ведущих зарубежных университетов. Так, в Стэндфордском университете в базовую программу обучения студентов четвертого курса, специализирующихся в области информатики, входит десятинедельный курс *Probability Theory for Computer Scientists* [3]. В Гарвардском университете в программу обучения аспирантов по биомедицине включены курсы *Introduction to Probability, Introduction to Theoretical Statistics*, каждый из которых изучается в течение семестра[1]. В Оксфордском университете студенты технического факультета, специализирующиеся в области робототехники и искусственного интеллекта изучают курс *Advanced probability theory* [2].

Таблица 1

Теория вероятностей в государственных стандартах по математическим и информационным направлениям подготовки

Область науки	Направление подготовки	Учебный цикл	Требования к знаниям и умениям	Дисциплина	Год утверждения
010000 Физико-математические науки	010400 Прикладная математика и информатика	Математический и естественнонаучный цикл. Базовая часть	Знать методы теории вероятностей и математической статистики	Теории вероятностей и математическая статистика	2010
		Профессиональный цикл. Базовая часть	Знать и уметь применять на практике методы теории вероятностей и математической статистики		
	010300 Фундаментальная информатика и информационные технологии	Математический и естественнонаучный цикл. Базовая часть	Знать основные понятия и методы теории вероятностей и математической статистики	Теория вероятностей и математическая статистика	2009
230000 Информатика и вычислительная техника	230100 Информатика и вычислительная техника	Математический и естественнонаучный цикл. Базовая часть	Знать основы теории вероятностей и математической статистики. Владеть методами теории вероятностей и математической статистики	Математика	2009
	23400 Информационные системы и технологии	Математический и естественнонаучный цикл. Базовая часть	Знать основные понятия и методы теории вероятностей и математической статистики	Математика	2010
	23700 Прикладная информатика	Математический и естественнонаучный цикл. Базовая часть	Знать методы статистического анализа. Уметь вычислять вероятности случайных событий	Теория вероятностей и математическая статистика	2009

[1] https://www.hms.harvard.edu/dms/bbs/documents/Quantitative%20Course%20Resources.pdf

[2] http://www.robots.ox.ac.uk/~mosb/c24/

Не будет преувеличением сказать, что знание теории вероятностей является обязательным для квалифицированных специалистов в области информационных технологий. В этой связи приобретает особую актуальность системное описание методов и процедур обработки информации на основе теории вероятностей. Данная статья посвящена описанию простых вероятностных величин, которые лежат в основе ряда других вероятностно-статистических методов обработки информации.

ОБЩАЯ ХАРАКТЕРИСТИКА ТЕОРИИ ВЕРОЯТНОСТЕЙ И ПРОСТЫЕ ВЕРОЯТНОСТНЫЕ ВЕЛИЧИНЫ

К компонентам современной теории вероятностей, которые используются наиболее широко для решения теоретических и прикладных задач, относятся теорема Байеса, Марковские цепи, а также вероятностно-статистические метрики, такие как хи-квадрат, прирост информации (information gain), отношение шансов (odds ratio). На их основе в разных предметных областях решаются классификационные и прогностические задачи, определяются пороговые уровни, находятся пределы, выполняется интервальный анализ.

Одним из основных понятий теории вероятностей является понятие случайного события – события, которое может произойти либо не произойти [4, с.17]. Случайные события отличаются от достоверных (неизбежных) и невозможных. Если взять какой-нибудь предмет, а затем разжать пальцы, то он неизбежно упадёт вниз, а его полёт вверх невозможен в соответствии с законом всемирного тяготения. Соответственно, теория вероятностей применяется к случайным событиям, а её применение к неизбежным и невозможным событиям исключено. Мы также полагаем, что возможно выделить закономерные события, вероятность наступления/ненаступления которых очень высока. Применять теорию вероятности к таким событиям возможно, но малоцелесообразно. К закономерным относятся события, обусловленные законами и принципами развития личности и общества, которые, в отличие от законов природы, предусматривают исключения. В соответствии с законом спроса существует обратная зависимость между повышением цены и спросом на товар, повышение цены снижает спрос и наоборот. Однако возможны ситуации, когда, напротив, при повышении цены увеличивается спрос, а при понижении цены спрос уменьшается. Другой пример: студент проучился три года, не имеет академических задолженностей. Какова вероятность того, что этот студент закончит университет и получит диплом? Очевидно, что эта вероятность не составляет 100%; известны случаи отчисления и хороших студентов с последнего курса. Однако это именно исключения. Теорию вероятности в данном случае можно применить, проанализировав данные исключения, однако, целесообразно ли это делать, если вероятность положительного исхода в соответствии со здравым смыслом составляет больше 90%?

Ещё одно основное понятие теории – само понятие вероятности, которое интерпретируется как количественная характеристика возможности наступления случайного события, вычисляемая на основе соотношения между количеством исходов данного события и общим количеством исходов совместных событий по формуле:

$$P(i)=\frac{f(i\in S)}{f(S)}\ , \qquad (1)$$

где: i – данное событие (событие для которого и высчитывается вероятностная величина), f – количество исходов; P – вероятность; S – множество всех совместных событий, т.е. событий, которые наступают совместно, поскольку входят в одно пространство.

Простая вероятностная величина имеет следующие свойства: 1) $f(i) \neq f(S)$. В случае равенства вероятность будет равна 1, что характеризует достоверное событие, а не случайное; 2) $f(i)$ не может быть больше $f(S)$, получение такого результата указывает на ошибку в вычислениях; 3) из первых двух свойств следует, что $0<P(i)<1$; 4) $\sum P(i)=1$. Другой результат указывает на ошибку в вычислениях (если не используется модифицированная формула); 5) вероятностные величины легко конвертируются в проценты по формуле $P(i) * 100$.

Обычно применение формулы (1) иллюстрируется на примерах опытов с подбрасыванием монетки и вытаскиванием карты из колоды. Если подбросить монету, то вероятность выпадения решки по формуле (1) составит 1/2, поскольку $f(S)=2$ (орёл и решка), а $f(i)=1$ (решка). Вероятность того, что из колоды будет вытащена дама составит 4/36=1/9, поскольку в колоде из 36 карт 4 дамы. Эти примеры показывают различие между элементарными и сложными событиями. Элементарные события наступают 1 раз; количество исходов сложных событий больше одного.

Понятие простой вероятностной величины часто применяется в лингвистической информатике для вычисления вероятности вхождения лингвистических единиц (символов алфавита, слов, словосочетаний) в некоторый текст. В этом случае i – лингвистическая единица, вероятностная величина для которой рассчитывается по формуле:

$$P_{(i)}=\frac{f(i\in S)}{\sum f(S)}\ , \qquad (2)$$

где f – частотность вхождения лингвистической единицы в текстовый документ.

Для вычислений следует сначала получить исходные статистические данные: частотности лингвистических единиц в данном тексте. Частотности символов алфавита некоторого текста можно получить с помощью формулы MS Excel:

(ДЛСТР(A$3)-
(ДЛСТР(ПОДСТАВИТЬ(СТРОЧН(A$3);C4;""))))), (3)

где A$3 – адрес ячейки с текстом, C4 – адрес символа. По этой формуле вначале подсчитывается общее количество символов в тексте (функция ДЛСТР),

Таблица 2

Статистические данные и вычисление простой вероятностной величины (выборка)

Программное обеспечение	Текст	Символ/ слово	частотность	$\Sigma f(S)$	$P(i)$
MS Excel	Bayes theorem is one of the foundations of contemporary theory of probability	o	11	62	0,177419
		e	7		0,112903
		t	6		0,096774
		r	5		0,080645
		a	4		0,064516
AntConc	Газетная статья *Weighing the risks of liposuction*[4]	the	68	1418	0,0479549
		a	51		0,0359661
		of	37		0,0260931
		to	28		0,0197461
		in	25		0,0176305

затем вместо указанного символа подставляется (функция ПОДСТАВИТЬ) пустое место (двойные кавычки формуле) и снова подсчитывается количество символов. Частотность символа определяется как разница между первой и второй величиной, при этом функция СТРОЧН позволяет получать результат без учёта регистра. Для получения статистических данных в ячейку вставляется текст, составляется в столбик список символов алфавита, формула (3) применяется для обработки первого символа и растягивается. В табл. 2 приводятся полученные статистические данные и результаты вычислений.

Ещё одна формула

```
(ДЛСТР(A$3)-
(ДЛСТР(ПОДСТАВИТЬ(СТРОЧН(A$3);          (4)
C4;"")))))/ДЛСТР(C4)
```

позволяет получать данные о частотности определённых сочетаний символов (биграм, триграм, тетраграм, пентаграм, гексаграм). Вычисление частотностей и вероятностных величины символов алфавита и их сочетаний имеют непосредственное значение для систем оптического распознавания символов, криптографии [5], а также социальных сетей и служб, в которых предусмотрены ограничения на размер сообщений.

Заметим, что размер текста в ячейке MS Excel ограничен 32767 знаками, поэтому для обработки текстов большого объёма целесообразно применять специализированное программное обеспечение. На различных интернет ресурсах приводятся исходные коды программ, подсчитывающих количество символов алфавита в тексте[3]. Для получения статистических данных о распределении слов в тексте целесообразно использовать программы статистического анализа – конкордансы (см. табл. 2). Например, приложение AntConc [6]. Приложение распространяется бесплатно, поддерживает английский, русский и ряд других языков, не уступая по функциональности платным аналогам. С его помощью можно также получать данные о распределении словосочетаний, выполнять удаление стоп-слов, подсчитывать коэффициенты терминов по метрикам хи-квадрат и логарифмическое подобие (log-likelyhood).

ПРИКЛАДНОЕ ЗНАЧЕНИЕ ВЕРОЯТНОСТНЫХ ВЕЛИЧИН

В лингвистической информатике вычисление простых вероятностных величин не только позволяет нейтрализовать разницу в размерах текстов, вычислять пороговые уровни, выступать в качестве основы классификации и авторской атрибуции текстов, но и применяется в качестве одного из методов взвешивания, способствующего распознаванию ключевых терминов тексте. Сырые частотности используются достаточно редко и только в случае, если анализируется распределение терминов в одном тексте. В качестве примера можно привести разработанную нами методику симметричного взвешивания терминов с целью автоматического реферирования [7]. Эта методика предусматривает определение числового коэффициента (веса) каждого предложения данного текста, составление ранжированного списка предложений и выбор из верхней части списка некоторого количества предложений, чьи коэффициенты превышают установленный пороговый уровень. Вес предложения определяется по сумме вхождений терминов словаря (основ слов), используемых в данном предложении, в другие предложения текста, то есть на основе сырых частотностей. В том случае, если анализ данных требует сопоставления диспропорциональных выборок, сырые частотности обычно не используются, так как частотность событий напрямую зависит от размера выборки. В биологии, медицине, эпидемиологии, политологии распределение некоторого признака/признаков по популяциям или группам зависит от их размера. В процессе информационного поиска, классификации и авторской атрибуции текстов, их интеллектуального анализа исследуется распределение терминов по текстовым документам разных размеров, что требует нормализации полученных сырых частотностей с целью сглаживания разниц в размерах,

[3] См., например, https://codereview.stackexchange.com/questions/63872/counting-the-number-of-character-occurrences ; https://mathematica.stackexchange.com/questions/110636/analyzing-frequency-of-individual-letters-in-a-large-body-of-text

[4] http://www.gutenberg.org/wiki/Main_Page

Таблица 3

Простые вероятностные величины и сглаживание разниц в размерах выборок

№	*i*	f1(*i*∈t1)	f2(*i*∈t2)	Δ1=f1/f2	P1(*i*∈t1)	P2(*i*∈t2)	Δ2=ABS(P1/P2)	Δ1/Δ2
1	*the*	8612	454	18,969	0,042473	0,03818	1,112441	17,05169
2	*to*	6350	362	17,541	0,031317	0,030443	1,028709	17,05147
3	*and*	6316	386	16,363	0,031149	0,032462	0,959553	17,05273
4	*of*	5543	272	20,379	0,027337	0,022874	1,195112	17,05196
5	*he*	4847	147	32,973	0,023904	0,012362	1,933668	17,05205

устранения зависимости полученных статистических данных от величины выборок. Вероятностно-статистические методы, в том числе и вычисление простых вероятностных величин, предусматривают такую нормализацию. В качестве примера можно привести распределение определённого артикля в двух корпусах английского языка.

В *Corpus of Contemporary American English* наиболее частотное слово английского языка *the* встречается 30 890 521 раз, а *British National Corpus* – 5 973 437 раза, что более чем в пять раз меньше. Однако если соотнести эти сырые частотности с общим количеством токенов, т.е. с суммой частотностей слов (560 млн и 100 млн соответственно), то получается, что плотность распределения определённого артикля в британском варианте английского языка даже несколько (на 0,047) выше, чем в американском варианте, поскольку для него $P(the \in BNC)$=0,597, в то время как $P(the \in COCA)$=0,55. Таким образом, использование вероятностных величин позволяет получать более достоверную информацию о распределении событий.

Наглядно значение вероятностных величин для сглаживания разниц размеров выборок можно представить, проведя анализ распределения слов по частотности в текстах разных размеров. Для этого нами с сайта *Project Gutenberg*[5] были взяты роман Т. Драйзера *The Financier* (t1) объёмом 202766 токенов (10929 уникальных слов) и рассказ У.С. Моэма *The Merry-go-round* (t2) размером 11891 токен (2605 уникальных слов). В табл. 3 представлены статистические данные, показывающие, что с помощью простых вероятностных величин разница (Δ) между сырыми частотностями пяти наиболее частотных слов двух текстов была уменьшена примерно в 17 раз.

Данные, приводимые в табл. 3, также показывают, что простые вероятностные величины не позволяют полностью сгладить разницу в размерах текстов, и это может сказаться на достоверности получаемых результатов. Соответственно, возникает потребность разработки дополнительных методов сглаживания и выравнивания размеров выборок. Наиболее простым методом, используемым в разных предметных областях, является выравнивание по нижнему или верхнему пределу [8]. Выравнивание по нижнему пределу (*undersampling*) предполагает уменьшение большей по объёму выборки до размеров меньшей выборки. Выравнивание по верхнему пределу (*oversampling*) предусматривает увеличение меньшей по объёму выборки до размера большей выборки. Если брать лингвистический материал, то в первом случае удаляется часть большего по объёму текста для того, чтобы уравнять его с размером текста, меньшего по объёму. Во втором случае части меньшего по объёму текста копируются, чтобы уравнять его с размером текста большего по объёму. И в том, и в другом случае происходит искусственное изменение структуры текста, что может отрицательно сказаться на конечных результатах анализа. Поэтому в лингвистической информатике с целью сглаживания разниц в размерах текстов часто применяются нормализация по косинусу и величины, вычисляемые по соотношению частотностей данного термина и термина с первым рангом, которые находятся по формуле:

$$P_{r(ij)} = \frac{f_{(ij)}}{f\left(R1_j\right)}\ , \qquad (5)$$

где $R1$ – термин с первым рангом в документе j-ом. Вес самого этого слова либо игнорируется, либо вычисляется делением его частотности на количество уникальных слов в данном документе. Получившиеся величины характеризуются теми же свойствами, что и простые вероятностные величины, но поскольку они вычисляются на основе соотношения с частотностью первого элемента в ранжированном списке, их можно назвать ранжированными вероятностными величинами (P_r). Если применить эти величины к статистическим данным, указанным в табл. 3, то окажется, что разница между сырыми частотностями слов 2-5 (*to*, *and*, *of*, *he*) уменьшается соответственно в 16,2208, 14,1146, 21,8933, 57,3147 раз. Сравним для слова to∈t1 и to∈t2: 6350/8612≅0,737343; 362/454≅0.7973568, следовательно, Δ2≅1,0813919709, что даёт 17,541/1,0813919709≅16,2208. Нормализация по косинусу выполняется по формуле [9]:

$$WL_j = \frac{1}{\sqrt{\sum_{i=0}^{m}\left(WC_i W_{ij}\right)^2}}\ , \qquad (6)$$

где: WLj – нормализованный весовой коэффициент документа j-го; Wij – весовой коэффициент термина i-го в данном документе; WC – весовой коэффициент

[5] https://www.sfgate.com/health/article/Weighing-the-risks-of-liposuction-S-F-woman-s-2826015.php

термина в коллекции (корпусе) документов (глобальный вес термина), который часто находится по формуле *IDF* (*inverse document frequency*):

$$WC_i = \log_2\left(\frac{N}{n}\right), \quad (7)$$

где N – количество документов в корпусе, а n – количество документов, в которых данный термин встречается хотя бы один раз. В [9] приводятся и другие варианты вычисления глобального весового коэффициента.

Заметим, что формула (6) применяется после того, как было проведено взвешивание и построены модели документов, в отличие от методики выравнивания, которая, как правило, выполняется до проведения статистического анализа. В информационном поиске эта формула позволяет устранять зависимость распределения терминов запроса от размеров документов [10].

Нами была предложена оригинальная методика логарифмического выравнивания [11], которая также применяется к результатам статистического анализа. Эта методика предусматривает вычисление коэффициента логарифмического соотношения между размерами выборок по формуле:

$$Q = \log_{T1} T2, \quad (8)$$

где $T1$ – количество элементов в меньшей по размеру выборке, а $T2$ – количество элементов в большей по размеру выборке. Применительно к рассмотренным выше текстам $Q=\log_{1189} 202766=1{,}30226$. Далее коэффициент Q применяется в качестве поправки к результатам статистического анализа по формуле:

$$A_U = \sqrt[Q]{P_U}, \quad (9)$$

где P_U – коэффициент меньшей выборки, выраженный простой вероятностной величиной. Таким способом можно, например, выровнять длины векторов моделей двух текстовых документов разных размеров. Логарифмическое выравнивание статистических выборок может применяться в любой предметной области.

ЗАКЛЮЧЕНИЕ

Последние десятилетия ознаменовались интенсивным развитием предметно-ориентированных информационных технологий, которые разрабатываются в рамках отдельных направлений научных дисциплин, и в названии которых обычно используется термин "информатика". Очевидно, что необходима обобщающая научная дисциплина, в рамках которой изучаются закономерности развития предметно-ориентированных информатик. Такой дисциплиной, как мы полагаем, является информационная наука. С позиций информационной науки можно выделить следующие основные задачи, которые реализуются в рамках предметно-ориентированных информатик: 1) создание специализированного аппаратного обеспечения; 2) разработка специализированное программного обеспечения, языков программирования и сред; 3) создание онтологий, отражающих структуру предметной области; 4) разработка предметно-ориентированных информационно-поисковых языков и систем, в первую очередь фактографического типа; 5) разработка методов моделирования предметной области; 6) создание специализированного математического аппарата, математических методов и метрик, в первую очередь вероятностно-статистических.

В настоящей статье предпринята попытка обобщить применение простых вероятностных величин с целью обработки информации. Проведённый анализ позволяет сделать следующие выводы.

Предложенная нами формула простой вероятностной величины (1), как мы считаем, более адекватно отражает взаимосвязи между переменными, чем приводимая в некоторых источниках [4, с. 18-19] формула "классического определения" статистической вероятности:

$$P(A) = \frac{m}{n}. \quad (10)$$

Вместе с тем, термин "классический" который используется и в зарубежных работах[6], можно применять с целью разграничения формулы (1) и ранжированной формулы (5).

Полагаем, что можно также выделить редуцированную формулу:

$$P_{(i)} = \frac{f(i \in S)}{\sum f(S) - f(i)}, \quad (11)$$

т.е. частотность события i-го не учитывается в сумме частотностей всех событий.

Таким образом, простые вероятностные величины могут быть представлены тремя формулами: классической, ранжированной и редуцированной. Последние две формулы характеризуются теми же (перечисленными выше) свойствами, что и классическая формула за исключением четвёртого: сумма вероятностных величин совместных событий не равна единице.

Одним из преимуществ простых вероятностных величин является возможность использования их как для интрастатистического, так и для интерстатистического анализа. Под интрастатстическим анализом мы понимаем анализ распределений в рамках одной статистической выборки, интерстатистический анализ предполагает сопоставление распределения событий в двух и более выборках. Типичный пример интерстатистического анализа – формула (7).

Простые вероятностные величины широко используются с целью сглаживания разниц в размерах статистических выборок. В среднем наибольшей силой сглаживания обладает ранжированная формула, а наименьшей – редуцированная. Классическая формула даёт более равномерное распределение результатов.

