УДК 303-722-4
https://www.doi.org/10.47813/nto.5.2024.5005

EDN SSJOOA

Персонализация адаптивно-обучающей методики Л.А. Растригина на базе частотного словаря с использованием кластерного анализа корпусов текстов

К.В. Полянский[1*], И.В. Ковалев[1,2,3,4]

[1]Сибирский федеральный университет, пр. Свободный, 79, Красноярск, 660041, Россия
[2]Красноярский краевой Дом науки и техники Российского союза научных и инженерных общественных объединений, ул. Урицкого, 61, Красноярск, 660049, Россия
[3]Навоийский государственный горно-технологический университет, Навои, Узбекистан
[4]Красноярский государственный аграрный университет, Красноярск, Россия

*E-mail: k.v.polyansky@gmail.com

Аннотация. Рассмотрено применение адаптивно-обучающей методики Л.А. Растригина к частотному мультилингвистическому словарю по системному анализу. Выявлены недостатки данной методики: отсутствие персонализации по отношению к обучаемому при выдаче порций обучающей информации. В качестве решения предложено модифицировать критерий качества обучения. В его формулу вводятся коэффициенты значимости, полученные из данных на основе кластерного анализа обучающей коллекции корпусов текстов. Для этого проведено TF-IDF взвешивание терминов из коллекции и составлена матрица корреляции на основе косинусных расстояний их векторов. Как результат, получен новый критерий качества обучения, учитывающий терминологические предпочтения ученика и опирающийся на значимость терминов в обучающей коллекции корпусов текстов. Модифицированная адаптивно-обучающая методика персонализирует процесс обучения, делая его более гибким и современным.

Ключевые слова: адаптивно-обучающая методика, частотный словарь, кластерный анализ.

Personalization of the adaptive learning method of L.A. Rastrigin based on a frequency dictionary using cluster analysis of text corpora

K.V. Polyansky[1*], I.V. Kovalev[1,2,3,4]

[1]Siberian Federal University, 79 Svobodny pr., Krasnoyarsk, 660041, Russia
[2]Krasoyarsk Science & Technology City Hall, 61, Uritskogo street, Krasnoyarsk, 660049, Russia
[3]Navoi State University of Mining and Technologies, Navoi, Uzbekistan
[4]Krasnoyarsk State Agrarian University, Krasnoyarsk, Russia

*E-mail: k.v.polyansky@gmail.com

Abstract. The application of the adaptive learning method by L.A. Rastrigin to a frequency multilingual dictionary on system analysis is considered. The disadvantages of this method are revealed, such as the lack of personalization in relation to the student when issuing portions of training information. As a solution, it is proposed to modify the criterion of the quality of training. Its formula includes coefficients of significance obtained from data based on cluster analysis of the training collection of text corpora. For this purpose, TF-IDF weighting of terms from the collection was carried out and a correlation matrix was compiled based on the cosine distances of their vectors. As a result, a new criterion of the quality of training was obtained, taking into account the terminological preferences of the student and based on the significance of terms in the training collection of text corpora. The modified adaptive learning method personalizes the learning process, making it more flexible and modern.

Keywords: adaptive learning method, frequency dictionary, cluster analysis.

1. Введение

Современный человек в повседневной жизни сталкивается с большим потоком текстовой информации, существенная часть которой представлена на иностранных языках. Поэтому все большее значение для него имеют системы, способные выполнять интенсивное, а главное – персонализированное (учитывающее предметную область и интересы) обучение иностранной терминологии.

Для быстрого освоения лексики в узких предметных областях создаются специализированные словари иностранных терминов. Примером может служить частотный мультилингвистический словарь по системному анализу и информационным технологиям [1], небольшой фрагмент которого представлен в таблице 1.

Таблица 1. Фрагмент частотного мультилингвистического словаря по системному анализу и информационным технологиям.

	RU		EN	GE
1	1227	тип, 5	type, 3	Typ, 3
2	683	функция, 113	function, 127	Funktion, 77
3	1378	порт, 4	port, 76	Port, 5
4	122	анализ, 327	analysis, 197	Analyse, 169
5	106	алгоритм, 126	algorithm, 169	Algorithmus, 89
6	280	байт, 47	byte, 146	Byte, 23
7	506	энергия, 15	energy, 267	Energie, 13
8	745	цель, 94	goal, 17	Ziel, 9
9	919	язык, 101	language, 110	Sprache, 97
10	941	слой, 2	layer, 39	Schicht, 2

На базе данного словаря проводится обучение иностранной терминологии по системному анализу и в сфере IT. В качестве алгоритма обучения выступает адаптивно-обучающая методика Л.А. Растригина [2].