[6] http://highhopes.com/probability.html

Рассматриваемые величины могут выступать основой для построения моделей объектов и предметных областей. На основе простых вероятностных величин вычисляются условные вероятностные величины, которые также широко используются в информационных направлениях различных предметных областей, что подтверждает анализ государственных стандартов высшего профессионального образования и учебных планов ведущих зарубежных университетов. Обзор и анализ условных вероятностных величин – задача нашей следующей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. Яцко В.А. Принципы исследования исторического развития информатики // Научно-техническая информация. Сер.1. – 2017. – № 9. – С. 1-9; Yatsko V.A. The Principles for the Investigation of the Historical Development of Computer Science // Scientific and Technical Information Processing. – 2017. – Vol. 44, №. 3. – P. 207–214.
2. Федеральные государственные образовательные стандарты высшего профессионального образования по направлениям подготовки бакалавриата / Министерство образования и науки Российской Федерации. – 2016. – URL: https://минобрнауки.рф/документы/924.
3. Sahami M. A course on probability theory for computer scientists // SIGCSE 11. Proceedings of the 42nd ACM technical symposium on computer science education. – NY, 2011. – P. 263-268. – URL: http://robotics.stanford.edu/users/sahami/papers-dir/SIGCSE11-Probability.pdf.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов. – 9-е изд. – М.: Высш. шк., 2003. – 479 с.
5. Egg A. Cryptographic cipher-attack through statistical frequency analysis. – 2013 – URL: http://www.eggie5.com/46-cryptographic-cipher-attack-through-statistical-frequency-analysis
6. Anthony L. AntConc version 3.4.4w. – Tokyo, Japan: Waseda University, 2016. – URL: http://www.laurenceanthony.net/software.
7. Яцко В.А. Методика симметричного взвешивания предложений // Научно-техническая информация. Сер. 2. – 2016. – № 2. – С. 36–41.
8. García S., Herrera F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy // Evolutionary Computation. – 2009. – Vol. 17, Iss. 3. – P. 275–306. – URL: https://www.mitpressjournals.org/doi/pdf/10.1162/evco.2009.17.3.275.
9. Chisholm E., Kolda T.G. New term weighting formulas for the vector space method in information retrieval. – 1999. – URL: http://www.sandia.gov/~tgkolda/pubs/pubfiles/ornl-tm-13756.pdf.
10. Lee D.L., Huei C., Kent S. Document ranking and the vector space model // IEEE software. 1997. – Vol 14, № 2. – P. 67-75. – URL: http://sir-lab.usc.edu/cs585/sr/Document%20Ranking%20and%20the%20vector_space%20model.pdf.
11. Yatsko V. Zonal text processing // Digital scholarship in the humanities. – 2016. – Vol. 31, Iss. 4. – P. 773–781.

Материал поступил в редакцию 05.04.18.

Сведения об авторе

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

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 343.982.43.065:[044.89:510.64]

С.М. Гусакова, И.А. Лапшина, А.Н. Охлупина

Идентификация подписи: постановка задачи и вариант решения с помощью интеллектуальной ДСМ-системы

Для объективизации выводов эксперта-почерковеда при проведении им почерковедческой экспертизы, сокращения времени и уменьшения трудоемкости этой работы актуально применение компьютерных методов и, в частности, методов интеллектуального анализа данных. Описывается интеллектуальная ДСМ-система, разрабатываемая для решения задачи идентификации подписи и проведения исследований в области почерковедения.

Ключевые слова: *идентификация подписи, признаки почерка, ДСМ-система, модифицированный ДСМ-метод, операция сходства*

ВВЕДЕНИЕ

Компьютерная поддержка работы эксперта-криминалиста при решении им задачи идентификации подписи является для криминалистического сообщества насущной проблемой. Все создаваемые автоматизированные системы проверки подписи используют либо статическую (написанную на бумаге) подпись, либо динамическую, исполненную на экране графического планшета. Динамическая подпись позволяет программно определить скорость, длину, силу нажима, координаты точек начала и конца фрагментов, экстремумов, а также ряд других признаков в процессе выполнения подписи. Это дает возможность как отечественным, так и зарубежным специалистам создать целый спектр методов и алгоритмов верификации подписи [1–4].

Однако на практике эксперт, как правило, имеет дело со статичными подписями.

Существуют компьютерные системы и отдельные программы для объективизации выделения признаков и автоматического применения решающего правила [5], автоматизации работы с методиками [6].

В последние годы в России и за рубежом активно разрабатываются автоматические системы идентификации подписи, использующие различные методы распознавания образов, в которых отсканированная подпись рассматривается как рисунок [7–9]. Несмотря на то, что эти системы дают хорошие результаты, полностью заменить эксперта они не могут, поскольку, во-первых, при некоторых способах фальсификации подписи методы распознавания образов одинаково описывают и подлинник и фальсификат, а во-вторых, суд признает только обоснованное заключение эксперта. Поэтому эти системы являются вспомогательными. В связи с этим актуальной задачей остается создание не автоматической, а автоматизированной системы, моделирующей рассуждения эксперта и позволяющей проводить исследовательскую работу.

В [10] для решения идентификационной задачи почерковедческой экспертизы использовался модифицированный ДСМ-метод [11]. Идентификационная задача формулировалась следующим образом: определить неизвестного исполнителя краткого текста среди нескольких известных исполнителей текстов. В результате проведенных с помощью ДСМ-системы экспериментов, было установлено, что метод дает хорошую точность прогноза и для решения задачи надо использовать прописные буквы.

Однако гораздо чаще эксперт-криминалист вынужден решать другую идентификационную задачу – выполнена ли конкретная подпись лицом, подписи которого представлены в образцах на исследование. При решении этой задачи эксперт сталкивается с рядом трудностей, связанных, в первую очередь, с выделением признаков подписи. Одна из проблем – это субъективизм процесса выделения и описания признаков подписи, который определятся множеством факторов. К ним относят личностные качества эксперта, методику обучения при получении допуска на право производства подобного рода исследований, стаж экспертной деятельности в области проведения почерковедческих экспертиз и другие. Анализ экспертных заключений показывает, что один и тот же вывод обосновывается совокупностью различных признаков.

Еще одна трудность, возникающая при работе эксперта, связана с возможностью описания одного и того же признака по-разному. Например, если первый элемент буквы «А» исполнен в подписи вертикально, возможны варианты описания: «1-й элемент буквы «А» или «вертикальный элемент буквы «А».

Подписи традиционно подразделяются на буквенные, смешанные и безбуквенные, а «современная» подпись все более стремится к упрощению, как на качественном, так и на количественном уровнях. Буквенные элементы заменяются безбуквенными – отсюда и дефицит полезной информации. Здесь встает вопрос о расширении признакового поля применительно к безбуквенным элементам подписей.

Если при проведении почерковедческих исследований подписей по традиционной качественно-описательной методике указанные проблемы не столь очевидны, то при постановке вопроса об автоматизации подобных исследований они являются препятствием для осуществления данного процесса.

В Федеральном исследовательском центре «Информатика и управление» РАН и Московском Университете МВД РФ имени В.Я. Кикотя разрабатывается ДСМ-система, которая должна не только облегчить работу эксперта при решении этой задачи, но и, учитывая тот факт, что ДСМ-метод является методом автоматизированной поддержки научных исследований [12], дать возможность проводить такие исследования в области почерковедения.

ПОСТАНОВКА ЗАДАЧИ

Задача идентификации исполнителя подписи ставится следующим образом. Есть два набора подписей, выполненных от имени одного лица. Исполнитель одного из них известен, про второй из наборов подписей известно, что они выполнены не лицом, от имени которого значатся. Имеется также образец подписи, представленный на идентификацию. Требуется определить, является ли представленный образец подписью указанного исполнителя или он выполнен иным лицом. Сразу следует заметить, что с точки зрения реальной криминалистической деятельности эксперта такая постановка является искусственной. Целый набор подписей, заведомо выполненных не лицом, от имени которого они значатся, эксперт может получить только в каких-то особых случаях. Но эта постановка позволяет получить множество отрицательных примеров, наличие которых необходимо для работы ДСМ-системы. Поскольку работа ДСМ-систем позволяет не только выдвигать гипотезы и находить закономерности, но и проводить исследования в предметной области, некоторая искусственность постановки допустима, а полученные результаты исследований позволят помочь эксперту принимать более обоснованное решение и в реальной постановке задачи идентификации. Кроме того, система позволяет давать ответ и в случае наличия только подписей, выполненных известным лицом, используя усеченный вариант ДСМ-метода. Выполнение условий применения ДСМ-метода к рассматриваемой задаче было показано в [13].

Следует заметить, что при проведении почерковедческих исследований подписей по традиционной качественно-описательной методике эксперт старается выделить совокупность признаков, необходимых для идентификации и позволяющих сделать вывод в соответствии с используемой методикой. При этом разные эксперты выделяют разные наборы признаков. Особенностью использования ДСМ-системы является возможность учета максимального количества признаков, причем в процессе работы системы одни из них могут быть оценены как значимые, а другие – как мало значимые.

В связи с указанными выше проблемами, возникающими при выделении экспертом признаков подписи, а также в связи с необходимостью увеличения количества выделяемых признаков эксперту должна быть предоставлена возможность автоматизированного ввода признаков подписи. С этой целью к ДСМ-системе добавляется подсистема, позволяющая в автоматизированном режиме вводить признаки новой подписи, сохраняя при этом единообразие в описании признаков.

ОПИСАНИЕ ДАННЫХ

База фактов ДСМ-системы для решения задачи идентификации подписи состоит из множества объектов, множества свойств и отношения «объект обладает (не обладает) свойством», обозначаемого через $\Rightarrow_1$. Объектами являются подписи и их расшифровки (фамилия, инициалы), записанные в базе фактов в языке представления данных. Этот язык, разработанный криминалистами-почерковедами (см., например, [14]), был расширен добавлением некоторых признаков. Предполагается дальнейшее расширение этого языка. Подпись описывается на нескольких уровнях – уровне транскрипции, уровне общих признаков и уровне частных признаков. Для описания расшифровок используются общие и частные признаки почерка.

Под транскрипцией подписи традиционно понимается общее построение ее графического изображения, качественный и количественный состав. В транскрипции перечисляются последовательно все буквы и безбуквенные элементы.

Общие признаки подразделяются на четыре группы, характеризующие подпись в целом с разных сторон, но в разрабатываемой системе описание общих признаков происходит без привязки конкретного признака к определенной группе.

Частные признаки подписи раскрывают информацию об особенностях выполнения букв, элементов букв, безбуквенных элементов подписи и их соединений. Представлены они в системе на данном этапе десятью группами, характеризующими особенности и форму движений при выполнении букв и элементов. Предусмотрена возможность увеличения количества групп при необходимости.

Свойства – это фамилии исполнителей подписей, представленных в базе фактов. Поскольку в базе фактов присутствуют как подписи, выполненные определенным лицом от своего имени, так и подписи, от имени этого же лица, но выполненные дру-

гим лицом, база фактов состоит из примеров вида $J_{\langle \nu, 0\rangle}$ $(X_i \Rightarrow_1 Y_i)$, $\langle \nu, 0\rangle$ – истинностное значение, где $\nu \in \{1, -1, \tau\}$, 0 означает нулевой шаг вычислений. Если $\nu = 1$, объект (подпись) X_i выполнен субъектом Y_i; при $\nu = -1$ объект X_i выполнен не Y_i; $\nu = \tau$ означает, что неизвестно, выполнена ли подпись лицом, от имени которого значится, или нет.

Особенностью задачи является тот факт, что вопрос о выполнении подписи лицом, от имени которого она значится, решается каждый раз для одного конкретного человека, и, следовательно, информация о подписях других людей не используется. Поэтому реально база фактов разбивается на самостоятельные подбазы в количестве, равном числу исполнителей подписей и задача решается в каждой такой подбазе отдельно.

ПОДСИСТЕМА ВВОДА

Подсистема ввода создана для предоставления возможности эксперту вводить, удалять, редактировать описание подписей, используя при этом единообразие в обозначении признаков.

Поскольку при описании подписи в языке представления данных используются три уровня, в подсистеме, соответственно, присутствуют три вкладки: «Транскрипция», «Общие признаки» и «Частные признаки».

В данной версии системы количество объектов для одного свойства равно десяти (десять подписей, выполненных одним человеком от своего имени и десять подписей, выполненных от имени указанного лица иным лицом), но оно может меняться в случае необходимости.

В базе данных подсистемы ввода содержатся списки людей, чьи подписи будут исследоваться, всех букв и безбуквенных символов, наборов элементов каждой буквы, общих и частных признаков и множеств их значений (см. в разделе *Программная реализация*).

Процесс исследования подписи и выделения ее признаков, с одновременным занесением их в подсистему ввода, начинается с введения данных о транскрипции конкретной подписи. Для человека, подписи которого будут описываться с помощью языка представления данных, эксперт вводит все встретившиеся буквы и безбуквенные символы из соответствующего списка с указанием как места в подписи, так и связи со следующей буквой или безбуквенным символом. При описании транскрипции следующей подписи этого же человека, буквы и их признаки, уже выделенные в предыдущем образце подписи, не повторяются, а только добавляется единица к счетчику. В случае если в новом образце определенные буквы и безбуквенные элементы отсутствуют или выполнены по-другому, предусмотрено окно «Вариант», которое указывает на вариационность проявления указанной буквы (безбуквенного элемента) в данном множестве образцов подписей.

Для внесения общих и частных признаков подписей предусмотрены два пути. Первый используется при наличии признака в базе данных. В этом случае выбранный признак добавляется в описание подписи. Второй путь предполагает отсутствие признака в базе данных. Тогда эксперт сначала должен его ввести с соблюдением определенных правил, исключающих неверное описание.

Таким образом, происходит автоматизация процесса описания признаков подписи. За счет использования заранее установленных параметров и их конкретизации эксперту удается соблюсти однозначность описания признаков подписей и заметно сократить время работы. Под конкретизацией здесь понимается указание на определенную часть выбранного элемента выбранной буквы. Например, если при описании формы движений при выполнении 1-го элемента буквы «Б», требуется акцентировать внимание именно на его начальной части, эксперт открывает окно «Конкретизация» с имеющимся списком определенных локализаций элементов (начальная часть, нижняя часть, верхняя часть, левая точка, правая точка, верхняя точка, нижняя точка и другие).

Следует отметить, что в результате работы эксперта с подсистемой ввода формируется совокупное описание признаков подписи по всем десяти образцам с указанием для каждого элемента количества подписей, в которых он встретился.

ВАРИАНТ ДСМ-МЕТОДА ДЛЯ ИДЕНТИФИКАЦИОННОЙ ЗАДАЧИ

Как уже было указано, при решении задачи с помощью ДСМ-системы мы имеем дело с подбазой фактов, относящейся к одному лицу. Объекты – образцы подписей, записанные с помощью языка представления данных – это три множества: признаки транскрипции, общие и частные признаки. Учитывая, что почерку присущи, с одной стороны, как устойчивость, так и вариативность, а с другой – нет абсолютной уникальности, можно предположить, что не все признаки повторяются от образца к образцу, а в подписи, выполненной другим лицом, могут присутствовать и многие признаки подписей известного исполнителя, тем более, что другое лицо именно к этому и стремится. Отношение $(V \Rightarrow_2 W)$, читаемое как «результат V операции сходства объектов, является причиной проявления свойства W» в данном случае означает, что V характеризует почерк данного лица. Поэтому естественно предположить, что почерк характеризует вся совокупность признаков, встретившихся во всех образцах подписей, а не только общие для всех образцов признаки. А это значит, что в качестве операции сходства должна быть выбрана операция объединения, а не пересечения множеств. Если V – результат операции сходства не совпадает с $V1$ – результатом операции сходства на множестве объектов из отрицательных примеров (множествах признаков подписей, выполненных другим лицом), то V является положительной, а $V1$ отрицательной гипотезами. Если V и $V1$ совпадают, то гипотеза противоречива, и делать вывод по аналогии на ее основании нельзя.

Отношение вложения меняет при этом направление – объект вкладывается в гипотезу, а не наоборот. Следовательно, наиболее адекватным методом для решения идентификационной задачи является модифицированный ДСМ-метод. В этом случае предикат положительного сходства имеет вид:

$$\tilde{G}^{+}_{a,n}(V,W,k) \rightleftharpoons \exists X_1,...,\exists X_k \exists Y$$
$$\left(\&_{i=1}^{k} J_{(1,n)}(X_i \Rightarrow_1 Y) \,\&\, \forall i,j\left((i \neq j) \rightarrow (X_i \neq X_j)\right)\right) \&$$
$$\&\left(V = \bigcup_{i=1}^{k} X_i\right) \&$$
$$\&\forall X\left(\left(J_{(1,n)}(X \Rightarrow_1 Y) \,\&\, (Y \subseteq W)\right) \rightarrow \bigcup_{i=1}^{k}(X = X_i)\right) \&$$
$$\&(k \geq k_0)).$$

Правила правдоподобного вывода I-го рода такие же, как в классическом варианте ДСМ-метода. Но при проведении рассуждений по аналогии следует учитывать свойство вариативности почерка, а также тот факт, что значения признаков не являются абсолютно уникальными. Если в образце, предложенном на идентификацию, некоторый признак дан в варианте, не встретившимся ни в одной из подписей, попавших в базу фактов, полного вложения объекта в гипотезу не будет. Поэтому в рассуждении по аналогии используется квазивложение – мощность пересечения объекта с гипотезой. Вывод по аналогии делается на основании того, с какой из двух гипотез – положительной или отрицательной – мощность пересечения исследуемого объекта больше. Если эти мощности совпадают или разность их находится в пределах некоторого порога, обоснованный вывод сделан быть не может.

Таким образом, правила правдоподобного вывода II-го рода имеют вид:

$$\left(\left(J_{(1,n)}(V \Rightarrow_2 W) \,\&\, J_{(-1,n)}(V1 \Rightarrow_2 W) \,\&\right.\right.$$
$$\left.\left.\&\left(\left|X \cap V\right| - \left|X \cap V1\right| > \delta\right)\right) \rightarrow J_{(1,n+1)}(X \Rightarrow_1 W)\right).$$

В случае, когда $\left|X \cap V1\right| - \left|X \cap V\right| > \delta$, имеет место $J_{(-1,n+1)}(V1 \Rightarrow_2 W)$.

Если же $\left|X \cap V\right| - \left|X \cap V1\right| \leq \delta$, обоснованный вывод сделать нельзя. (Здесь δ – выбранный порог).

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Разрабатываемая система состоит из трех частей:

1. Подсистема ввода с базой данных.
2. ДСМ-решатель, реализующий модифицированный ДСМ-метод на данных из базы фактов.
3. Интерфейс, позволяющий пользователю работать с базой данных и базой фактов и представляющий результаты анализа данных в удобной для пользователя форме.

База данных и база фактов

База данных создана на основе MySQL версия 5.7.21. Она расположена на локальном сервере компьютера (в дальнейшем планируется перенос базы на удаленный сервер). На логическом уровне база данных подразделяется на базу данных подсистемы ввода и базу фактов. База данных подсистемы ввода содержит данные, необходимые для формирования базы фактов.

В качестве основных объектов в базе данных были выделены:

- *Человек*. Сведения об объектах этого типа хранятся в таблице *«Persons»*, которая содержит фамилию, имя, отчество и дату рождения, а таблица «Positions» – должность лиц, чьи подписи рассматриваются.
- *Алфавит* – список всех букв и безбуквенных символов с уникальным набором элементов (1-й элемент, 2-й элемент и.т.п), уточнениями для элементов (левая часть, правая часть и т.п.) и конкретизацией.
- *Общие признаки* – список всех общих признаков, возможных для подписи (наклон, нажим, степень выработанности и т.п.) с их значениями.
- *Частные признаки* – список частных признаков с их значениями и реализациями в буквах.

Объекты базы фактов:

- *Подпись*. Хранится в таблице *«Sets»*, содержащей описание совокупности образцов подписей, относящихся к конкретному человеку (идентификатор берется из таблицы *«Persons»*), либо одной подписи, представленной на идентификацию. Для каждой такой совокупности или подписи указывается, кому она принадлежит/приписывается, количество подписей, тип (подпись/расшифровка).
- *Транскрипция.*

Таблица *«Transcription»*, в которой хранятся объекты этого типа, содержит подробную транскрипцию для каждой подписи/совокупности подписей. Для каждой буквы транскрипции указывается, к какой подписи/совокупности подписей она относится, что это за буква, место этой буквы в транскрипции, связь со следующей буквой (интервальная, слитная, монограмма или последняя буква), и является ли она вариативной (характерно только для совокупности подписей). Информация о связи между буквами берется из словарной таблицы *«LinkNext»*, относящейся к базе данных.

Имеется также таблица, связывающая частный признак и конкретные буквы из транскрипции конкретной подписи и таблица, связывающая конкретное множество образцов подписей и конкретный общий признак. Кроме того, дается информация о том, в скольких образцах этот признак содержится (необходимо для совокупности подписей).

Работая с подсистемой ввода и базой данных, эксперт, по сути, формирует базу фактов, с которой будет работать ДСМ-система.

Программная реализация подсистемы ввода

Программная часть представлена в виде динамической библиотеки libhandwriting_data_access.so, написанной на языке C++. Для связи с базой данных и базой фактов MySQL используется библио-

тека MySQL Connector C++, версия 1.1.9 для Linux. Библиотека скомпилирована с помощью GCC, версия 5.4.0. Написана в среде разработки NetBeans IDE 8.2.

Из содержащихся в библиотеке классов основным является класс data_access, при применении методов которого пользователь обращается к базе данных MySQL, получая, добавляя, удаляя, изменяя и анализируя данные. Условно методы можно разбить на несколько групп, часть из которых относится к подсистеме ввода, а часть реализует работу ДСМ-решателя.

К подсистеме ввода относятся методы, позволяющие получать данные по их уникальному ключу, добавлять информацию, введенную пользователем, в соответствующие таблицы, удалять данные, выбранные с помощью интерфейса и все связанные с этими данными записи в других таблицах.

ДСМ-решатель

ДСМ-решатель должен обеспечить реализацию правил правдоподобного вывода I-го и II-го родов, что позволит найти гипотезы о причинах и гипотезы о принадлежности предъявленного на идентификацию образца подписи определенному лицу. Поскольку ДСМ-решатель реализует в данной системе модифицированный ДСМ-метод, а также с учетом постановки задачи идентификации, приведенной выше, для каждой подбазы фактов, в которой работает ДСМ-решатель, существует одна положительная и одна отрицательная гипотеза. Эти гипотезы есть результат операции объединения всех признаков всех образцов подписи, приписанных (для положительной гипотезы) или не приписанных (для отрицательной гипотезы) конкретному лицу.

Как было отмечено выше, работа эксперта по введению признаков образцов подписей с подсистемой ввода формирует базу фактов. Но способ, которым реализована эта работа, приводит к тому, что в процессе ввода осуществляется нахождение операции сходства (объединения). Т.е. в базе фактов содержатся не отдельные образцы подписей, записанные с помощью языка представления данных, а их объединение. Таким образом, остается реализовать квазивложение и вывод по аналогии. Это достигается за счет применения методов квазивложения данных. Рассматривая эти методы, следует учитывать, что гипотеза состоит из трех частей – объединенной транскрипции, а также объединенных множеств общих и частных признаков.

На вложимость проверяется один конкретный образец, помеченный в качестве тестового. Вся информация о подписях хранится как совокупность признаков, поэтому каждый из методов вложения принимает два аргумента: уникальный ключ совокупности подписей и уникальный ключ проверяемого образца. Каждую проверку на вложимость по трем пунктам: вложимость транскрипций, вложимость общих признаков и частных признаков осуществляет один из методов в библиотеке.

Все эти процедуры проделываются для совокупности признаков как положительных, так и отрицательных примеров. Затем подсчитывается мощность пересечения множества признаков проверяемого образца с множествами признаков положительной и отрицательной гипотез и находится разность этих мощностей. Вывод делается в соответствии с правилами правдоподобного вывода II –го рода.

Все данные получаются или изменяются с помощью запросов на языке SQL. Для выполнения запросов используются объекты библиотеки MySQL Connector C++, такие как Driver и Connection (для подключения к базе данных на сервере и сохранения этого подключения), Statement и Recordset (для выполнения запросов и сохранения результатов запросов).

ИНТЕРФЕЙС

Интерфейс системы позволяет пользователю работать с базой данных в подсистеме ввода, с базой фактов в ДСМ-системе и проводить анализ полученных результатов. Он реализован на фреймворке *Qt,* версия 5.8.0. К проекту подключается библиотека libhandwriting_data_access.so. Проект содержит пять форм, соответствующих пяти классам.

С помощью интерфейса пользователь осуществляет просмотр, добавление и удаление данных по всем объектам из базы данных и базы фактов, для чего в главном окне он выбирает одну из соответствующих форм.

Используя форму, реализующую ДСМ-решатель, пользователь указывает подбазу, с которой он будет работать, получает результаты квазивложения отдельно по всем трем частям объекта, а также вывод, сделанный системой.

После завершения работы начинается процесс анализа результатов – сравниваются признаки проверяемого образца подписи, вошедшие в положительную и отрицательную гипотезы, определяется, какие из них устойчивые, а какие вариативные, выделяются неинформативные признаки и т.п. На основании этого анализа могут корректироваться различные части системы – состав признаков, порог, решающие правила.

ЗАКЛЮЧЕНИЕ

Результаты применения разрабатываемой ДСМ-системы могут быть использованы как в исследовательской работе специалиста-почерковеда, так и для проведения почерковедческой экспертизы идентификации подписи. Причем работа с ДСМ-системой предполагает итеративный процесс – сначала эксперт анализирует результаты, полученные на первом этапе. На основе этого анализа в систему вносятся изменения и процесс повторяется.

Для большей обоснованности выводов планируется проверять устойчивость полученных результатов в последовательности расширяющихся баз фактов [15].

Создаваемая система предоставляет возможность расширения, которое предполагается использовать, в частности, для установления влияния психофизиологических характеристик личности на особенности выполнения подписи. Предложенные в работе методы могут использоваться и в других задачах с использованием криминалистических данных.

СПИСОК ЛИТЕРАТУРЫ

1. Аникин И.В., Анисимова Э.С. Распознавание рукописной подписи на основе нечеткой логики // Вестник Казанского государственного энергетического университета – Изд-во Казанского государственного энергетического университета. – 2016. – №3(31). – С. 48–64.
2. Дорошенко Т.Ю., Костюченко Е.Ю. Система аутентификации на основе динамики рукописной подписи // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2014. – №2(32). – С.219–223.
3. Jain A.K., Griess F.D., Connell S.D. Online signature verification // Pattern Recognition. – 2002. – № 35. – P. 2963–2972.
4. Ortega–Garsia J., Fierrez–Aquilar J., Martin–Rell J. Complete Signal Modeling and Score Normalization for Function-Based Dynamic Signature Verification // Audio and Video-Based Biometric Person Authentication. – 2003. – P. 658–667.
5. Смирнов А.В. Программа «ОКО-1» для исследования кратких и простых почерковых объектов // Теория и практика судебной экспертизы. – Вып. 1. – М.: ГУ РФЦСЭ, 2006. – 121 с.
6. Кулик С.Д., Никонец Д.А. Вопросы автоматизации почерковедческой экспертизы // Современные тенденции развития криминалистики и судебной экспертизы в России и Украине. Материалы международной научно-практической конференции в рамках проекта «Российско-украинские криминалистические чтения на Слобожанщине» – 25–26 марта 2011 г. (научное издание) в 2-х т. – Т. 1. – Белгород: БелГУ, 2011. – С. 69–73.
7. Гороховатский А.В. Верификация подписи на основе инвариантов преобразования Радона // Радиоэлектроника и информатика. – 2007. – Вып. 4. – С. 95–100.
8. Daramola S. Person Identification System using Static-dyamic Signatures Fusion // International Journal of Computer Science and Information Security. – 2010. – Vol. 8, № 6. – С. 88–92.
9. Sheng He, Lambert Schomaker. Writer identification using curvature-free features // Pattern Recognition. – 2017. – Vol. 63. – P. 451–464.
10. Гусакова С.М., Комаров А.С. Интеллектуальная система для решения идентификационной задачи в почерковедении // Искусственный интеллект и принятие решений. – 2010. – №4. – С. 49–54.
11. Гусакова С.М. Подход к решению задач атрибуции исторических источников с помощью ДСМ-метода // Автоматическое порождение гипотез в интеллектуальных системах / под ред. проф. В.К. Финна. – М.: Книжный дом «Либроком», 2009. – С. 494–501.
12. Финн В.К. Эпистемологические основания ДСМ-метода автоматического порождения гипотез. Часть I. // Научно-техническая информация. Сер. 2. – 2013. – №9. – С.1–29; Часть II. – 2013. – №12. – С. 1–26; Finn V.K. Epistemic foundations of the JSM method automatic hypothesis generation // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, №2. – P. 96–148.
13. Охлупина А.Н. К вопросу совершенствования модельных методов криминалистического исследования подписей. // Вестник экономической безопасности – 2016. – №6. – С. 104–108.
14. Устинов В.В. Модельные методы судебно-почерковедческого исследования: дис. … канд. юрид. наук. – М.: Московский ун-т МВД РФ им. В. Я. Кикотя, 2011.
15. Финн В.К. Обнаружение эмпирических закономерностей в последовательностях баз фактов посредством ДСМ-рассуждений // Научно-техническая информация. Сер. 2. – 2015. – № 8. – С. 1-29; Finn V.K. Detecting Empirical Regularities in Bases of Facts Using JSM Reasoning // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, № 4. – P. 122–151.

Материал поступил в редакцию 18.05.18.

Сведения об авторах

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

ЛАПШИНА Ирина Андреевна – студент 4-го курса Отделения интеллектуальных систем в гуманитарной сфере РГГУ, Москва
e-mail: i14lap61@mail.ru

ОХЛУПИНА Анастасия Николаевна – эксперт-криминалист ГУ МВД России по Московской области, адъюнкт Московского университета МВД РФ им. В.Я. Кикотя, Москва
e-mail: stasya.zharova@inbox.ru

УДК 002:004.6.032.2

П.В. Голубцов

Переход от априорной к апостериорной информации: байесовские процедуры в распределенных крупномасштабных системах обработки данных

Рассматривается процедура перехода от априорной к апостериорной информации для линейного эксперимента в контексте систем Больших Данных. Этот процесс носит, на первый взгляд, принципиально последовательный характер, а именно: в результате наблюдения, априорная информация трансформируется в апостериорную, которая впоследствии трактуется как априорная для следующего наблюдения, и т.д. Показано, что такая процедура может быть распараллелена и унифицирована за счет преобразования как результатов измерений, так и исходной априорной информации к некоторому специальному виду. Исследуются и сравниваются свойства различных форм представления информации. Рассматриваемый подход позволяет эффективно масштабировать процедуру байесовского оценивания и, таким образом, адаптировать ее к проблемам обработки больших объемов распределённых данных.

Ключевые слова: *Большие Данные, априорная и апостериорная информация линейное оценивание, каноническая информация, распределенные системы сбора и обработки данных, алгебра информации, информационное пространство*

ВВЕДЕНИЕ

Специфика обработки информации в системах Больших Данных, когда исходные данные собираются, хранятся распределённо и могут постоянно пополняться, ранее была рассмотрена в работе [1], где было показано, что для эффективной обработки распределенных данных ключевую роль играет использование специальной промежуточной формы представления информации, обладающей определенными алгебраическими свойствами. В [2] с подобных позиций исследовалась задача линейного оценивания в контексте распределенных систем сбора и обработки информации. Было показано, что на построенном информационном пространстве естественным образом порождается не только алгебраическая структура, описывающая композицию отдельных фрагментов информации, но и согласованное с ней отношение предпорядка, отражающее феномен качества информации.

В настоящей работе рассматривается проблема линейного оценивания с априорной информацией, тесно связанная с байесовским переходом от априорного распределения к апостериорному в математической статистике [3–5] и совпадающая с байесовским переходом для нормальных распределений. Эта задача представляет возможность исследовать и сравнивать различные формы представления информации, такие как: исходная «сырая» информация; максимально удобная для интерпретации «явная» информация; специальная «каноническая» информация, максимально удобная для промежуточных манипуляций с информацией (таких как, слияние, обновление, передача, хранение и т.п.). Далее мы покажем, что эти три способа представления информации приводят к информационным пространствам, обладающим определенными алгебраическими свойствами, а также исследуем свойства этих пространств и связи между ними. Основной интерес для нас представляют особенности этих пространств в контексте распределенной обработки больших объемов данных.

Заметим, что байесовская процедура последовательного обновления информации, считается одним из важнейших инструментов в экспертных системах [6–8]. Особый интерес к различным вариантам этой процедуры наблюдается в контексте Больших Данных [9, 10], поскольку она позволяет обновлять информацию об объекте исследования по мере поступления данных, в результате чего отпадает необходимость накопления и хранения самих исходных данных. Как будет показано ниже, выбор адекватного канонического информационного пространства позволяет существенно повысить эффективность процесса обработки данных за счет унификации и минимизации вычислений. Более того, в последнее время в проблематике Больших

Данных особое внимание уделяется методам анализа данных, допускающим параллельную и распределенную обработку [11, 12]. Далее мы увидим, что введение подходящей промежуточной формы представления информации открывает возможность гибкого распараллеливания и масштабирования процедуры обновления информации в распределенных системах обработки данных.

ЛИНЕЙНОЕ ОЦЕНИВАНИЕ С АПРИОРНОЙ ИНФОРМАЦИЕЙ

Кратко приведем постановку и решение задачи линейного оценивания с априорной информацией. Более детальное и общее рассмотрение можно найти в [13–15].

Линейное измерение

Рассмотрим схему линейного измерения вида [13, 14],

$$y = Ax + v \tag{1}$$

где $x \in \mathcal{D}$ – объект измерения – вектор евклидова пространства, $y \in \mathcal{R}$ – результат измерения, $\mathcal{R}$ – пространство результатов измерения, $A : \mathcal{D} \to \mathcal{R}$ – линейное отображение, описывающее искажения измерительной системы, и $v \in \mathcal{R}$ – случайный вектор шума с нулевым средним $\mathsf{E}v = 0$ и заданным ковариационным оператором $\mathsf{D}v = S : \mathcal{R} \to \mathcal{R}$.

Ковариационный оператор случайного вектора $\mu \in \mathcal{R}$ – многомерное обобщение понятия дисперсии. Он определяется как

$$(\mathsf{D}\mu)(z) = \mathsf{E}\langle \mu - \mathsf{E}\mu, z\rangle(\mu - \mathsf{E}\mu)\,[1]$$

для любого $z \in \mathcal{R}$. Ковариационный оператор случайного вектора μ – самосопряженный положительно полуопределенный оператор. Его матрица в ортонормированном базисе представляет собой ковариационную матрицу координат вектора μ в этом базисе.

Вся информация об измерении – это модель измерения, описываемая парой (A, S) и результат измерения y. Будем рассматривать здесь лишь измерения, в которых оператор S положительно определен, $S > 0$ и, следовательно, обратим. По сути это означает, что шум v возможен во всех направлениях, т.е. не существует собственного подпространства $\tilde{\mathcal{R}} \subset \mathcal{R}$ такого, что $v \in \tilde{\mathcal{R}}$ с вероятностью единица. Таким образом, исходные (или сырые) данные об измерении элемента $x \in \mathcal{D}$, как и в [2], представляются тройками вида (y, A, S), где $y \in \mathcal{R}$, $A : \mathcal{D} \to \mathcal{R}$, $S : \mathcal{R} \to \mathcal{R}$, $S > 0$, а $\mathcal{R}$ – некоторое линейное пространство. Множество всех таких троек вида (y, A, S) будем обозначать $\mathfrak{R}$. Ниже мы рассмотрим структуру этого пространства более подробно.

[1] Все рассматриваемые линейные пространства являются конечномерными евклидовыми со скалярным произведением $\langle \cdot, \cdot \rangle$

Кроме того, в отличие от [2], будем считать, что имеется априорная информация относительно объекта измерения x. В математической статистике априорная информация об x задается некоторым вероятностным распределением на пространстве $\mathcal{D}$ [3, 4], т.е. x рассматривается как случайный вектор, независимый с v. Предположим, что известны лишь некоторые свойства априорного распределения, а именно, его априорное среднее $\mathsf{E}x = x_0$ и априорный ковариационный оператор $\mathsf{D}x = F : \mathcal{D} \to \mathcal{D}$. Также будем считать, что оператор F положительно определен, $F > 0$, т.е. не существует собственного подпространства $\tilde{\mathcal{D}} \subset \mathcal{D}$ такого, что $x - x_0 \in \tilde{\mathcal{D}}$ с вероятностью единица. Множество всех таких пар (x_0, F) обозначим $\mathfrak{E}$, т.е.

$$\mathfrak{E} = \left\{ (x_0, F) \middle| x_0 \in \mathcal{D}, F : \mathcal{D} \to \mathcal{D}, F > 0 \right\}.$$

Оптимальное линейное оценивание

Задача линейного оценивания с априорной информацией о векторе x состоит в построении оценки $\hat{x}$ вида:

$$\hat{x} = Ry + r\ , \tag{2}$$

определяемого линейным отображением $R : \mathcal{R} \to \mathcal{D}$ и вектором сдвига $r \in \mathcal{D}$. При этом оценка $\hat{x} = Ry + r$ должна быть в среднем максимально близка к x. Формально рассмотрим среднюю погрешность оценки $H(R, r) = \mathsf{E}\|\hat{x} - x\|^2$. Из (1) и (2) следует, что

$$\|\hat{x} - x\|^2 = \|(RA - I)x + r + Rv\|^2 = \|(RA - I)x + r\|^2 + \\ + 2\langle (RA - I)x + r, Rv\rangle + \|Rv\|^2.$$

Усредняя это выражение по v, и учитывая, что $\mathsf{E}_v v = 0$, получаем:

$$\mathsf{E}_v \|\hat{x} - x\|^2 = \|(RA - I)x + r\|^2 + \mathsf{E}_v \|Rv\|^2 = \\ = \|(RA - I)x + r\|^2 + \mathrm{tr} RSR^*.$$

В последнем равенстве мы воспользовались тем, что $\mathsf{D}(Rv) = RSR^*$ и $\mathsf{E}\|\mu\|^2 = \mathrm{tr}\mathsf{D}\mu$ для случайного вектора μ с нулевым средним.

Поскольку в выражении для $\mathsf{E}_v \|\hat{x} - x\|^2$ присутствует неизвестный вектор x, определим погрешность оценивания, обеспечиваемую парой (R, r), как усредненную по априорному распределению вектора x, погрешность $\mathsf{E}_v \|\hat{x} - x\|^2$, т.е.,

$$H(R, r) = \mathsf{E}_x \mathsf{E}_v \|\hat{x} - x\|^2.$$

Обозначая $\tilde{x} = x - x_0$ и учитывая, что $\mathsf{E}_x \tilde{x} = 0$, получаем

$$H(R,r) = \mathsf{E}_x \left\| (RA - I)(\tilde{x} - x_0) + r \right\|^2 + \mathrm{tr} RSR^* =$$
$$= \mathsf{E}_x \left\| (RA - I)\tilde{x} \right\|^2 + 2\mathsf{E}_x \left\langle (RA - I)\tilde{x}, (RA - I)x_0 + r \right\rangle +$$
$$+ \left\| (RA - I)x_0 + r \right\|^2 + \mathrm{tr} RSR^* = \mathrm{tr}(RA - I)F(RA - I)^* +$$
$$+ \left\| (RA - I)x_0 + r \right\|^2 + \mathrm{tr} RSR^*.$$

Итак, задача линейного оценивания с априорной информацией состоит в построении таких R и r, при которых средняя погрешность оценивания $H(R,r)$ минимальна:

$$\min_{R,r} H(R,r) = \min_{R,r} \left(\mathrm{tr}(RA - I)F(RA - I)^* + \right.$$
$$\left. + \left\| (RA - I)x_0 + r \right\|^2 + \mathrm{tr} RSR^* \right). \qquad (3)$$

Легко видеть, что r входит только во второе слагаемое $\left\| (RA - I)x_0 + r \right\|^2$, которое всегда неотрицательно и обращается в 0 только при

$$r = (I - RA)x_0 \ . \qquad (4)$$

Поэтому можно исключить r из задачи минимизации (3) и свести ее к проблеме минимизации только относительно R:

$$\min_R H(R) = \min_R \mathrm{tr}\left((RA - I)F(RA - I)^* + RSR^* \right).$$

Пусть $\mathbb{S}_{\mathcal{D}}$ – пространство всех самосопряженных операторов на $\mathcal{D}$. Определим частичный порядок на $\mathbb{S}_{\mathcal{D}}$ следующим образом:

$$P \geq Q \Leftrightarrow P - Q \geq 0 \ .$$

Заметим, что tr является строго монотонным отображением из пространства самосопряженных операторов в вещественную прямую, а именно, если $P \geq Q$ то $\mathrm{tr}P \geq \mathrm{tr}Q$, а если, кроме того $P \neq Q$ то $\mathrm{tr}P > \mathrm{tr}Q$. Поэтому, рассмотрим задачу минимизации оператора $Q = (RA - I)F(RA - I)^* + RSR^*$. Фактически, оператор Q, который можно определить как $Q(z) = \mathsf{E}\langle \hat{x} - x, z \rangle (\hat{x} - x)$ для $z \in \mathcal{D}$, описывает корреляционные свойства погрешности оценивания $\hat{x} - x$, поэтому будем называть его оператором погрешности оценивания. Докажем, что существует единственное линейное отображение R, доставляющее минимум оператору Q.

Сначала преобразуем Q к виду, в котором явно выделены «квадратичные» и «линейные» относительно R члены:

$$Q = R\left(AFA^* + S\right)R^* + RAF + FA^*R^* + F =$$
$$= RCR^* + RD^* + DR^* + F \ ,$$

где $C = AFA^* + S > 0$ и $D = FA^*$.