Основное преимущество данной методики в том, что алгоритм подстраивается под обучаемого на каждом сеансе обучения. Он выдает необходимые порции терминов в зависимости от частоты их употребления, а также от качества запоминания

информации учеником. Структурная схема процесса адаптивного обучения приведена на рисунке 1.

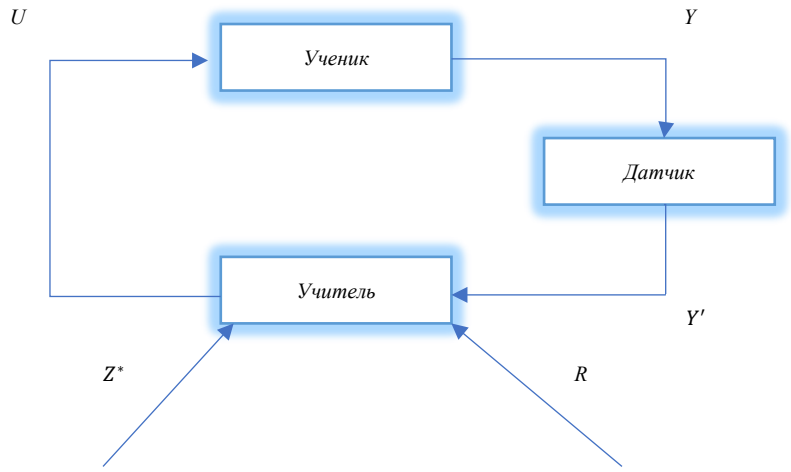

Рисунок 1. Схема процесса адаптивного обучения.

На схеме используются следующие обозначения:

- *Учитель* (обучающее устройство) – это программный комплекс (приложение), посредством которого происходит обучение иностранной терминологии.
- *Ученик* (объект управления) – человек (обучаемый), проходящий обучение.
- *Датчик* – подсистема программного комплекса (приложения), осуществляющая проверку знаний *Учеником* и отправляющая результаты *Учителю.*
- U – порция обучающей информации, которую *Учитель* подает на *Ученика*.
- Y – состояние *Ученика*, подаваемое на *Датчик* (ответы на тесты).
- Y' – информация о состоянии *Ученика*, получаемая *Учителем* (интерпретация ответов на тесты: оценки, баллы, вероятности незнания терминов и пр.).
- Z^* – цели обучения, сообщаемые *Учителю* (например, критерий качества, при котором сеанс обучения считается завершенным).
- R – ресурсы, которыми *Учитель* располагает для обучения (частотный словарь терминов, информация о сеансах обучаемого и т.д.).

Благодаря своей ориентированности на индивидуальные особенности запоминания терминов адаптивно-обучающая методика позволяет выстроить процесс обучения максимально эффективно. А в связке с частотным мультилингвистическим

словарем становится средством интенсивного обучения наиболее популярной иностранной терминологии.

Однако, у методики есть и существенный недостаток, она не учитывает персональные лексические интересы ученика. Обучаемый же почти всегда хочет изучать в первую очередь значимые для него (в профессиональном или бытовом контексте) термины, которые связаны с родом его деятельности, задачами или увлечениями.

Персонализация адаптивно-обучающей методики смогла бы повысить ее релевантность относительно лексических целей обучаемого. Персональный подход к обучению делает его более гибким и современным.

2. Постановка задачи (Цель исследования)

В основе адаптивно-обучающего алгоритма лежит критерий качества обучения. Благодаря этому критерию на *n*-ом сеансе обучения алгоритм предлагает ученику именно те термины, которые на (*n-1*)-ом сеансе ученик запомнил хуже всего. Также, в первую очередь он выдает термины, являющиеся наиболее популярными (с высокой частотностью в словаре).