Используя обратимость оператора C, и проводя процедуру, аналогичную «выделению полного квадрата» в выражении $RCR^* + RD^* + DR^*$, получаем:

$$Q = \left(R - DC^{-1}\right)C\left(R - DC^{-1}\right)^* + F - DC^{-1}D^* \ .$$

В этом выражении только первое слагаемое включает R, и для любого R оператор $\left(R - DC^{-1}\right)C\left(R - DC^{-1}\right)^* \geq 0$. Из невырожденности C сразу следует, что если $R - DC^{-1} \neq 0$, то и $\left(R - DC^{-1}\right)C\left(R - DC^{-1}\right)^* \neq 0$. Следовательно, Q и $H(R) = \mathrm{tr}Q$ достигают минимальных значений в единственной точке $R = DC^{-1}$. При этом оператор Q достигает минимального значения

$$Q = F - DC^{-1}D^* = F - RAF = (I - RA)F \qquad (5)$$

Отсюда следует, что оптимальное R определяется выражением

$$R = FA^*\left(AFA^* + S\right)^{-1} = \left(F^{-1} + A^*S^{-1}A\right)^{-1} A^*S^{-1}$$

Чтобы убедиться в справедливости последнего равенства, достаточно умножить обе части легко проверяемого тождества

$$\left(F^{-1} + A^*S^{-1}A\right)^{-1} FA^* = A^*S^{-1}\left(AFA^* + S\right)$$

на $\left(F^{-1} + A^*S^{-1}A\right)^{-1}$ слева и на $\left(AFA^* + S\right)^{-1}$ справа.

Несложно убедиться, что

$$I - RA = \left[I - \left(F^{-1} + A^*S^{-1}A\right)^{-1} A^*S^{-1}A \right] =$$
$$= \left(F^{-1} + A^*S^{-1}A\right)^{-1} F^{-1}$$

Используя это выражение в (4) и (5), получим явные выражения для оптимального вектора сдвига: $r = \left(F^{-1} + A^*S^{-1}A\right)^{-1} F^{-1}x_0$ и для минимального значения оператора погрешности:

$$Q = \left(F^{-1} + A^*S^{-1}A\right)^{-1}.$$

Таким образом, задача оптимального оценивания с априорной информацией о сигнале вида

$$\min_{R:\mathcal{R}\to\mathcal{D},r\in\mathcal{D}} \mathsf{E}\|Ry+r-x\|^2$$

имеет единственное решение

$$R=\left(F^{-1}+A^*S^{-1}A\right)^{-1}A^*S^{-1},$$
$$r=\left(F^{-1}+A^*S^{-1}A\right)^{-1}F^{-1}x_0$$

При этом оптимальная оценка вектора x

$$\hat{x}=Ry+r=\left(F^{-1}+A^*S^{-1}A\right)^{-1}\left(F^{-1}x_0+A^*S^{-1}y\right) \quad (6)$$

обладает наименьшим оператором погрешности оценивания

$$Q=RSR^*=\left(F^{-1}+A^*S^{-1}A\right)^{-1} \quad (7)$$

Отсюда, в частности, следует, что оценка $\hat{x}=Ry+r$ обладает минимальными погрешностями оценивания для каждой из координат x_j в некотором ортонормированном базисе:

$$\mathsf{E}\left(\hat{x}_j-x_j\right)^2=Q_{jj}=\left(\left(F^{-1}+A^*S^{-1}A\right)^{-1}\right)_{jj}$$

и $\mathsf{E}_v\|\hat{x}-x\|^2$ достигает минимального значения

$$\mathsf{E}\|\hat{x}-x\|^2=\mathrm{tr}Q=\mathrm{tr}\left(F^{-1}+A^*S^{-1}A\right)^{-1}.$$

Байесовское оценивание в случае нормальных распределений

Отметим, что в случае нормальных распределений, т.е. если распределение погрешности v, $P_v=N(0,S)$ и априорная информация $P_x=N(x_0,F)$, то условное распределение x при наблюдении y также нормально и $P_{x|y}=N(\hat{x},Q)$, где $\hat{x}$ и Q определяются формулами (6) и (7), см., напр., [3, 15, 16]. Таким образом, рассматриваемая процедура перехода от априорной информации к апостериорной полностью соответствует стандартному байесовскому переходу для нормальных распределений.

Исчезающая априорная информация

Рассмотрим предельный случай исчезающей априорной информации [14]. Для этого будем считать, что ковариационный оператор $F=F_\alpha$ стремится к ∞ «во всех направлениях» при $\alpha\to\infty$. А именно, пусть $F_\alpha\geq\alpha I$. Отсюда сразу следует, что $F_\alpha^{-1}\leq\frac{1}{\alpha}I$ и $F_\alpha^{-1}\to 0$. При этом для оценки вектора x и оператора погрешности оценивания Q получаем:

$$\lim_{\alpha\to\infty}\hat{x}_\alpha=\lim_{\alpha\to\infty}\left(F_\alpha^{-1}+A^*S^{-1}A\right)^{-1}\left(F_\alpha^{-1}x_0+A^*S^{-1}y\right)=$$
$$=\left(A^*S^{-1}A\right)^{-1}A^*S^{-1}y,$$
$$\lim_{\alpha\to\infty}Q_\alpha=\lim_{\alpha\to\infty}\left(F_\alpha^{-1}+A^*S^{-1}A\right)^{-1}=\left(A^*S^{-1}A\right)^{-1},$$

что отвечает оптимальному линейному оцениванию неизвестного вектора x, рассмотренному в [2].

Априорная информация как дополнительное измерение

Отметим, что априорная информация $(x_0,F)\in\mathfrak{R}$, описываемая средним x_0 и ковариационным оператором F может формально рассматриваться как дополнительное измерение [15]. Действительно, пусть в дополнение к измерению вида (1) производится независимое измерение, описываемое моделью (I,F), т.е.

$$x_0=x+\mu$$

сопровождающееся шумом $\mu\in\mathcal{D}$ с нулевым средним $\mathsf{E}\mu=0$ и заданным ковариационным оператором $\mathsf{D}\mu=F:\mathcal{D}\to\mathcal{D}$. Согласно [2] такая пара измерений может рассматриваться как одно измерение неизвестного вектора x:

$$\begin{pmatrix}x_0\\y\end{pmatrix}=\begin{pmatrix}I\\A\end{pmatrix}x+\begin{pmatrix}\mu\\v\end{pmatrix},$$

где $\begin{pmatrix}x_0\\y\end{pmatrix}\in\mathcal{D}\times\mathcal{R}$ – результат такого двойного измерения, $\begin{pmatrix}I\\A\end{pmatrix}:\mathcal{D}\to\mathcal{D}\times\mathcal{R}$ – линейное отображение, описывающее пару измерений, и $\begin{pmatrix}\mu\\v\end{pmatrix}\in\mathcal{D}\times\mathcal{R}$ – случайный вектор шума с ковариационным оператором $\mathsf{D}\begin{pmatrix}\mu\\v\end{pmatrix}=\begin{pmatrix}F&0\\0&S\end{pmatrix}:\mathcal{D}\times\mathcal{R}\to\mathcal{D}\times\mathcal{R}$.

Согласно [2] оптимальная линейная оценка неизвестного x и ее ковариационный оператор определяются выражениями:

$$\hat{x}=\left(F^{-1}+A^*S^{-1}A\right)^{-1}\left(F^{-1}x_0+A^*S^{-1}y\right),$$
$$\mathsf{D}\hat{x}=\left(F^{-1}+A^*S^{-1}A\right)^{-1}.$$

Эти формулы в точности совпадают с выражениями (6) и (7) для оптимальной оценки отвечающего ей оператора погрешности в задаче оценивания с априорной информацией. Таким образам, априорная информация о векторе x вида $(x_0,F)\in\mathfrak{E}$ может быть формально «заменена» дополнительным измерением, описываемым тройкой $(x_0,I,F)\in\mathfrak{R}$.

ПЕРЕХОД ОТ АПРИОРНРОЙ К АПОСТЕРИОРНОЙ ИНФОРМАЦИИ

Эксперимент с априорной информацией нередко трактуется (см. например, [4]) как процедура перехода от априорной информации к апостериорной. Более того, полученная апостериорная информация рассматривается как априорная по отношению следующему измерению [3, 10, 17]. Строго говоря, понятия априорной и апостериорной информации концептуально различны [18]. Поэтому, чтобы такой переход был обоснован, необходимо убедиться, что использование апостериорной информации в качестве априорной для следующего измерения приведет к тому же результату, что и использование исходной априорной информации для пары измерений.

Итак, пусть априорная информация о векторе x имеет вид $(x_0, F_0) \in \mathfrak{E}$. Согласно (6) и (7), измерение вида:

$$y_1 = A_1 x + v_1, \;\; \mathsf{D}v_1 = S_1 \,, \tag{8}$$

описываемое данными $(y_1, A_1, S_1) \in \mathfrak{R}$, обеспечит оценку:

$$x_1 = \left(F_0^{-1} + A^* S_1^{-1} A_1\right)^{-1} \left(F_0^{-1} x_0 + A_1^* S_1^{-1} y_1\right) \tag{9}$$

и оператор погрешности оценивания:

$$F_1 = \left(F_0^{-1} + A^* S_1^{-1} A_1\right)^{-1}. \tag{10}$$

Мы намеренно использовали обозначения x_1 и F_1 для обозначения оценки и ее оператора погрешности (вместо $\hat{x}$ и Q), чтобы подчеркнуть наше намерение использовать апостериорную информацию $(x_1, F_1) \in \mathfrak{E}$ в качестве априорной для следующего измерения:

$$y_2 = A_2 x + v_2, \;\; \mathsf{D}v_2 = S_2 \,. \tag{11}$$

Аналогично предыдущему шагу, получаем оценку x_2 и оператор погрешности F_2:

$$\begin{aligned} x_2 &= \left(F_1^{-1} + A_2^* S_2^{-1} A_2\right)^{-1} \left(F_1^{-1} x_1 + A_2^* S_2^{-1} y_2\right) = \\ &= \left(F_0^{-1} + A_1^* S_1^{-1} A_1 + A_2^* S_2^{-1} A_2\right)^{-1} \left(F_0^{-1} x_0 + \right. \\ &\left. + A_1^* S_1^{-1} y_1 + A_2^* S_2^{-1} y_2\right) \end{aligned} \tag{12}$$

и оператор погрешности оценивания:

$$\begin{aligned} F_2 &= \left(F_1^{-1} + A_2^* S_2^{-1} A_2\right)^{-1} = \\ &= \left(F_0^{-1} + A_1^* S_1^{-1} A_1 + A_2^* S_2^{-1} A_2\right)^{-1} \end{aligned} \tag{13}$$

Нетрудно убедиться, что задача оценивания с априорной информацией $(x_0, F_0) \in \mathfrak{E}$ и комбинированным измерением

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \;\; \mathsf{D}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$$

описываемым данными $\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}\right) \in \mathfrak{R}$ и представляющим пару независимых измерений (8) и (11), также приводит к результату оценивания, представленному выражениями (12) и (13). Это формально подтверждает правомерность использования апостериорной информации в качестве априорной для последующих измерений.

ПОСЛЕДОВАТЕЛЬНОЕ ОБНОВЛЕНИЕ ИНФОРМАЦИИ ДЛЯ СЕРИИ ИЗМЕРЕНИЙ

Рассмотрим серию независимых измерений:

$$y_i = A_i x + v_i, \;\; \mathsf{D}v_i = S_i, \;\; i = 1,...,n \tag{14}$$

Исходные данные, представляющие отдельное измерение, описываются тройкой $(y_i, A_i, S_i) \in \mathfrak{R}$. Рассмотрим схему последовательного «обновления» информации о векторе x, состоящую в переходе от априорной к апостериорной информации при поступлении очередного фрагмента данных (y_i, A_i, S_i).

Итак, пусть имеется исходная априорная информация $(x_0, F_0) \in \mathfrak{E}$. При поступлении первого измерения (y_1, A_1, S_1) оно «преобразует» априорную информацию (x_0, F_0), в апостериорную (x_1, F_1), согласно (9) и (10), которая теперь выступает в качестве априорной для второго измерения (y_2, A_2, S_2) и т.д. (рис. 1).

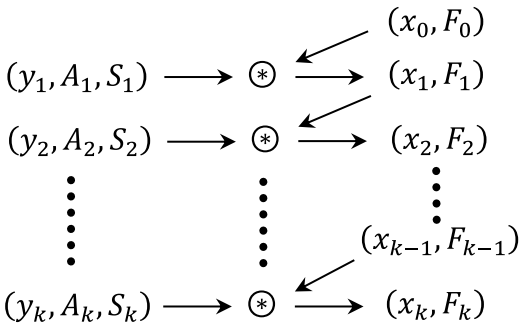

Рис. 1. Последовательное обновление информации

На k-м шаге добавление данных (y_k, A_k, S_k) к текущей, априорной для данного шага, информации

(x_{k-1}, F_{k-1}) и преобразовании ее в апостериорную (x_k, F_k) описывается как

$$(x_k, F_k) = (y_k, A_k, S_k) \circledast (x_{k-1}, F_{k-1}),$$

где

$$F_k = \left(F_{k-1}^{-1} + A_k^* S_k^{-1} A_k\right)^{-1},$$
$$x_k = F_k \left(F_{k-1}^{-1} x_{k-1} + A_k^* S_k^{-1} y_k\right)^{-1}.$$

Такая процедура последовательного обновления информации считается особенно важной в задачах обработки потоков Больших Данных (*Big Data streams*, [9, 10]) поскольку позволяет избежать накопления и хранения больших наборов данных.

Отметим, однако, что такое обновление «явной» информации о векторе *x* выглядит неоправданно трудоемко, поскольку на каждом шаге требует обращения линейных операторов. Кроме того, в этом процессе комбинируется информация, представленная двумя различными формами: явной, $(x_k, F_k) \in \mathfrak{E}$ и сырой $(y_k, A_k, S_k) \in \mathfrak{R}$. Формально, операция обновления $\circledast$ определена на $\mathfrak{R} \times \mathfrak{E}$, т.е., $\circledast : \mathfrak{R} \times \mathfrak{E} \to \mathfrak{E}$.

ПОСЛЕДОВАТЕЛЬНОЕ ОБНОВЛЕНИЕ ИНФОРМАЦИИ В ЯВНОЙ ФОРМЕ

Процесс обновления информации, описанный выше, можно сделать более однородным путем преобразования исходной информации в явную форму, перед добавлением ее к накопленной информации.

Выше мы видели, что если оператор $A_k^* S_k^{-1} A_k$ обратим, то сырую информацию $(y_k, A_k, S_k) \in \mathfrak{R}$, можно представить в явном виде $(x_k, F_k) \in \mathfrak{E}$, где

$$F_k = \left(A_k^* S_k^{-1} A_k\right)^{-1},\ x_k = F_k A_k^* S_k^{-1} y_k.$$

В данном случае x_k и F_k есть не что иное, как оптимальная линейная оценка [2] вектора *x* и ее ковариационный оператор, построенные на основании измерения (y_k, A_k, S_k) [2]

Если все исходные данные, полученные в результате измерений (14), допускают подобное представление, то перед добавлением к накопленной информации $(\bar{x}_{k-1}, \bar{F}_{k-1})$ сырой информации (y_k, A_k, S_k) преобразуем последнюю в явную форму (x_k, F_k). Модифицированная схема обновления информации представлена на рис. 2.

Здесь добавление информации в явной форме (x_k, F_k) к накопленной $(\bar{x}_{k-1}, \bar{F}_{k-1})$, представлено выражением

$$(\bar{x}_k, \bar{F}_k) = (\bar{x}_{k-1}, \bar{F}_{k-1}) \oplus (x_k, F_k)$$

где

$$\bar{F}_k = \left(\bar{F}_{k-1}^{-1} + \bar{F}_k^{-1}\right)^{-1},$$
$$\bar{x}_k = \bar{F}_k \left(\bar{F}_{k-1}^{-1} \bar{x}_{k-1} + F_{k-1}^{-1} x_k\right).$$

Отметим особенности такого подхода. Обновление базируется на операции композиции $\oplus$ двух элементов одного вида – информации в явной форме, т.е., $\oplus$ – бинарная операция на $\mathfrak{E}$. Использование явной формы в качестве основной представляется привлекательным, поскольку явная форма информации представляет собой оценку и ее погрешность, определяемые соответствующим набором данных. Однако, как и для предыдущей схемы, накопление информации сопровождается многократными обращениями линейных операторов. Кроме того, представление сырой информации (y_k, A_k, S_k) в явной форме возможно не всегда, а лишь тогда, когда $A_k^* S_k^{-1} A_k$ обатим, что существенно ограничивает применимость такого подхода. Фактически это означает, что явная форма не может быть использована как универсальная и эффективная форма представления информации.

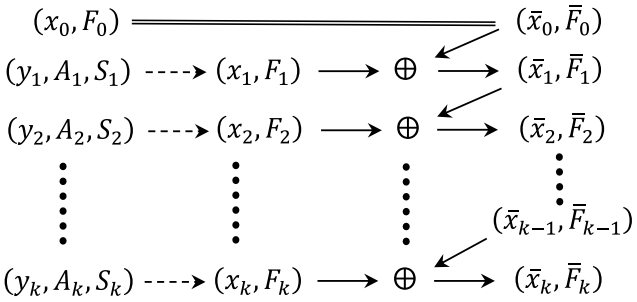

Рис. 2. Последовательное обновление информации с предварительным преобразованием сырой информации в явную форму (пунктирные стрелки означают, что соответствующее преобразование не всюду определено)

ПОСЛЕДОВАТЕЛЬНОЕ ОБНОВЛЕНИЕ ИНФОРМАЦИИ В КАНОНИЧЕСКОЙ ФОРМЕ

В работе [2] было показано, что удобной промежуточной формой представления результатов измерения $(y, A, S) \in \mathfrak{R}$ является каноническая форма $(u, T) = \left(A^* S^{-1} y, A^* S^{-1} A\right) \in \mathfrak{I}$. При выборе ее в качестве основной формы представления информации схема последовательного обновления информации принимает вид, представленный на рис. 3.

[2] Заметим, что здесь (x_k, F_k) обозначает «частичную» информацию, определяемую только *k*-м измерением, в отличие от предыдущего раздела, где так обозначалась «полная» накопленная после *k*-го измерения информация.

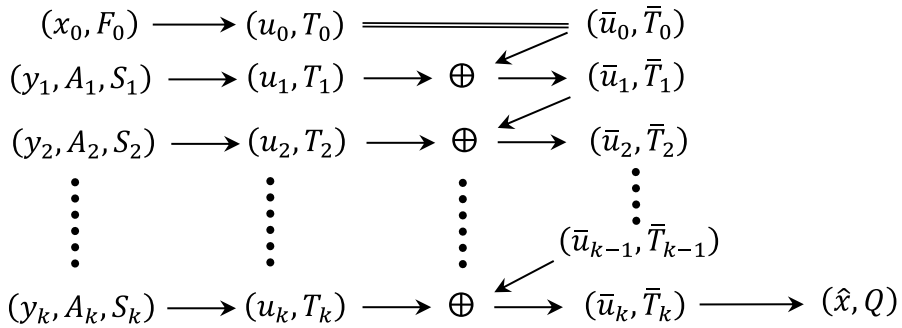

Рис. 3. Последовательное обновление канонической информации с предварительным преобразованием сырой информации в каноническую форму

Отметим что в канонической форме все шаги обработки данных максимально упрощаются. Действительно, преобразования, представленные на рис. 3, описываются следующими формулами:

Преобразование исходной априорной информации к каноническому виду:

$$(x_0, F_0) \to (u_0, T_0) = \left(F_0^{-1} x_0, F_0^{-1}\right);$$

Преобразование сырой информации к каноническому виду:

$$(y_k, A_k, S_k) \to (u_k, T_k) = \left(A_k^* S_k^{-1} y_k, A_k^* S_k^{-1} A_k\right);$$

Композиция фрагментов канонической информации:

$$(\overline{u}_k, \overline{T}_k) = (\overline{u}_{k-1}, \overline{T}_{k-1}) \oplus (u_k, T_k) = (\overline{u}_{k-1} + u_k, \overline{T}_{k-1} + T_k);$$

Построение результата оценивания из канонической информации:

$$(\overline{u}_k, \overline{T}_k) \to (\hat{x}, Q) = \left(\overline{T}_k^{-1}, \overline{T}_k^{-1} \overline{u}_k\right).$$

Полученная схема обновления информации, использующая в качестве основной формы информации каноническую, значительно превосходит предыдущие:

- все формы представления информации легко преобразуются в каноническую;
- каноническая информация (в отличие от явной) определена для любой сырой информации;
- композиция информации в канонической форме наиболее эффективна, поскольку описывается покомпонентной суммой пар вида (u,T), где $u \in \mathcal{D}$, $T : \mathcal{D} \to \mathcal{D}$,
- наиболее трудоемкая часть – построение оценки $\hat{x}$ и оператора погрешности Q может производиться лишь после того, когда накоплена каноническая информация на основе большого количества данных. При поступлении новых данных эта процедура может проводиться время от времени по мере необходимости получения обновленной оценки.

ИНФОРМАЦИОННЫЕ ПРОСТРАНСТВА

Выше мы видели, что в задаче линейного оценивания можно использовать разные формы представления информации:

- исходная *сырая* форма (y, A, S), которая описывает исходные данные;
- окончательная *явная* форма $(\hat{x}, Q)$, в которой представляется результат оценивания или априорная информация (x_0, F);
- промежуточная *каноническая* форма (u,T), максимально удобная для накопления информации.

Определим формально соответствующие информационные пространства $\mathfrak{R}$, $\mathfrak{E}$, и $\mathfrak{I}$. Для некоторого линейного пространства $\mathcal{R}$ будем обозначать $\mathbb{S}_{\mathcal{R}}$ – пространство самосопряженных операторов на пространстве $\mathcal{R}$, $\mathbb{S}_{\mathcal{R}}^{+} = \left\{S \in \mathbb{S}_{\mathcal{R}} \middle| S > 0\right\}$ – выпуклый конус положительно определенных операторов на $\mathcal{R}$, $\overline{\mathbb{S}}_{\mathcal{R}}^{+} = \left\{S \in \mathbb{S}_{\mathcal{R}} \middle| S \geq 0\right\}$ – замкнутый выпуклый конус положительно полуопределенных операторов на $\mathcal{R}$.

Каноническое информационное пространство

Пусть $\mathcal{D}$ – фиксированное пространство объекта измерения x. В [2] мы видели, что элементы канонической информации образуют хорошо организованную алгебраическую структуру – каноническое информационное пространство

$$\mathfrak{I} = \left\{(u,T) \middle| T \in \overline{\mathbb{S}}_{\mathcal{R}}^{+}, u \in \mathcal{R}(T)\right\}$$

с операцией композиции

$$(u_1, T_1) \oplus (u_2, T_2) = (u_1 + u_2, T_1 + T_2).$$

А именно, пусть $0 = (0,0) \in \mathfrak{I}$ (т.е. $u = 0 \in \mathcal{D}$ и $T = 0 : \mathcal{D} \to \mathcal{D}$) – элемент, определяющий «отсутствие информации». Тогда $(\mathfrak{I}, \oplus, 0)$ является *комму-*

тативным моноидом с сокращением, т.е., для всех $a,b,c \in \mathfrak{I}$ выполняются соотношения:

- $a \oplus b = b \oplus a$ – коммутативность,
- $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ – ассоциативность,
- $a \oplus 0 = a$ – свойство нейтрального элемента,
- $a \oplus b = a \oplus c \Rightarrow b = c$ – свойство сокращения.

Пусть $\mathfrak{I}^+$ – множество всех элементов (u,T) пространства $\mathfrak{I}$, для которых оператор $T > 0$ и, следовательно, обратим, т.е.,

$$\mathfrak{I}^+ = \{(u,T) \in \mathfrak{I} | T > 0\} = \\ = \{(u,T) | u \in \mathcal{D}, T \in \mathbb{S}^+_{\mathcal{R}}\} \subset \mathfrak{I}.$$

Очевидно, $\mathfrak{I}^+$ является *коммутативной подполугруппой* моноида $\mathfrak{I}$, но не подмоноидом, поскольку $\mathfrak{I}^+$ не содержит нейтральный элемент. Более того, $\mathfrak{I}^+$ является *идеалом* в $\mathfrak{I}$, т.е., если $a \in \mathfrak{I}^+$ и $b \in \mathfrak{I}$, то $a \oplus b \in \mathfrak{I}^+$.

Исходное информационное пространство

Аналогично все элементы сырой информации, т.е., тройки вида (y,A,S), можно рассматривать как элементы другой алгебраической структуры – исходного (сырого) информационного пространства $\mathfrak{R}$. Определим его формально. Пусть

$$\mathfrak{R}_{\mathcal{R}} = \{(y,A,S) | y \in \mathcal{R}, A : \mathcal{D} \to \mathcal{R}, S \in \mathbb{S}^+_{\mathcal{R}}\}$$

обозначает множество всех возможных измерений в пространстве $\mathcal{R}$. Определим исходное информационное пространство $\mathfrak{R}$ как объединение всех пространств $\mathfrak{R}_{\mathcal{R}}$, а именно, $\mathfrak{R} = \bigcup_{n=0}^{\infty} \mathfrak{R}_{\mathbb{R}^n}$.

Теперь мы можем формально определить операцию композиции на $\mathfrak{R}$ как

$$(y_1,A_1,S_1) \oplus (y_2,A_2,S_2) = \left\{ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \right\}.$$

Нетрудно убедиться, что исходное информационное пространство $(\mathfrak{R}, \oplus, 0)$ является (некоммутативным) *моноидом с сокращением:*

- $(a \oplus b) \oplus c = a \oplus (b \oplus c)$,
- $a \oplus 0 = a = 0 \oplus a$,
- $a \oplus b = a \oplus c \Rightarrow b = c \;\&\; b \oplus a = c \oplus a \Rightarrow b = c$.

Нейтральным элементом $0 \in \mathfrak{R}$ является тройка $0 = (0,0,0) \in \mathfrak{R}_{\mathbb{R}^0}$, т.е., $y = 0 \in \mathbb{R}^0$ – единственный элемент 0-мерного = пространства, $A = 0 : \mathcal{D} \to \mathbb{R}^0$ – единственное линейное отображение из $\mathcal{D}$ в 0-мерное пространство и $S = I : \mathbb{R}^0 \to \mathbb{R}^0$ – единственный линейный оператор в 0-мерном пространстве.

Пусть $\mathfrak{R}^+$ – множество всех элементов (y,A,S) пространства $\mathfrak{R}$, для которых оператор $A^* S^{-1} A > 0$, т.е.

$$\mathfrak{R}^+ = \{(y,A,S) \in \mathfrak{R} | A^* S^{-1} A > 0\} \subset \mathfrak{R}$$

Очевидно, $\mathfrak{R}^+$ является (некоммутативной) *подполугруппой* моноида $\mathfrak{R}$, но не подмоноидом, т.к. не содержит нейтральный элемент. Более того, $\mathfrak{R}^+$ является двусторонним *идеалом* в $\mathfrak{R}$, т.е., если $a \in \mathfrak{R}^+$ и $b \in \mathfrak{R}$, то $a \oplus b \in \mathfrak{R}^+$ и $b \oplus a \in \mathfrak{R}^+$.

Явное информационное пространство

Как мы видели выше, все элементы явной информации, т.е. пары вида (x_0,F), также можно рассматривать как элементы алгебраической структуры – явного информационного пространства:

$$\mathfrak{E} = \{(x,F) | x \in \mathcal{D}, F \in \mathbb{S}^+_{\mathcal{D}}\}$$

с операцией композиции:

$$(x_1,F_1) \oplus (x_2,F_2) = \\ = \left(\left(F_1^{-1} + F_2^{-1}\right)^{-1} \left(F_1^{-1} x_1 + F_2^{-1} x_2\right), \left(F_1^{-1} + F_2^{-1}\right)^{-1} \right).$$

Несложно видеть, что пространство $\mathfrak{E}$ не имеет нейтрального элемента и образует *коммутативную полугруппу с сокращением:*

- $(a \oplus b) \oplus c = a \oplus (b \oplus c)$,
- $a \oplus 0 = a$,
- $a \oplus b = a \oplus c \Rightarrow b = c$.

Сравнение информационных пространств

Вкратце обсудим достоинства и недостатки работы с информацией в этих формах.

Сырая информация:

1. Тривиальным образом представляет всю информацию, содержащуюся в исходных данных.

2. По мере поступления новых данных размер памяти, необходимый для их хранения, будет расти и потенциально неограничен.

3. Комбинирование информации в этой форме не требует специальных вычислений, однако, с ростом размера, временные затраты на организацию соответствующих массивов данных будут неограниченно расти.

4. Для вычисления окончательного результата оценивания необходимы значительные вычислительные ресурсы в связи с необходимостью производить вычисления с матрицами огромных размеров.

Явная информация:

1. Не всегда позволяет представить информацию, содержащуюся в исходных данных. Для возможности представления в явном виде необходима обратимость оператора $A^* S^{-1} A$.

2. Размер памяти не зависит от объема представленных данных ($\frac{m(m+3)}{2}$ чисел).

3. Комбинирование информации в этой форме предполагает многократное обращение матриц фиксированного размера $m \times m$ и умножения таких матриц на столбцы.

4. Определение окончательного результата не требует никаких вычислений, поскольку такое представление информации содержит результат оценивания в явной форме.

Каноническая информация:

1. Всегда может представить информацию, содержащуюся в исходных данных.

2. Размер памяти не зависит от объема представленных данных ($\frac{m(m+3)}{2}$ чисел).

3. Комбинирование информации в этой форме максимально упрощено и предполагает только сложения матриц фиксированного размера $m \times m$ и сложения m-мерных столбцов.

4. Вычисление окончательного результата требует умеренных вычислений (решения системы m уравнений с m неизвестными) и может выполняться лишь тогда, когда требуется получить результат оценивания по накопленной информации.

Таким образом, каноническая форма представления информации является самой универсальной и эффективной среди рассмотренных. Использование ее в качестве основной для проведения манипуляций с информацией позволяет повысить эффективность обработки данных.

Связь с достаточными статистиками и информационными матрицами

Заметим, что в случае нормальных распределений компоненты u и T канонической информации (u,T) имеют интересный теоретико-статистический смысл. Вектор u является *минимальной достаточной статистикой*, а оператор T представляет собой *информационную матрицу Фишера* [4, 5] для измерения y (и достаточной статистики u), см. напр., [15]. Как известно, матрица Фишера описывает количество (возможно, правильнее сказать, качество) информации, содержащейся в измерении. Таким образом, каноническая информация (u,T) в данном контексте представляет собой минимальную достаточную статистику плюс детальную характеризацию ее информативности.

РАБОТА С ИНФОРМАЦИЕЙ В РАЗЛИЧНЫХ ФОРМАХ

На разных стадиях обработки требуется осуществлять преобразования между различными формами представления информации. Рассмотрим эти преобразования более подробно (рис. 4).

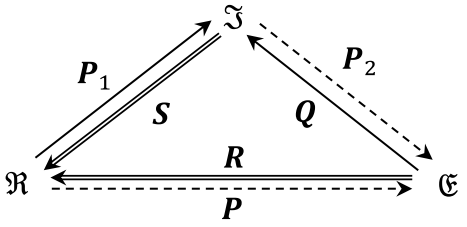

Рис. 4. Преобразования между разными формами информации (отображения, обозначенные пунктиром, определены не всюду, двойными линиями – определены не однозначно)

Преобразования одних форм в другие описывается следующими формулами:

1. *Преобразование* $\boldsymbol{P}:\mathfrak{R}\to\mathfrak{E}$ *сырой информации* (y,A,S) *в явную.* Фактически, реализует полное решение задачи оптимального оценивания [2], предоставляющее в качестве результата оптимальную линейную оценку $x_0=\hat{x}$ и оператор погрешности оценивания $F=\mathsf{D}\hat{x}$. Отображение $\boldsymbol{P}$ является частично определенным, точнее, оно определено лишь если оператор $A^*S^{-1}A$ обратим, т.е., на подполугруппе $\mathfrak{R}^+\subset\mathfrak{R}$:

$$\boldsymbol{P}:\mathfrak{R}^+\to\mathfrak{E},$$
$$\boldsymbol{P}:(y,A,S)\mapsto(x_0,F)=\left(\left(A^*S^{-1}A\right)^{-1}A^*S^{-1}y,\left(A^*S^{-1}A\right)^{-1}\right)$$

Нетрудно убедиться, что на своей области определения $\mathfrak{R}^+$ отображение P сохраняет алгебраическую структуры пространств, а именно, является частично определенным *гомоморфизмом полугрупп*, т.е.,

$$\forall d_1,d_2\in\mathfrak{R}^+\quad \boldsymbol{P}(d_1\oplus d_2)=\boldsymbol{P}(d_1)\oplus\boldsymbol{P}(d_2)$$

2. *Преобразование* $\boldsymbol{P}_\mathbf{I}:\mathfrak{R}\to\mathfrak{I}$ *сырой информации* (y,A,S) *в каноническую.* Определено всюду (мы считаем, что оператор S всегда обратим):

$$\boldsymbol{P}_\mathbf{I}:\mathfrak{R}\to\mathfrak{I},$$
$$\boldsymbol{P}_\mathbf{I}:(y,A,S)\mapsto(u,T)=\left(A^*S^{-1}y,A^*S^{-1}A\right)$$

и является (всюду определенным) *гомоморфизмом моноидов*, т.е.,

$$\boldsymbol{P}_\mathbf{I}(0_\mathfrak{R})=0_\mathfrak{I}$$

и

$$\forall d_1,d_2\in\mathfrak{R}\quad \boldsymbol{P}_\mathbf{I}(d_1\oplus d_2)=\boldsymbol{P}_\mathbf{I}(d_1)\oplus\boldsymbol{P}_\mathbf{I}(d_2)$$

3. *Преобразование* $\boldsymbol{P_2}:\mathfrak{I}\to\mathfrak{E}$ *канонической информации* (u,T) *в явную.* Определено лишь если оператор T обратим, т.е., на подполугруппе $\mathfrak{I}^+\subset\mathfrak{I}$:

$$\boldsymbol{P_2}:\mathfrak{I}^+\to\mathfrak{E},$$
$$\boldsymbol{P_2}:(u,T)\mapsto(x_0,F)=\left(T^{-1}u,T^{-1}\right).$$

и является *гомоморфизмом коммутативных полугрупп*, т.е.,

$$\forall c_1,c_2\in\mathfrak{I}^+\quad \boldsymbol{P_2}(c_1\oplus c_2)=\boldsymbol{P_2}(c_1)\oplus\boldsymbol{P_2}(c_2).$$

Отображения $\boldsymbol{P_1}$ и $\boldsymbol{P_2}$ обеспечивают факторизацию отображения $\boldsymbol{P}$, т.е., $\boldsymbol{P}=\boldsymbol{P_2}\circ\boldsymbol{P_1}$, позволяющую разбивать обработку данных (отображение $\boldsymbol{P}$) на две фазы: выделение канонической информации (отображение $\boldsymbol{P_1}$) и построение результата обработки на основе канонической информации (отображение $\boldsymbol{P_2}$).

4. *Преобразование* $\boldsymbol{Q}$ *явной информации* (x_0,F) *в каноническую.* Определено всюду (мы считаем, что оператор F всегда обратим):

$$\boldsymbol{Q}:\mathfrak{E}\to\mathfrak{I},$$
$$\boldsymbol{Q}:(x_0,F)\mapsto(u,T)=\left(F^{-1}x_0,F^{-1}\right).$$

Отображение $\boldsymbol{Q}$ позволяет представить любую явную (и в частности, априорную) информацию в каноническом виде. Нетрудно видеть, что отображения $\boldsymbol{Q}$ и $\boldsymbol{P_2}$ являются, в некотором смысле, взаимно обратными, точнее, они являются *изоморфизмами коммутативных полугрупп* $\mathfrak{E}$ и $\mathfrak{I}^+$, а именно, $\boldsymbol{P_2}\circ\boldsymbol{Q}=\boldsymbol{I}_{\mathfrak{E}}$ и $\boldsymbol{Q}\circ\boldsymbol{P_2}=\boldsymbol{I}_{\mathfrak{I}^+}$.

5. *Преобразование* $\boldsymbol{R}$ *явной информации* (x_0,F) *в сырую.* Заметим, что существует бесконечное множество элементов сырой информации (троек вида (y,A,S)), обеспечивающих результат оценивания (x_0,F). Наиболее очевидный из них $(x_0,I,F)\in\mathfrak{R}$. Определим отображение $\boldsymbol{R}$ как

$$\boldsymbol{R}:\mathfrak{E}\to\mathfrak{R},$$
$$\boldsymbol{R}:(x_0,F)\mapsto(y,A,S)=(x_0,I,F).$$

Отображение $\boldsymbol{R}$ позволяет представить любую явную (и в частности, априорную) информацию в виде результата некоторого гипотетического измерения. Нетрудно проверить, что $\boldsymbol{R}$ является правым обратным к, $\boldsymbol{P}$ т.е., $\boldsymbol{P}\circ\boldsymbol{R}=\boldsymbol{I}_{\mathfrak{E}}$, но не является гомоморфизмом полугрупп, т.к. формально не сохраняет операцию $\oplus$.

6. *Преобразование* $\boldsymbol{S}$ *канонической информации* (u,T) *в сырую.* Как и в предыдущем случае существует бесконечное множество элементов сырой информации, приводящих к канонической информации (u,T). Один из вариантов можно определить следующим образом:

$$\boldsymbol{S}:\mathfrak{I}\to\mathfrak{R},$$
$$\boldsymbol{S}:(u,T)\mapsto(y,A,S)=\left(u,P,T+(I-P)\right).$$

Здесь $P=TT^-$ – ортогональный проектор на $\mathcal{R}(T)$ – пространство значений оператора T [13, 19]. Отображение $\boldsymbol{S}$ позволяет представить каноническую информацию в виде результата некоторого гипотетического измерения, но, по-видимому, не представляет практической ценности. Подобно $\boldsymbol{R}$, отображение $\boldsymbol{S}$ является правым обратным к $\boldsymbol{P_1}$, т.е., $\boldsymbol{P_1}\circ\boldsymbol{R}=\boldsymbol{I}_{\mathfrak{E}}$, но не является гомоморфизмом моноидов.

ПАРАЛЛЕЛЬНАЯ РАСПРЕДЕЛЕННАЯ ОБРАБОТКА ДАННЫХ В ЗАДАЧЕ ЛИНЕЙНОГО ОЦЕНИВАНИЯ С АПРИОРНОЙ ИНФОРМАЦИЕЙ

Отображения информационных пространств, упомянутые выше, обеспечивают широкие возможности преобразования информации в процессе обработки. Наиболее универсальным и эффективным представляется преобразование как априорной информации, так и всех имеющихся данных в каноническую форму, комбинирование информации в этой форме и вычисление результата оценивания $(\hat{x},Q)$ на основании накопленной канонической информации (рис. 5).

Эта схема обладает всеми преимуществами предыдущей, представленной на рис. 3. Кроме того, она допускает существенное снижение времени обработки, поскольку преобразование различных фрагментов данных и априорной информации может производиться независимо и параллельно на различных компьютерах. Сложение фрагментов канонической информации требует минимальных вычислительных ресурсов. При появлении новых фрагментов данных необходимо преобразовать их в каноническую форму, добавить к накопленной канонической информации и пересчитать результат, используя обновленную каноническую информацию.

Одним из существенных недостатков Байесовского подхода считается сильная зависимость результата оценивания от априорного распределения. В результате этого ошибочная априорная информация может приводить к ошибочным результатам оценивания [20]. Однако наличие сократимости и коммутативности в пространствах $\mathfrak{I}$ и $\mathfrak{E}$ позволяет в любой момент «вычесть» исходную априорную информацию из накопленной и «заменить» ее на другую. Аналогично из накопленной информации можно «вычесть» любую предварительно включенную информацию, если впоследствии выяснится, что по тем или иным причинам соответствующее измерение было недостоверно.

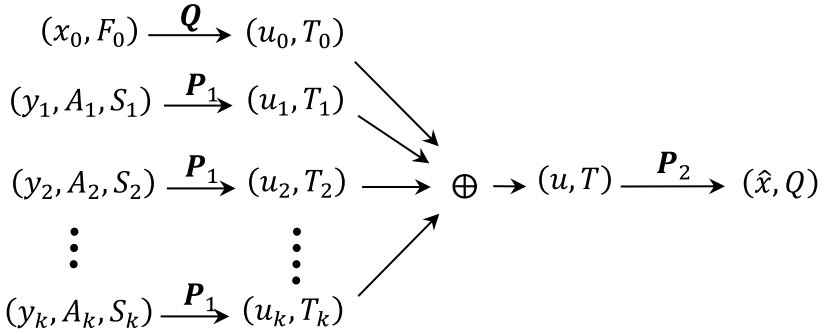

Рис. 5. Параллельная распределенная обработка данных и априорной информации

Особо отметим, что благодаря введению специальной промежуточной формы представления информации появляется возможность преобразовать последовательную по своей природе процедуру байесовского обновления информации в форму, допускающую высокую степень параллелизации и масштабирования. В результате этого процедура накопления информации органично «вписывается» в архитектуру систем распределенного хранения и анализа данных, таких как, например, *Hadoop MapReduce* [21–24] или *Spark* [25].

ЗАКЛЮЧЕНИЕ

Рассмотренная задача линейного оценивания с априорной информацией предоставляет целый спектр моделей информационных пространств с интересными соотношениями между ними. Два типа информационных пространств – исходное и явное – являются, в некотором смысле, «естественными» и, фактически, определяются самой постановкой задачи. Первое формализует пространство исходных данных, второе – априорные и апостериорные виды информации. Однако достаточно бедные свойства этих пространств ограничивают возможности оптимизации алгоритмов обработки данных, опирающихся только на такие формы представления информации. В связи с этим (особенно в контексте параллельной распределенной обработки данных) возникает потребность в построении некоторого «искусственного» информационного пространства, обладающего максимально богатой и универсальной структурой. В определенном смысле такая специальная форма представления информации отражает саму суть информации, содержащейся в данных.