Исследуем подробно данный критерий, рассмотрим, как он работает при классическом подходе и предложим решение по его модификации с целью персонализации адаптивно-обучающей методики. Формула критерия качества обучения Q_n имеет вид [3]:

$$Q_n = \sum_{i=1}^{N} p_i(t_i^n) \times q_i \rightarrow min, \quad (1)$$

где *n* – номер сеанса обучения,

N – общее количество терминов информационно-терминологического базиса (ИТБ) частотного словаря,

$p_i(t_i^n)$ – вероятность незнания *i*-го термина из *n*-го набора обучающей информации (ОИ),

q_i – относительная частота, выражающая долю лексической единицы (термина) в тексте, подвергшемуся статистической обработке при составлении частотного словаря.

Адаптивно-обучающий алгоритм стремится минимизировать произведение $p_i(t_i^n) \times q_i$ для каждого *i*-го термина к концу сеанса обучения. Этого можно достичь, подавая в порцию обучающей информации к концу сеанса максимальные значения данного произведения.

Если учесть, что в начале обучения вероятность незнания любого i-го термина $p_i(t_i^1) = 1$ (т.к. обучение еще не начиналось, и все термины из набора ученику незнакомы), то минимизация критерия качества обучения Q_1 в самом начале сводится к минимизации q_i по формуле 2 и включению в порцию ОИ терминов с максимальной относительной частотой.

$$Q_1 = \sum_{i=1}^{N} q_i \rightarrow min, \quad (2)$$

Используем в качестве ИТБ адаптивно-обучающего алгоритма фрагмент частотного словаря из таблицы 1 (в примере - его русскоязычная часть). Алгоритм с учетом формулы 2 в качестве порции ОИ будет выбирать термины в порядке, указанном в таблице 2: от наиболее частотных к менее частотным. Это классический подход, когда относительная частота употребления терминов играет главную роль при выдаче обучающей информации.

Таблица 2. Выдача терминов к обучению. Классический подход**.**

	RU	q^{o}	$q^{к}$	$q^{o}+q^{к}$
1	анализ, 327	0.327	0.000	0.327
2	алгоритм, 126	0.126	0.000	0.126
3	функция, 113	0.113	0.000	0.113
4	язык, 101	0.101	0.000	0.101
5	цель, 94	0.094	0.000	0.094
6	байт, 47	0.047	0.000	0.047
7	энергия, 15	0.015	0.000	0.015
8	тип, 5	0.005	0.000	0.005
9	порт, 4	0.004	0.000	0.004
10	слой, 2	0.002	0.000	0.002

Этот подход хорош тем, что в первую очередь выдает к обучению термины с высокой частотностью, а значит – наиболее популярные. Однако, зачастую в начале обучения для ученика важно изучить термины, актуальные именно для него, связанные с его предметной областью (образующие единый предметный кластер). Классический подход не может обеспечить персонализацию обучения и становится менее

эффективным. Актуальные для обучаемого термины вытесняются наиболее частотными и могут изучаться в последнюю очередь, что при больших размерах ИТБ может стать существенным недостатком и повлиять на качество обучения.

Предположим, что обучаемому наиболее интересны термины его предметной области (образующие кластер [«*цель*», «*энергия*», «*слой*»] и выделенные в таблице 2), и он хотел бы в первую очередь получать к обучению именно эти термины, а также термины каким-то образом с ними связанные. Следовательно, нужно повысить вес этих кластерных терминов относительно остальных. Этого можно добиться, рассматривая q_i как комплексный показатель, состоящий, собственно, из частотного $q_i^{\text{о}}$, а также некоторого добавочного (кластерного) показателя $q_i^{\text{к}}$.

$$q_i = q_i^{\text{о}} + q_i^{\text{к}} \quad (3)$$

Тогда критерий качества на первом шаге обучения примет вид:

$$Q_1 = \sum_{i=1}^{N} q_i^{\text{о}} + q_i^{\text{к}} \rightarrow min, \quad (4)$$

Для обеспечения минимизация критерия качества Q_1 в самом начале обучения в порцию ОИ будут включены термины с максимальным значением суммы $q_i^{\text{о}} + q_i^{\text{к}}$.

В классическом подходе (таблица 2) $q_i^{\text{к}}$ для каждого термина равно 0. Поэтому фактически расчетная частота равна относительной: $q_i = q_i^{\text{о}}$. В случае подхода, основанного на персонализации обучения, для достижения цели необходимо повышать значение кластерного показателя $q_i^{\text{к}}$ для терминов, которые в первую очередь интересны ученику и понижать это значение для всех остальных.