Выше было показано, что использование канонической формы представления информации в качестве основной для манипуляций с информацией позволяет не только унифицировать процессы обработки данных, но и повысить их эффективность. Более того, благодаря богатым алгебраическим свойствам канонического информационного пространства традиционно последовательная процедура байесовского уточнения информации (перехода от априорной к апостериорной информации) допускает различные варианты распараллеливания. Это открывает возможности гибкого и эффективного масштабирования процедуры накопления информации в распределенных системах обработки данных.

СПИСОК ЛИТЕРАТУРЫ

1. Голубцов П.В. Понятие информации в контексте задач обработки Больших Данных // Научно-техническая информация. Сер. 2. – 2018. – № 1. – С. 31-36; Golubtsov P.V. The Concept of Information in Big Data Processing // Automatic Documentation and Mathematical Linguistics – 2018. – Vol. 52, №1. – P. 38–43.
2. Голубцов П.В. Задача линейного оценивания и информация в системах Больших Данных // Научно-техническая информация. Сер. 2. – 2018. – № 3. – С. 23-30; Golubtsov P.V. The Linear Estimation Problem and Information in Big-Data Systems // Automatic Documentation and Mathematical Linguistics – 2018. – Vol. 52, № 2. – P. 73–79.
3. Lindley D. Bayesian statistics: A review. – Philadelphia, PA: SIAM, 1972. – 89 p.
4. Барра Ж.-Р. Основные понятия математической статистики. – М.: Мир, 1974. – 280 с.
5. Боровков А.А. Математическая статистика. – Новосибирск: Наука, 1997. – 772 с.
6. Efron B. Bayes' theorem in the 21st century // Science. – 2013. – Vol. 340, № 6137. – P. 1177–1178.
7. Spiegelhalter D.J., Dawid A.P., Lauritzen S.L., Cowell R.G. Bayesian analysis in expert systems // Statistical Science. – 1993. – Vol. 8. – P. 219–247.
8. Spiegelhalter D.J., Lauritzen S.L. Sequential updating of conditional probabilities on directed graphical structures // Networks. – 1990. – Vol. 20, № 5. – P. 579–605.
9. Zhu J., Chen J., Hu W. Big learning with Bayesian methods // National Science Review. – 2017. – Vol. 4, № 4. – P. 627–651.
10. Oravecz Z., Huentelman M., Vandekerckhove J. Sequential Bayesian updating for Big Data. Chapter 2 of Big Data in Cognitive Science / ed. Michael N. Jones. – NY: Taylor & Francis, 2016. – P. 13–33.

11. Bekkerman R., Bilenko M., Langford J. Scaling up machine learning: Parallel and distributed approaches. – NY: Cambridge University Press, 2012. – 492 p.
12. Fan J., Han F., Liu H. Challenges of big data analysis // National Science Review. – 2013. – Vol. 1, № 2. – P. 293–314.
13. Пытьев Ю.П. Псевдообратный оператор. Свойства и применения // Мат. сборник. – 1982. – Т. 118, № 5, – С. 19–49; Pyt'ev Yu.P. Pseudoinverse operators. Properties and applications // Math. USSR Sb. – 1983. – Vol. 46, № 1. – P. 17–50.
14. Пытьев Ю.П. Математические методы интерпретации эксперимента. – М.: Высшая школа, 1989. – 351 с.
15. Пытьев Ю.П. Методы математического моделирования измерительно-вычислительных систем. – М.: Физматлит, 2012. – 428 с.
16. Lindley D.V., Smith A.F.M. Bayes Estimates for the Linear Model // Journal of the Royal Statistical Society. Series B. – 1972. – Vol. 34, № 1. – P. 1–41.
17. Robert C.P. On the Relevance of the Bayesian Approach to Statistics // Review of Economic Analysis. – 2010. – Vol. 2, № 2. – P. 139–152.
18. Vasudevan A. On the a priori and a posteriori assessment of probabilities // Journal of Applied Logic. – 2013. – Vol. 11, № 4. – P. 440–451.
19. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание – М.: Наука, 1977. – 224 с.
20. Little R.J. Calibrated Bayes: A Bayes/Frequentist Roadmap // The American Statistician. – 2006. – Vol. 60, № 3. – P. 213–223.
21. White T. Hadoop: The Definitive Guide. – Sebastopol, CA: O'Reilly, 2015. – 754 p.
22. Dean J., Ghemawat S. Mapreduce: simplified data processing on large clusters // Communications of the ACM. – 2008. – Vol. 51, № 1. – P. 107–113.
23. Palit I., Reddy C.K. Scalable and Parallel Boosting with MapReduce // IEEE Transactions on Knowledge and Data Engineering. – 2012. – Vol. 24, № 10. – P. 1904–1916.
24. Ekanayake J., Pallickara S., Fox G. MapReduce for Data Intensive Scientific Analyses // Fourth IEEE International Conference on eScience. – Indianapolis: IN, 2008. – P. 277–284.
25. Ryza S., Laserson U., Owen S., Wills J. Advanced Analytics with Spark: Patterns for Learning from Data at Scale. – Sebastopol, CA: O'Reilly, 2015. – 276 p.

Материал поступил в редакцию 10.05.18.

Сведения об авторе

ГОЛУБЦОВ Петр Викторович – доктор физико-математических наук, доцент, профессор Московского государственного университета им. М.В. Ломоносова; ведущий специалист ВИНИТИ РАН; профессор Национального исследовательского университета «Высшая школа экономики», Москва.
e-mail: golubtsov@physics.msu.ru

УДК 004.774–047.44:[001.891–051:002]

П.А. Калачихин

О разработке вебометрического критерия ранжирования исследователей*

Рассматривается возможность извлечения наукометрических данных из социальных сетей, перечисляются основные подходы к типизации и классификации альтметрик, анализируются преимущества и недостатки использования альтметрик в качестве наукометрических показателей и вводится понятие вебофикации библиометрических показателей. Приводятся примеры вебометрических показателей. Конструируются вебометрические индикаторы, предназначенные для оценки продуктивности работы исследователей. Предлагается критерий ранжирования исследователей по продуктивности их деятельности при помощи вебометрических индикаторов, вычисляемых на основании альтметрик.

Ключевые слова: *альтметрики, вебофикация, полувебометрические показатели, продуктивность, публикационная активность, социальные сети*

ВВЕДЕНИЕ

В последнее время бурный рост продемонстрировали методы мониторинга в сфере исследований и разработок, основывающиеся на анализе научной информации специализированных баз данных. Это обусловлено развитием технологий обработки текстовых электронных документов и web-ориентированных приложений. При развитии информационно-коммуникационных технологий продолжает формироваться новая парадигма отношений в научно-технической сфере.

Цель настоящего исследования – формирование основ для разработки комплексной системы наукометрических показателей, актуализированных с учетом современных тенденций развития научных коммуникаций. Задачи исследования включают анализ практики оценки и стимулирования публикационной активности, а также выявление наиболее распространенных недостатков существующих систем наукометрических показателей, допускающих искажение представлений о реальных достижениях как отдельных исследователей, так и образовательных организаций и научных сообществ.

Анализ показателей, определяющих недостатки сложившихся систем стимулирования публикационной активности, выполнен с использованием метода мониторинга разнородных электронных информационных ресурсов по научным коммуникациям. Особенности функционала социальных сетей фигурируют в качестве материала для проведения настоящего исследования.

СПОСОБЫ КЛАССИФИКАЦИИ АЛЬТМЕТРИК

Термин «альтметрика» (т.е. альтернативная метрика) охватывает широкий спектр различных web-платформ. Альтметрика является отражением любого события в некоторой форме, которая показывает уровень обязательств между исследователями и результатами исследований. Альтметрики принимают многообразные формы, и поэтому крайне важно учитывать, что не все альтметрики равнозначны и не всегда отражают одинаковый уровень обязательств. Огромное количество различных потенциальных источников наукометрических данных распределено по множеству платформ, прослеживающих альтметрики. Многочисленные способы классификации альтметрик связаны с неопределенностью понимания того, как различные альтметрики отражают результаты исследований, а также с с проблемой отсутствия стандартов для альтметрик [1].

В данном исследовании предлагается собственная классификация альтметрик с точки зрения интерактивных действий пользователей:

- лайки $@_{like}$ (от англ. *like* – «*то, что нравится*») могут быть бинарными оценками «*хорошо*» / «*плохо*» или множественными эмоциями «*интересно*», «*смешно*» и т.п.;
- загрузки $@_{load}$ файлов на локальный компьютер;
- просмотры $@_{view}$ электронной публикации;

* Работа выполнена в рамках исследования по теме 0003-2015-0008 Госзадания ВИНИТИ РАН

- ответы на вопросы других пользователей $@_{reply}$;
- новые темы в разделах форумов $@_{theme}$;
- упоминания в виде тэгов на специальных языках разметки $@_{\#}$;
- опрос $@_{select}$ по качественной или количественной, упорядоченной или неупорядоченной шкале.

Помимо этого существуют так называемые «технологические» альтметрики, которые относятся не к отдельным авторам, публикациям, журналам или организациями, а к социальным платформам для исследователей в целом. Технологические альтметрики подсчитывают: подписчиков или читателей $@_{CR}$; зарегистрированных пользователей $@_{CU}$; посещения $@_{CI}$; рейтинги в поисковой системе $@_{SSRP}$.

Срезы альтметрик удобно осуществлять по отдельным измерениям, например, по характеристикам, занесенным в профили авторов, являющихся пользователями наукометрических платформ, участниками социальных сетей для исследователей или представителями иных форм сетевых сообществ. Подобные срезы касаются именно пользователей, потому что для других объектов достаточно использования библиометрических индикаторов. Следует также учитывать ограничения, которые влияют на результат с понижающими коэффициентами: информационные ресурсы с платным доступом, с необходимостью регистрации, с локальными ограничениями доступа и др.

Альтметрики отличаются динамичными изменениями в течение коротких промежутков времени, поэтому они не используются в календарных отчетах. Альтметрики приходится отслеживать регулярно. В связи с этим мониторинг должен происходить с высокой частотой. Распространено «накручивание» альтметрик, в том числе при помощи программного обеспечения типа *adware* или «*ботов*» – программных агентов с активным поведением [2]. Обычно альтметрики могут быть узнаны только через интерактивный диалог, тем не менее, разработчиками создан программный интерфейс, выполняющий запросы для получения сведений об альтметриках.

ВОЗМОЖНОСТЬ ИЗВЛЕЧЕНИЯ НАУКОМЕТРИЧЕСКИХ ДАННЫХ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

По сравнению с обычными социальными сетями академические сайты социальных сетей (*Academic Social Networking Sites – ASNS*) предлагают более специфичные функции [3]. Научные публикации достаточно полно представлены на платформах социального общения, что делает их более значимым источником для оценки влияния публикаций по сравнению с наукометрическими базами данных. Несмотря на то, что использование в научной деятельности средств социального общения остается незначительным, пользователи постепенно узнают о потенциале альтметрик, извлекаемых из социальных платформ [4].

Зависимость альтернативных метрик от базовых платформ социального общения не может быть установлена однозначно. Эти платформы и их возможности допускают интерактивные действия при выполнении задач, для которых созданы. Большинство пользовательских действий не могут выполняться за пределами определенной платформы, которая оснащается множеством конкретных и зависящих от базового инструмента индикаторов. В то время, как первая волна оцифровки научных коммуникаций, в которую попала электронная почта, а также электронные журналы, выражалась в возможностях оперативного проведения обсуждений научным сообществом, вторая волна оцифровки включила использование инструментов, допускающих действительно широкое обсуждение вне научного сообщества. Но присутствие одних только платформ не гарантирует широкого охвата. Средства социального общения скорее открывают новые каналы неофициальных обсуждений среди исследователей, а не «строят мосты» между научным сообществом и обществом в целом. Активизация использования средств социального общения требует тщательного согласования между инвесторами, научными учреждениями и управленцами.

Альтметрики служат индикаторами для измерения следующих категорий:

- внимание, обязательства или влияние, а также социальное воздействие исследований на различную аудиторию;
- взаимодействие, контекст и сети;
- значимость количественных оценок и уникальности исследований.

Альтметрики характеризуются следующими особенностями:

- быстро накапливаются после выхода публикаций с результатами исследований;
- обладают более разнообразным набором для измерения различных типов влияния по сравнению с обычным цитированием;
- не ограничены научными публикациями и формальными цитированиями как форматами изложения результатов исследований;
- демонстрируют контекст, в котором исследование имеет некоторое влияние или воздействие на аудиторию;
- используют открытые данные; поэтому альтметрики проще тиражировать, чем сведения из баз данных.

Время покажет, являются ли средства социального общения и альтметрики побочным явлением научной среды или они станут центральными элементами методов распространения и оценки исследований. Возможно исчезновение некоторых индикаторов и платформ из-за недостатка практической значимости и уместности, в то время как возможно дальнейшее совместное использование. Исход в обоих случаях будет одинаковым из-за завершения сроков действия платформ и сервисов, на которых они базируются [5].

Альтметрические данные предназначены для отслеживания научных исследований на множестве web-платформ, включая новостные сайты, платформы социального общения, блоги и инструменты управления ссылками. Существует очень сложный контекст для систематического и исчерпывающего анализа альтметрик посредством оценки средств социального общения. Если предположить, что корреляция альтметрик и цитирований действительно име-

ет место быть, возникает следующий вопрос: как генераторы знаний и пользователи смогут извлечь из этого выгоду. Измерение связи между количеством цитирований и альтметрическими факторами требует построения регрессионной модели. Регрессионную модель следует дополнить анализом корреляции между традиционными, базирующимися на библиометрии, и вебометрическими индикаторами [6].

ПЛЮСЫ И МИНУСЫ ИСПОЛЬЗОВАНИЯ АЛЬТМЕТРИК В КАЧЕСТВЕ НАУКОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ

В современных исследованиях наукометрических показателей приоритетны две темы:

1) анализ разработок, основанных на производительности систем финансирования и того, как они включают и ограничивают методы оценки ответственности;

2) анализ практик использования метрик оценки производительности на разных уровнях исследовательской системы [7].

Альметрики и вебометрики применяются, чтобы подчеркнуть традиционный технологический аспект, которым обладают изначально различные особенности метрик [8]. C. Hoffmann, C. Lutz и M. Meckel указывают на различие между вебометриками и альтметриками, хотя и те, и другие, как правило, имеют числовую форму [9].

Не все исследователи используют платформы социального общения, поэтому измерение влияния научных публикаций на общество и экономику всегда касается только определенной выборки людей, упоминающих публикацию более или менее часто. Поскольку нет никакой точной пользовательской статистики или демонстрационных описаний для отдельных платформ социального общения, этот полезный вклад не подлежит количественной оценке. Количественные значения альтметрик часто делаются видимыми как количество соответствующих упоминаний в информационной среде социальной платформы. Величина объема информации о группах пользователей, имеющих отношение к научной публикации, важна для измерения социального влияния. Описания достигнутого социального влияния обычно сильно не хватает на практике. Публикации часто существуют в различных версиях, например, таких, как предварительная печать и постпечать от издателя. Цитирования соответствуют простым упоминаниям или обсуждениям процитированных публикаций. Значения альтметрик также возрастают, когда альтметрики относятся к переговорам через средства социального общения. Каждый исследователь точно знает, что измеряется количеством цитирований, например, сколько раз публикация процитирована. В альтметриках этот показатель часто является нечетким, измеряющим ту же самую величину.

Выделим ряд преимуществ, которыми альтметрики обладают по сравнению с традиционными метриками:

- широта – альтметрики измеряют влияние не только внутри, но и за рамками «нормальной» науки;
- разнообразие – альтметрики измеряют эффект влияния от разных видов результатов интеллектуальной деятельности;
- быстрота – альтметрики позволяют измерить влияние сразу после выхода публикации или окончания работы над результатами;
- доступность – как правило, альтметрические данные относительно легко получить [10].

Существует масса людей, регулярно общающихся через web-порталы. Электронные письма постоянно привлекают внимание одних пользователей к другим потенциально интересным пользователям. Отсутствие коммуникаций или нежелание пользователей связываться между собой может привести к тому, что web-портал перестанет приносить коммерческую выгоду. До сих пор нет ясного понимания, какой вклад в продвижение коммуникаций вносят альтметрики. Оценка экономического эффекта данного вида деятельности не может быть сопоставлена ни с одной традиционной библиометрической метрикой. Дело в том, что цитирование друг друга у исследователей не поощряется. Значение альтметрик со временем будет возрастать, если принимать во внимание, что платформы могут тиражироваться. Поэтому необходимо прикладывать дополнительные усилия для нормализации значений альтметрик.

Альтметрики не только предлагают возможности, но и создают трудности. Основная возможность, обеспечиваемая альтметриками, их разнообразием и неоднородностью, представляет основную проблему. Альтметрики включают различные типы метрик, многообразие которых мешает установить ясное понимание того, что метрики из себя представляют. Проблемы, связанные с их неоднородностью и неосмысленностью, вызваны отсутствием концептуальной основы, а также множественностью платформ социального общения, пользователей и мотивов поведения.

Точность, непротиворечивость и воспроизводимость данных характеризуют основные признаки качества. Качество данных обычно не выдвигается на первое место, особенно в контексте оценки исследований. В альтметриках качество данных является главной проблемой, которая существеннее ошибок и статистических погрешностей. В контексте цитирования ошибки, главным образом, представляют несоответствие между действиями и зарегистрированными событиями. Ошибки могут быть обнаружены и измерены посредством обращения к различным агрегаторам данных или к первоисточникам. В то время, как библиометрические источники являются статическими документами, большинство источников данных в контексте альтметрик динамичны, так как следы альтметрик подвержены изменениям или полному удалению. Еще более сложной, чем зависимость от агрегаторов, является зависимость альтметрик от платформ социального общения как провайдеров данных. Сильная зависимость от платформ социального общения достигает кульминации, когда характер платформ непосредственно влияет на поведение пользователей, и технологические особенности платформ определяют фактические действия пользователей [11].

ПРЕОБРАЗОВАНИЕ БИБЛИОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ В АЛЬТМЕТРИКИ

Цитирование является стандартным источником данных в наукометрии и ориентированной на конкретные задачи метрикой, которая измеряет научное влияние публикаций. Индикатор воздействия на читателей может отображать способность журналов, государств и академических учреждений выпускать печатные работы, которые находятся ниже или выше среднего воздействия публикаций на выбранный сегмент общества [12].

С одной стороны, открытость данных из платформ социального общения обеспечивает легкодоступный источник для статистических анализов, который охотно принимается научным сообществом. Значимые данные о влиянии большого набора публикаций обычно не так просто получить в библиометрии. С другой стороны, ответвление этого нового источника данных соответствует желанию измерить широкий эффект от научных исследований.

Чем больше средств социального общения оказываются в распоряжении людей, фокусирующихся на исследованиях, тем выше корреляция между соответствующими альтметриками и традиционным цитированием. Результаты частых исследований, измеряющих корреляции между альтметриками и традиционным цитированием, рассматриваются как первый шаг в сторону проведения исследований в области альтметрик. Альтметрики обеспечивают удобную возможность фокусировать внимание на более прибыльных проектах. Низкие корреляции указывают на альтметрики, которые особенно интересны для широкого измерения эффекта от исследований, т. е. влияния на другие сферы общественной жизни, нежели наука. Будущие исследования должны фокусироваться на этих альтметриках, чтобы оценить их потенциал и широту воздействия [13]. Поэтому преждевременно применять альтметрики в фундаментальных задачах по оценке исследований [14].

Для обозначения преобразования библиометрических индикаторов в альтметрике будем использовать термин «вебофикация». Вебофикация w возможна потому, что между альтметриками $\mathcal{A}$ и библиометрическими показателями $\mathcal{B}$ прослеживается некоторое соответствие, пусть и не всегда четкое, как это видно на рис. 1.

Переход с обычных метрик на альтметрики представляется если не сомнительной, то, по крайней мере, рискованной инициативой. Однако вебофикацию возможно осуществлять при помощи специальной таблицы.