Задача повышения веса терминов из предметной области обучаемого сводится к формированию кластерного вектора $\left[q_1^k, q_2^k, \dots, q_N^k\right]$ для терминов всего ИТБ. Кластерные показатели терминов предметной области в этом векторе должны иметь больший вес по сравнению с остальными. Кластерный вектор отображает взаимосвязь между терминами, то, как они сгруппированы (кластеризованы) по этим показателям.

Для выявления взаимосвязей между терминами ИТБ, а также нахождения объединяющих их групп (кластеров) необходимо провести кластерный анализ обучающей коллекции корпусов текстов, основанной на терминах частотного словаря из таблицы 1.

Этот анализ позволит найти кластерный вектор $[q_1^k, q_2^k, \dots, q_N^k]$, изменить на его основе критерий качества адаптивного обучения, сделав само обучение персонализированным по отношению к обучаемому.

3. Методы и материалы исследования

Для проведения кластерного анализа на основе частотного словаря (таблица 1) разобьем его термины на 3 условные корпуса текстов: «Математика» (*тип, функция, анализ, алгоритм, цель*), «IT» (*тип, функция, порт, байт, язык*) и «Энергетика» (*тип, порт, анализ, энергия, слой*) и запишем в таблицу так, чтобы каждой строке соответствовали термины соответствующего корпуса. Для простоты будем рассматривать только термины русского языка.

Таблица 3. Коллекция корпусов текстов**.**

Математика	тип	функция		анализ	алгоритм	цель
IT	тип	функция	порт		байт	язык
Энергетика	тип		порт	анализ	энергия	слой

Термины были подобраны таким образом, чтобы можно было сгруппировать их по признаку уникальности для каждой корпуса. Рассматриваемые корпуса являются идеализированными моделями, не содержащими стоп-терминов и других, характерных для реальных корпусов лексем.

Актуальные для обучаемого термины «*цель*», «*энергия*» и «*слой*» (которые мы рассматривали ранее) вошли в корпуса «Математика» и «Энергетика». Введем некоторые обозначения:

- *d* - корпус текста (в таблице 3 приведено 3 корпуса текста: «Математика», «IT» и «Энергетика»)
- *D* - коллекция корпусов текста – множество рассматриваемых корпусов текстов d
- *t* – термин, принадлежащий тому или иному корпусу текста d

Теперь проанализируем частотную составляющую терминов *t* для коллекции корпусов *D* из таблицы 3. Для начала будем рассчитывать частоту вхождения каждого термина в своем корпусе.

$$tf(t,d) = \frac{количество\ появлений\ термина\ t\ \ корпусе\ d}{общее\ количество\ терминов\ в\ корпусе\ d} \quad (1)$$

Далее рассчитаем «инверсную частоту документа». Эта характеристика указывает значимость термина уже не в рамках одного корпуса, но в рассматриваемой коллекции корпусов.

$$idf(t,D) = log(\frac{общее\ количество\ корпусов\ в\ коллекции\ корпусов\ D}{количество\ корпусов\ с\ термином\ t\ в\ коллекции\ корпусов\ D}) \quad (2)$$

Произведение значений $tf(t,d)$ и $idf(t,D)$ дает сводную характеристику TF-IDF, отражающую как значимость термина для корпуса, так и для коллекции корпусов в целом [4].

$$TF - IDF = tf(t,d) \times idf(t,D) \quad (3)$$

	алгоритм	анализ	байт	порт	слой	тип	функция	цель	энергия	язык
0	0.534	0.406	0.000	0.000	0.000	0.315	0.406	0.534	0.000	0.000
1	0.000	0.000	0.534	0.406	0.000	0.315	0.406	0.000	0.000	0.534
2	0.000	0.406	0.000	0.406	0.534	0.315	0.000	0.000	0.534	0.000

Рисунок 2. Термины, взвешенные по алгоритму TF-IDF.

Значения TF-IDF, вычисленные (значения приведены с корректировкой на алгоритм TfidfVectorizer [5] модуля sklearn языка python) для всех терминов коллекции корпусов приведены на рисунке 2. Здесь строками являются корпуса текстов: 0 - «Математика», 1 - «IT», 2 - «Энергетика», а столбцами – термины частотного словаря.

Чтобы произвести векторное сравнение пары терминов воспользуемся их косинусным расстоянием [6].