Таблица содержит далеко не полный список соответствий, поэтому на представленных данных не видно, что вебофикация, вообще говоря, не является строгим преобразованием. Неточность вебофикации записывается как $@m \cong w(m)$. Особенности преобразования вебофикации связаны с тем, что возможны случаи, когда преобразование является:

неоднозначным:

$$m \xrightarrow{w} \{@m_x, @m_y\}, \quad (1)$$

где m – библиометрический показатель; $@m_x$ – x-ая альтметрика; $@m_y$ – y-ая альтметрика; w – преобразование вебофикации;

– *неинъективным*:

$$\{m_x, m_y\} \xrightarrow{w} @m, \quad (2)$$

где m_x – x-й библиометрический показатель; m_y – y-й библиометрический показатель; $@m$ – альтметрика; w – преобразование вебофикации;

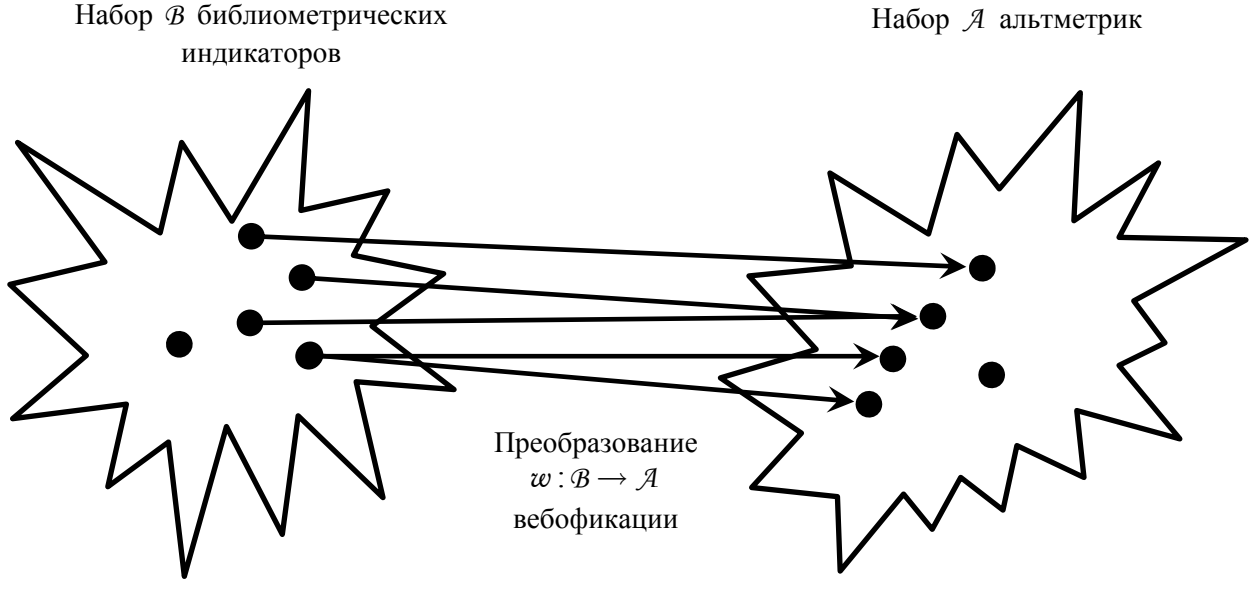

Рис. 1. Диаграмма преобразования библиометрических показателей в альтметрики

Преобразование библиометрических показателей в альтметрики табличным методом

№	Объекты оценки	Библиометрические показатели $\mathcal{B}$	Вебометрические индикаторы (с альтметриками $\mathcal{A}$)
1	Автор	Количество публикаций (*CP*)	Количество размещенных электронных текстов ($@_{CP}$)
2	Автор	Количество грантов (*CG*)	Количество коллективных, открытых, сетевых проектов ($@_{CG}$)
3	Публикация	Количество упоминаний в списках литературы (*CUR*)	Количество загрузок файлов в текстовых форматах ($@_{CUR}$)
4	Автор, публикация	Количество цитирований (*TCC*)	Количество переходов по ссылкам на электронные тексты работ ($@_{TCC}$)
5	Автор, публикация	Количество соавторов (*CCA*)	Количество подписчиков ($@_{CCA}$)
6	Автор, публикация, журнал, организация	Импакт-факторы и др. композитные индикаторы (*CCI*)	Рейтинги, привязанные к сетевым платформам ($@_{CCI}$)

несюръективным:

$$@m \xrightarrow{w^{-1}} \{\varnothing\}, \qquad (3)$$

где $@m$ – альтметрика; $\varnothing$ – пустое множество библиометрических показателей; w^{-1} – преобразование, обратное преобразованию вебофикации;

не везде определенным:

$$m \xrightarrow{w} \{\varnothing\}, \qquad (4)$$

где m – библиометрический показатель; $\varnothing$ – пустое множество альтметрик; w – преобразование вебофикации.

Таким образом, из формул (1) – (4) вытекает, что преобразование вебофикации принимает вид отношения, но не функции.

ПРИМЕРЫ ВЕБОМЕТРИЧЕСКИХ ИНДИКАТОРОВ

Альтметрики не дают полного устранения любых недостатков, содержащихся в традиционных метриках оценки воздействия. Однако альтметрики действительно позволяют производить оценку непосредственно на уровне продуктов, а не публикаций. Кроме того, они покрывают растущее разнообразие в академических продуктах, платформах и персонале. Общая неопределенность в наукометрии и за пределами наукометрии, связанная со значениями альтметрик, обусловлена тенденцией использования составных индикаторов без целевых ограничений относительно того, как альтметрики могут использоваться [15].

Относительная цитируемость разнородного потока публикаций является известным библиометрическим индикатором в оценке исследовательских работ. Цели индикатора заключаются в том, чтобы нормализовать цитирование, а также значащие различия среди областей. Критическое исследование теоретического основания механизма нормализации, примененного в индикаторе относительной цитируемости разнородного потока публикаций, создает предпосылки для использования альтернативного механизма нормализации. В результате возможно создание нового индикатора относительной цитируемости потока публикаций, который будет полагаться на альтернативный механизм нормализации [16].

Чтобы измерить социальное влияние научных публикаций, предлагается использовать альт-индекс, аналогичный *h*-индексу. Альт-индекс является допустимой альтметрикой для оценки исследований и способен устранить некоторый разрыв. Альт-индекс вычисляется с использованием той же формулы, которая используется для *h*-индекса. Единственная разница между *h*-индексом и альт-индексом заключается в том, что он основывается на социальных, а не академических значениях цитирования. Таким образом, формула для альт-индекса выглядит так: «*если у x количество публикаций будет, по крайней мере, x социальных цитирований, его альт-индекс будет x*». Социальный индекс исследователей вычисляется на основе предложенного альт-индекса [17].

У программных продуктов управления цитированием есть функции со средствами социального об-

щения, позволяющие пользователям находить и следить друг за другом. Данная аналитика является новой и считается частью изменений альтметрик, которая отслеживает нетрадиционные библиографические метрики. Как отмечалось ранее, у *ResearchGate* есть свой собственный показатель, названный RG-Score, который присваивает участникам сети рейтинги, основанные на взаимодействиях с контентом и рейтингами участников, взаимодействующих с контентом. Контент, такой как информация в профиле и отвеченные или заданные вопросы, влияет на рейтинг RG-Score в дополнение к информации о публикациях, такой как как размещения, загрузки и цитирования. Рейтинг RG-Score не имеет стандартного библиографического способа измерения. Таким образом, его применение зависит от учреждения [18].

КОНСТРУИРОВАНИЕ ВЕБОМЕТРИЧЕСКИХ ИНДИКАТОРОВ ДЛЯ ОЦЕНКИ ПРОДУКТИВНОСТИ ИССЛЕДОВАТЕЛЕЙ

Композитные показатели включают иные виды первичных показателей, поэтому справедливо называть их *полувебометрическими (semi-webometric)*. Конструируя вебометрические индикаторы, следует обратиться к паттернам показателей научного цитирования. При этом альтметрики конструируются на основании паттерна альтметрик. Композитные вебометрические индикаторы должны конструироваться по смешанному паттерну. Паттерны конструирования показателей научного цитирования [19] в данном исследовании трактуются шире – как паттерны конструирования произвольных наукометрических показателей. При построении вебометрических индикаторов следует ограничиться сложением, вычитанием, умножением, дискретным суммированием и произведением первичных показателей, в роли которых выступают вебометрические индикаторы низшего порядка или простые альтметрики.

При помощи преобразования вебификации на основании таблицы сконструируем композитный показатель $@W$ продуктивности деятельности исследователей, являющийся полувебометрическим (рис. 2).

Начиная с компоненты $@h$ полувебометрического показателя $@W$, которая характеризует влиятельность (в переводе на англ. язык – *impact*) исследователя, положим, что хиршеподобный альтиндекс $@h_\alpha$ – это максимальное n такое, что n работ одного и того же автора получили k^{n-1} просмотров (скачиваний $\alpha = @_{load}$, «лайков» $\alpha = @_{like}$ и т.п.), где $n \in \mathbb{N}$, $k \in \mathbb{N}$, $\alpha \in \mathcal{A}$. Определение «хиршеподобный» указывает на то, что данный показатель сконструирован по подобию библиометрического индекса Хирша. Приставка «альт» дает понять, что индекс является вебометрическим, т. е. сконструирован на основании альтметрик. Для каждой альтметрики следует подобрать и зафиксировать параметр k. Например, для количества просмотров электронных текстов научных работ параметр k взят равным 10. Параметры оценивания k и α задаются для полного множества объектов оценки, которыми выступают исследователи, так как оценка продуктивности каждого из исследователей, согласно принципу объективности, осуществляется в равных условиях.

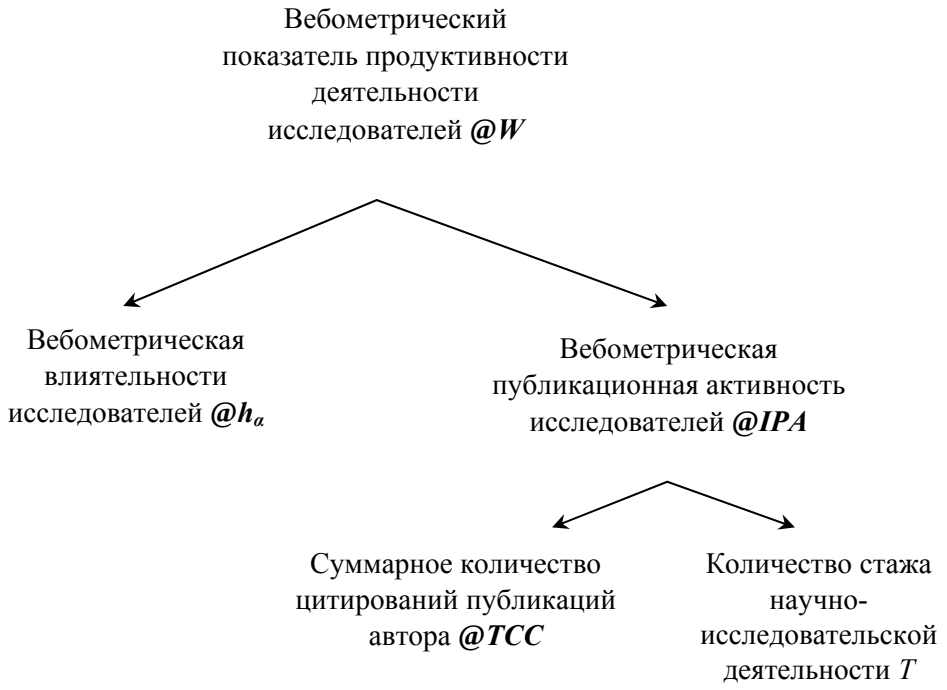

Рис. 2. Структура полувебометрического показателя продуктивности научно-исследовательской деятельности исследователей

Перейдем к следующей компоненте *@IPA* полувебометрического показателя *@W*, которая характеризует публикационную активность исследователя. Сначала необходимо сконструировать обычный индикатор *IPA* публикационной активности автора. Для этого предлагается использовать следующую формулу:

$$IPA = h \cdot a_h \cdot m = h \cdot \frac{TCC}{h^2} \cdot \frac{h}{T} = \frac{TCC}{T}\ , \qquad (5)$$

где h – индекс Хирша; a_h – так называемый a_h-индекс; m – так называемый m-квотиент; TCC – суммарное количество цитирований публикаций исследователя; T – количество лет стажа научно-исследовательской деятельности автора.

Первичный показатель TCC суммарного количества цитирований может быть вебофицирован при помощи таблицы в показатель $@TCC$ суммарного количества размещений ссылок на электронный текст работы, подсчитываемый на основании альтметрики $@_{TCC}$ при помощи поискового робота. Показатель T никак не вебофицируется, что вполне приемлемо для структуры полувебометрических показателей. Количество лет стажа научно-исследовательской деятельности T возможно условно принять за разницу между текущим годом и годом выхода первой научной публикации рассматриваемого исследователя.

Таким образом, получаем формулу оценки вебометрического индикатора *@IPA* публикационной активности автора:

$$@IPA = \frac{@TCC}{T}\ , \qquad (6)$$

где $@TCC$ – суммарное количество ссылок на электронные тексты работы исследователя; T – количество лет стажа научно-исследовательской деятельности этого же автора.

На практике значения вебометрических индикаторов подвержены сильным динамическим изменениям почти ежедневно. Кроме того, выводы делаются не на основании значений вебометрического индикатора конкретного объекта, а на основании положения (рейтингового места в общем списке) данного объекта среди других объектов относительно значения выбранного вебометрического индикатора. Именно поэтому вместо того, чтобы задавать целевую функцию (например, сюда подходит умножение $@h_\alpha$ и $@IPA$) лучше использовать бинарное отношение $\prec$ упорядочивания объектов O_i и O_j относительно вебометрического индикатора $@W$:

$$\begin{aligned}
&\left(@h_\alpha^{O_i} < @h_\alpha^{O_j} \Rightarrow @W_{O_i} \prec @W_{O_j}\right) \vee \\
&\vee \left\{ \begin{array}{l} \left(@h_\alpha^{O_i} = @h_\alpha^{O_j}\right) \wedge \left(@IPA_{O_i} < @IPA_{O_j}\right) \Rightarrow \\ \Rightarrow @W_{O_i} \prec @W_{O_j} \end{array} \right\} \vee \\
&\vee \left\{ \begin{array}{l} \left(@h_\alpha^{O_i} = @h_\alpha^{O_j}\right) \wedge \left(@IPA_{O_i} = @IPA_{O_j}\right) \Rightarrow \\ \Rightarrow @W_{O_i} = @W_{O_j} \end{array} \right\},
\end{aligned} \qquad (7)$$

где $@h_\alpha^{O_i}$ и $@h_\alpha^{O_j}$ – влиятельность, соответственно, *i*-го и *j*-го исследователей по α-ой альтметрики; $@IPA_{O_i}$ и $@IPA_{O_j}$ – публикационная активность, соответственно, *i*-го и *j*-го исследователей; $@W_{O_i}$ и $@W_{O_j}$ – публикационная активность, соответственно, *i*-го и *j*-го исследователей; $\prec$ – бинарное отношение упорядочивания по значению вебометрического индикатора $@W$.

Вебометрический критерий (7) работает следующим образом:

- более высокой признается продуктивность того исследователя, у которого выше влиятельность, задаваемая хиршеподобным альтиндексом по заранее выбранной альтметрике;
- при равной влиятельности следует считать продуктивнее того исследователя, у которого выше публикационная активность по состоянию на момент сравнения.

Критерий (7) позволяет установить, у какого исследователя выше продуктивность его деятельности по состоянию на текущий момент времени. Предложенный вебометрический критерий не позволяет упорядочить сразу весь список исследователей, но критерий можно использовать при определении бинарного отношения порядка в попарном сравнении, ранжируя исследователей таким же образом, как сортируются одномерные массивы.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании рассматриваются вопросы из относительно молодого направления информатики – *вебометрии*, которую также называют «сетевой наукометрией». В контексте эксперимента по переходу на новые способы материального поощрения труда научных работников среди прочих интересных трендов весомую практическую значимость приобретает проблема создания вебометрического рейтинга авторов научных публикаций. Если вебометрический рейтинг будет коллегиально принят и одобрен отечественным научным сообществом, то появится возможность включить его в систему показателей государственной наукометрической системы. Предпринятая попытка построения методики формирования вебометрического рейтинга авторов или, по крайней мере, проектирования такого вебометрического рейтинга на концептуальном уровне, не затрагивает техническую сторону вопроса, которая касается реализации конкретных программных решений.

Однако вместо шаблонного построения очередной рейтинговой модели выбран иной путь создания критерия ранжирования исследователей относительно показателей продуктивности их собственной научной деятельности. Предложенный критерий опирается на вебометрические индикаторы, которые формируются на основании подсчета альтметрик. Также уделяется внимание общим аспектам академических социальных сетей, разнообразие которых порождает различные типологии и классификации альтметрик. На этом фоне выявляются сильные и слабые стороны альтметрик, дающие обоснование разработке нового вебометрического критерия.

Модель *IPPA* (*Impact, Productivity & Publication Activity*) способствует формулированию рекомендаций по исправлению параметров систем оценки публикационной активности. Оптимизация достигается посредством уточнения представлений о реальных достижениях отдельных исследователей, а вместе с ними научно-образовательных организаций и научного сообщества в целом. Вебометрический критерий предназначается для использования в качестве методологического компонента обслуживания инфраструктурных потребностей управления научно-исследовательской деятельностью. В более дальней перспективе вебометрический критерий полезен при осуществлении интеграции механизма контроля за финансово-хозяйственным положением науки и научно-исследовательской деятельности как одной из производственных составляющих инновационной экономики.

СПИСОК ЛИТЕРАТУРЫ

1. MLE on Open Science Altmetrics and Rewards – Different types of Altmetrics. – URL: https://rio.jrc.ec.europa.eu/en/library/mle-open-science-%E2%80%93-report-different-types-altmetrics.
2. Gutmann P. The commercial malware industry // DEFCON conference. – 2007. – URL: http://stacy-konkiel.org/sociologybot/?platform=hootsuite.
3. Jeng W., DesAutels S., He D., Li L. Information exchange on an academic social networking site: A multidiscipline comparison on ResearchGate Q&A // Journal of the Association for Information Science and Technology. – 2017. – Vol. 68, № 3. – P. 638–652.
4. Haustein S., Peters I., Bar-Ilan J., Priem J., Shema H., Terliesher J. Coverage and adoption of altmetrics sources in the bibliometric community // Scientometrics. – 2014. – Vol. 101, № 2. – P. 1145–1163.
5. Sugimoto C. R., Work S., Larivière V., Haustein S. Scholarly use of social media and altmetrics: a review of the literature // Journal of the Association for Information Science and Technology. – 2017. – Vol. 68, № 9. – P. 2037–2062.
6. Hassan S.U., Imran M., Gillani U., Aljohani N.R., Bowman T.D., Didegah F. Measuring social media activity of scientific literature: an exhaustive comparison of scopus and novel altmetrics big data // Scientometrics. – 2017. – Vol. 113, № 2. – P. 1037–1057.
7. Valuing science and scholarship. Relations between quality, impact and indicators. – URL: https://www.cwts.nl/research/cwts-research-program-2017-2022.
8. Martín-Martín A., Orduña-Malea E., Ayllón J.M., López-Cózar E.D. The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter // arXiv preprint arXiv:1602.02412. – 2016. – URL: https://arxiv.org/ftp/arxiv/papers/1602/1602.02412.pdf.
9. Hoffmann C.P., Lutz C., Meckel M. A relational altmetric? Network centrality on ResearchGate as an indicator of scientific impact // Journal of the Association for Information Science and Technology. – 2016. – Vol. 67, № 4. – P. 765–775.
10. Bornmann L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics // Journal of informetrics. – 2014. – Vol. 8, № 4. – P. 895–903.
11. Haustein S. Grand challenges in altmetrics: heterogeneity, data quality and dependencies // Scientometrics. – 2016. – Vol. 108, № 1. – P. 413–423.
12. Bornmann L., Haunschild R. Measuring field-normalized impact of papers on specific societal groups: An altmetrics study based on Mendeley Data // Research Evaluation. – 2017. – Vol. 26, № 3. – P. 230–241.
13. Bornmann L. Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics // Scientometrics. – 2015. – Vol. 103, № 3. – P. 1123–1144.
14. Zahedi Z., Costas R., Wouters P. How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications // Scientometrics. – 2014. – Vol. 101, № 2. – P. 1491–1513.
15. Lapinski S., Piwowar H., Priem J. Riding the crest of the altmetrics wave: How librarians can help prepare faculty for the next generation of research impact metrics // arXiv preprint arXiv:1305.3328. – 2013. – URL: https://arxiv.org/ftp/arxiv/papers/1305/1305.3328.pdf.
16. Waltman L., van Eck NJ., van Leeuwen T.N., Visser M.S., van Raan A.F. Towards a new crown indicator: Some theoretical considerations // Journal of informetrics. – 2011. – Vol. 5. – № 1. – P. 37–47.
17. Hassan S.U., Gillani U.A. Altmetrics of "altmetrics" using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wiki // arXiv preprint arXiv:1603.07992. – 2016. – URL: https://arxiv.org/ftp/arxiv/papers/1603/1603.07992.pdf.
18. Ovadia S. ResearchGate and Academia.edu: Academic social networks // Behavioral & Social Sciences Librarian. – 2014. – Vol. 33, № 3. – P. 165–169.
19. Калачихин П.А. Паттерны конструирования показателей научного цитирования // Научно-техническая информация. Сер. 2. – 2017. – № 7. – С. 1-10; Kalachikhin P.A. Patterns for constructing scientific citation index // Automatic documentation and mathematical linguistics. – 2017. – Vol. 51, № 4. – P. 171–179.

Материал поступил в редакцию 28.03.18.

Сведения об авторе

КАЛАЧИХИН Павел Андреевич – кандидат экономических наук, старший научный сотрудник ВИНИТИ РАН, Москва
e-mail: pakalachikhin@viniti.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 004.85.021:[025.4.025:82 – 1/–9]

Н. Н. Буйлова

Классификация текстов по жанрам с помощью алгоритмов машинного обучения*

Рассмотрена проблема классификации документов по жанрам, выделены основные характеристики текста, используемые для распознавания его жанра, и описаны наиболее применяемые алгоритмы машинного обучения. Приведенные методы служат для классификации научных, технических, публицистических и художественных текстов.

Ключевые слова: *классификация текстов, определение жанра, машинное обучение*

ВВЕДЕНИЕ

Классификация документов – одна из основных подзадач информационного поиска. Из неструктурированных данных пользователь стремится получить документы, релевантные его запросу. Часто поисковый запрос включает не только ключевые слова, но требования к функциональному стилю документа, удобочитаемости текста, гендерной принадлежности автора, его возрасту и т. д. В этом случае документ должен иметь набор метатегов, описывающих его характерные особенности. Для крупных хранилищ текстов, таких как корпуса и библиотеки, определение тегов является ключевым требованием метаразметки.

Одним из способов классифицирования является определение функциональных стилей текстов. В свою очередь, функциональные стили делятся на жанры [1]. Согласно определению М.М. Бахтина, жанр – это «устойчивый тематически, композиционно и стилистически тип высказывания» [2, с. 255]. В отличие от литературной формы (обладающей четкими критериями) понятие жанра текста с трудом поддается формализации: при том, что параметры, описывающие конкретный жанр, вывести не удается, экспертная оценка человека, как и машинный классификатор, пользуясь неявными признаками, в большинстве случаев верно определяют жанр текста. Выявление скрытых признаков подробно рассматривается в статье [3] на примере коллекции медицинских текстов.

Автоматическое определение функциональных стилей текста – одно из приоритетных направлений классификации документов, однако в настоящий момент эта задача в значительной степени решена – в том числе в связи с особо актуальной для Всемирной паутины проблемой фильтрации спама, например [4]. Однако жанровое многообразие (особенно для художественной литературы) все еще недостаточно изученная область, что, по всей вероятности, связано с отсутствием четкого набора формальных признаков жанра; более того, анализ жанровой принадлежности затрудняется тем, что каждая пара жанров различается уникальными параметрами – от длины текста в символах до структуры предложения.

Несмотря на значительную сложность, задача автоматической классификации текстов по жанрам является привлекательной сферой исследований как в прикладном (вышеупомянутая фильтрация спама, индексация документов при поиске в корпусе или библиотеке, «родительский контроль» Интернета и т. д.), так и в теоретическом плане, например, четкое определение стилистических особенностей того или иного жанра может использоваться в обучении студентов словесности.

В прикладной области особое значение приобретает автоматическое определение жанра для структурирования библиотек и баз данных – постоянно пополняющиеся коллекции научных публикаций и художественной литературы растут со скоростью, которая не позволяет классифицировать тексты вручную, а объемы подобных корпусов на сегодняшний день таковы, что и поиск без привлечения алгоритмов не принесет плодов. Стоит отметить, что эта задача существует для корпусов на различных языках, а для большей части европейских и азиатских языков уже

* Исследование выполнено на основании гранта Российского научного фонда (проект № 16-18-02071 «Пограничный русский: оценка сложности восприятия русского текста в теоретическом, экспериментальном и статистическом аспектах»).

есть корпуса текстов, сравнимые с первым репрезентативным корпусом Брауна (т. е. содержащие не менее одного миллиона словоупотреблений) [5-7].

Актуальность рассматриваемой нами проблемы настолько высока, что уже имеются обзоры методов классификации текстов [8-10]. Однако эти обзоры сосредотачивались либо на технической стороне реализации алгоритмов, либо на рассмотрении общих подходов к классификации текста. Согласно нашим сведениям, не было предпринято попытки рассмотреть классификацию по жанрам в отдельности. Наш обзор посвящен методам определения жанра в научной, технической, публицистической и художественной литературе.

РАННИЕ ПУБЛИКАЦИИ, ПОСВЯЩЕННЫЕ ПРОБЛЕМЕ КЛАССИФИКАЦИИ ЖАНРА

Первые прикладные работы по распознаванию жанра текста появились в середине 1990-х гг. [11-12], в их основе лежали теоретические работы Д. Бибера [13]. Проведенные на Корпусе текстов университета Брауна эксперименты применяли методы дискриминантного анализа [12] и логистической регрессии, в том числе с использованием нейросети [11]. Базовым описанием текстов были частеречные характеристики текста, а также различные меры удобочитаемости. На их основе текстам присваивалось три типа признаков – сложность текста, наличие нарратива и жанр. Несмотря на высокое качество исходной разметки, доля правильного распознавания жанра не поднималась выше 83%. Более того, размеры этих выборок были недостаточны для обсуждения качества алгоритма.

Представляется необходимым упомянуть работу [14], в которой описанная комбинация алгоритмов позволила значительно повысить качество снятия неоднозначности. В дальнейшем этот подход лег в основу современных методов «мешка слов и деревьев решений», получивших широкое развитие в середине 2000-х гг. из-за стремительного разрастания Интернета и необходимости классифицировать все большее количество текстов. Тогда же появляются способы классификации, основанные на частеречной разметке [15]. На тот момент использование частеречной разметки позволяло распознавать жанр текста с высокой (96,9%) точностью в случае качественных исходных данных (газетная статья), в противном случае точность падала до 85,7% (сообщения форумов). Другим крупным ответвлением классификации документов стала классификация с использованием HTML-разметки, позволяющей комбинировать квантитативные методы описания самого текста с нетекстовыми элементами разметки гипертекста [16, 17]. Кроме того, следует упомянуть использование синтаксически размеченных корпусов (*treebank*), позволяющих проводить анализ дискурсивных связей [18].

Сегодня существует ряд способов машинной классификации текстов по жанрам, которые используют классические методы машинного обучения. Далее мы рассмотрим основные из них: наивный байесовский классификатор, деревья решений, случайный лес, метод опорных векторов.

ХАРАКТЕРИСТИКИ ТЕКСТА, ИСПОЛЬЗУЕМЫЕ ПРИ КЛАССИФИКАЦИИ

Функционирование классификаторов любой природы предполагает предобработку текстов для получения машиночитаемых данных. На сегодняшний день создано много способов описания отдельных репрезентативных признаков документа. Эту задачу можно описать как представление текстов в виде векторов, атрибуты которых делятся на два типа – частотные (каждое значение в векторе d соответствуют количеству вхождений признаков в документ d) и бинарные (каждое значение в векторе – бинарное и отражает факт присутствия признака в документе) [19].

Наиболее простым способом представления текста являются униграммы, также называемые «мешком слов» (*bag of words*), которые представляют собой набор слов документа без каких-либо связей между ними. Несмотря на свою простоту, модель имеет ряд недостатков: так, не учитываются грамматические связи между словами, порядок токенов и вероятность совместной встречаемости слов [20, 21].

Метод «мешка слов» обладает большим адаптационным потенциалом: например, в работе [22] помимо классического «мешка слов» использовалось расширение этого понятия, а именно подстроки из каждого предложения (если в стандартном методе «мешка слов» документ разбивается на токены от пробела до пробела, то в этой работе выделялись последовательности из нескольких слов, идущих в предложении подряд).

Помимо униграмм, применяются также би- и триграммы, которые могут представлять собой как букво- так и словосочетания из двух или трех элементов (эти комплексы являются частным случаем n-грамм, однако сочетания с n больше трех встречаются реже из-за больших объемов корпусов, необходимых для сбора достаточно репрезентативной выборки). В работе [23] критерий кластеризации, основанный на близости между двухбуквенными распределениями текстов, позволяет правильно идентифицировать автора с ошибкой не более 5%, а жанр – с ошибкой не более 15%.

Кроме простого «мешка слов» возможно использование процессированного списка слов, характеризующих документ. Такая метрика называется TF-IDF (TF – *term frequency*, IDF – *inverse document frequency*), и она оценивает важность слова в определенном документе относительно других текстов коллекции [24, 25]. Синтаксическая аннотация текста – еще одна важная характеристика, используемая для описания стиля [26] и, таким образом, имеющая потенциал в качестве признака при классификации жанров. Синтаксис, понятый как способ связи слов в предложении и предложений в тексте, характеризует строение фраз и иных структурных единиц, что позволяет выделить важные черты документа и жанра в целом [27].

Еще одним интересным типом признаков для классификации являются дискурсивные связи, а именно – способы объединения текста в единое целое при помощи вспомогательных лексических еди-

ниц. В работе [18] использовались эксплицитные (союзы и прочие служебные слова) и имплицитные (подразумеваемые, но не выраженные явно) дискурсивные связи.

Параметры удобочитаемости (*readability*) также могут быть использованы в качестве параметров классификации. В их состав входят такие атрибуты, как длина предложения, длина текста в символах, количество единиц определенных частей речи и т.д. Такой формат описания дает краткую характеристику документа [28].

Современные документы, существующие в сети Интернет, обладают еще одним классифицирующим параметром, а именно – HTML-разметкой. Особый способ структурной организации текста позволяет использовать метаинформацию о документе для определения жанра [24, 25].

Более подробно проблема выделения наиболее информативных признаков для классификации текстов описана в статье [3]. Все перечисленные характеристики текста используются в качестве информации для классификаторов.

КЛАССИФИКАТОРЫ ЖАНРА ТЕКСТА

Наивный байесовский классификатор

Метод наивного байесовского классификатора основывается на предположении, что слова независимы друг от друга (появление в тексте слова *A* не влияет на появление в тексте слова *B)*. В этом случае можно вычислить вероятность существования какого-либо списка слов (текста) при условии, что текст относится к определенному классу документов и заданы априорные вероятности появления каждого класса. Максимальная из вычисленных вероятность будет соответствовать классу документа [29].

Классификация более чем 9000 текстов, относящихся к семи различным жанрам – теле- и радионовостям, рекламе, репортажам и т.д. была проведена в работе [30]. Для этого использовались частотности слов, лингвистические параметры текста (временные формы глаголов, синтаксическая сложность), а также комбинация обоих вариантов. Использование наивного баейсовского классификатора при учете частотности слов позволило получить 76,7% точности распознавания, тогда как применение лингвистических параметров в отдельности снизило точность до 33,9%. Это хорошо отражает такие недостатки наивного баейсовского классификатора, как низкая восприимчивость к грамматической информации, в том числе, этот метод не учитывает вероятность появления в документе слов одного семантического поля («убийство» и «преступник» с большей вероятностью окажутся в одном тексте, нежели «любовь» и «пистолет») и тот факт, что вероятность встретить слово в разных местах текста также различается.

Тем не менее, модификация наивного байесовского классификатора, такая как линейный многоклассовый классификатор с отбором признаков, показывает высокие результаты распознавания при классификации научных текстов по отраслям знания [31].

Деревья решений

Дерево решений представляет собой граф или модель решений, учитывающий их возможные исходы и их вероятности. В применении к классификации документов алгоритм дерева решений начинается с выбора разделяющего слова, затем коллекция делится на две части и процедура выполняется заново до тех пор пока все документы коллекции не будут рассортированы. В листьях разрешающего дерева размещаются значения целевой функции, в прочих узлах — условия перехода, определяющие направление движения вдоль ребер дерева. Для классификации каждого примера алгоритму необходимо пройти все дерево от корня до одного из листов и тем самым получить значение целевой функции [32].

Такой подход реализован в статье [33], авторы которой предполагают, что пространство текста частотно (т.е. образовано частотами появления в тексте наборов признаков, к которым относятся служебные слова, биграммы, буквосочетания и т.д.). Полученные данные классификации предполагается использовать для определения метапараметров текста – жанра, автора, стиля и т.д.

Случайный лес

Дальнейшим развитием метода дерева решений является алгоритм «случайный лес» (*random forest*) – ансамблевый метод машинного обучения, который использует совокупность решающих деревьев, построенных независимо друг от друга. Финальная классификация документов проводится с помощью «голосования», т. е. итоговым классом объекта объявляется тот класс, который был решением большинства деревьев [34]. Известной сложностью применения алгоритма «случайный лес» является значительное число решающих деревьев, требующееся для большинства задач, что предъявляет высокие требования к объему памяти.

Оценка качества классификации текстов с помощью алгоритма «случайный лес» была проведена в работе [35]. Классификация материалов, представленных в сети Интернет, имеет практическую ценность как для оценки их содержания, так и для поиска и извлечения специализированной информации (например, научных публикаций по заданной теме). В этом исследовании выполнялась классификация неструктурированной информации по наличию в ней тем, связанных с противозаконной деятельностью. Примененный метод «случайный лес» показал высокую точность при классификации изучаемых данных, что особенно хорошо проявилось при использовании сбалансированных положительных и отрицательных выборок при обучении. Таким образом следует отметить, что алгоритм «случайный лес» склонен к переобучению при неравновесных выборках, что заметно сказывается на точности классификации.

В настоящее время алгоритм «случайный лес» широко используется при решении самых разнообразных задач классификации жанров на корпусах разных языков, примером чего может служить клас-

сификации по жанрам корпуса турецких газет, где с помощью этого алгоритма жанр распознавался правильно в диапазоне от 88% до 93% в зависимости от имеющихся наборов признаков [36].

Метод опорных векторов

Одним из мощных алгоритмов обучения с учителем является метод опорных векторов (*Support Vector Machine* – SVM). Классификация с помощью этого метода происходит благодаря поиску оптимальной разделяющей гиперплоскости в пространстве векторов высокой размерности [37].

Работа [30], посвященная классификации новостных текстов, наравне с наивным байесовским классификатором использует подход SVM, который показывает хороший результат распознавания (82%) даже с простыми признаками (частотности слов), а комбинация частотностей слов с лингвистическими параметрами только улучшает результаты работы классификатора.

ЗАКЛЮЧЕНИЕ

Определение жанра текста – это одна из необходимых задач, решаемых при организации электронных библиотек научных публикаций или художественной литературы. С середины 1990-х гг. до настоящего времени как теоретические основы, так и практическое применение классификации текстов по жанрам неуклонно расширяются. Нами были рассмотрены способы описания текста такими метриками, как «мешок слов», би- и триграммы, TF-IDF, дискурсивные связи, удобочитаемость и HTML-разметка. Применяемые как по отдельности, так и в комбинациях, эти характеристики служат надежной базой для работы машинных классификаторов. В задаче распознавания жанра текста используются алгоритмы различной природы и сложности – от наивного байесовского классификатора и деревьев решений, до методов «случайного леса» и SVM.

СПИСОК ЛИТЕРАТУРЫ

1. Голубева И.Б. Стилистика русского языка. – 2-е изд., испр. – М.: Рольф, 1999. – 448 с.
2. Бахтин М. М. Эстетика словесного творчества. – М.: Искусство, 1986. – 556 с.
3. Мангалова Е.С., Агафонов Е.Д. О проблеме выделения информативных признаков в задаче классификации текстовых документов // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2013. – №1.
4. Скляренко Н.С. Обзор алгоритмов машинного обучения, решающих задачу обнаружения спама // Новые информационные технологии в автоматизированных системах. – 2017. – №20. – С. 26-31.
5. Ehsani R., Muzaffer E.A., Gülsen E. at all. Disambiguating Main POS tags for Turkish // ROCLING. – 2012. – №20. – P. 1121-1128.
6. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author // Computational linguistics. – 2000. – Vol. 26, № 4. – P. 56-63.
7. Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M., Al-Rajeh A. Automatic Arabic text classification // JADT'08. – 2008. – P. 77–83.
8. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. – 2010. – №1. – С. 72-76.
9. Agarwal B., Mittal N. Text Classification Using Machine Learning Methods-A Survey // Proceedings of the Second International Conference on Soft Computing for Problem Solving. – 2012. – Vol. 236. – С. 89-95.
10. Sebastiani F. Machine learning in automated text categorization // ACM Comput. – 2002. – Surv. 34(1). – P. 1–47.
11. Kessler B., Nunberg G., Schutze H. Automatic Detection of Text Genre // Computing Research Repository. – 1997. – Vol. 29, № 1. – P. 1224-1229.
12. Karlgren J., Cutting D. Recognizing text genres with simple metrics using discriminant analysis // Proceedings of Coling. – 1994. – Vol. 4, № 3. – P. 12-19.
13. Biber D. The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding // Computers in the Humanities. – 1992. – № 3 – С. 58-64.
14. Schütze H. Automatic word sense discrimination // Computational Linguistics. – 1998. – Vol. 24, №1. – С. 29-36.
15. Giesbrecht E., Evert S. Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus // Proceedings of the 5th Web as Corpus Workshop (WAC5). – NY: NYPublish, 2009.
16. Rehm G. Towards Automatic Web Genre Identification // Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02). – 2002. – Vol. 4.
17. Boese E.S., Howe A.E. Effects of web document evolution on genre classification // Proceedings of the 14th ACM international conference on Information and knowledge management. – 2005. – Vol.6. – P. 89-95.
18. Webber B. Genre distinctions for Discourse in the Penn TreeBank // Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. – 2009. – P. 674–682.
19. Maas A., Daly R., Pham P. and all. Learning word vectors for sentiment analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. - 2011.
20. Wallach H. Topic modeling: beyond bagofwords // Proceedings of the 23rd International Conference on Machine Learning. – 2006. – P. 977-984.
21. McCallum A., Wangand X., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the Seventh IEEE International Conference on Data Mining. – 2007. – P. 697-702.
22. Radošević D., Dobša J., Mladenić D. at all. Genre Document Classification Using Flexible Length Phrases // Information and Intelligent Systems. – 2006. – Vol. 6 – P . 66-75.

23. Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Прикладная информатика. – 2013. – Т. 26, № 2. – С. 95-108.
24. Lee Y.-B., Myaeng S.H. Text genre classification with genre-revealing and subjectrevealing features // Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. – 2002. – P. 145–150.
25. Snyman D.P., Van Huyssteen G.B., Daelemans W. Automatic Genre Classification for Resource Scarce Languages // Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa. – 2011. – P. 132–137.
26. Biber D. Dimensions of Register Variation: A Cross-Linguistic Comparison. – Cambridge: Cambridge University Press, 1995. – 428 p.
27. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author // Computational linguistics. – 2000. – Vol. 26, № 4. – P. 471–495.
28. Falkenjack J., Santini M., Jonsson A. An Exploratory Study on Genre Classification using Readability Features // The Sixth Swedish Language Technology Conference. – 2016. – Vol. 6. – P. 72-78.
29. Li Y.H., Jain A.K. Classification of text documents // Computer Journal. – 1998. – № 41(8). – P. 537-46.
30. Dewdney N., Van Ess-Dykema C., MacMillan R. The form is the substance: classification of genres in text // Proceedings of the workshop on Human Language Technology and Knowledge Management. – 2001. – Vol. 7. – P. 56-65.
31. Сухарева А.В., Царьков С.В. Классификация научных текстов по отраслям знаний // Машинное обучение и анализ данных. – 2014. – Т. 1, № 8. – С. 22-25.
32. Quinlan J.R. Simplifying decision trees // International Journal of Man-Machine Studies. – 1987. – Vol.7 – P. 152-160.
33. Кубарев А.И., Кукушкина О.В., Поддубный В.В. и др. Построение таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений // Вестн. Томск. гос. ун-та. Сер. Управление, вычислительная техника и информатика. – 2012. – № 4. – С. 79–88.
34. Kam T. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition. – 1995. – P. 278–282.
35. Веретенников И., Карташев Е., Царегородцев А. Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «Случайный лес» // Известия АлтГУ. – 2017. – №4 (96). – С. 25-36.
36. Amasyali F. M., Banu D. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. – Springer-Verlag Berlin Heidelberg, 2006.
37. Cortes C., Vapnik V. Support vector networks // Machine Learning. – 1995. – Vol. 20. – P. 237–297.

Материал поступил в редакцию 16.03.18.

Сведения об авторах

БУЙЛОВА Надежда Николаевна – преподаватель Школы лингвистики Факультета гуманитарных наук Национального исследовательского университета – Высшая Школа Экономики, Москва
e-mail: nbujlova@hse.ru

Центр (Отдел) научно-информационного обслуживания (ЦНИО) ВИНИТИ РАН

Информационные услуги, предоставляемые ЦНИО ВИНИТИ РАН:

– проведение тематического поиска и консультации поисковых экспертов;
– подготовка списков научной литературы;
– подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
– библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
– информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

– фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
– базами данных и Интернет-ресурсами: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
– полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНИТИ www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНИТИ http://catalog.viniti.ru** и **служба электронной доставки документов.**

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНИТИ.

Обращаться в ЦНИО ВИНИТИ:
- адрес: 125190, Россия, г. Москва, ул. Усиевича, 20;
- телефоны: 8(499) 155 -42 -43, 8(499) 155 -42 -17;
- эл. почта cnio@viniti.ru, fdk@viniti.ru;
- факс 8(499) 930 -60 -00 (для ЦНИО).

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru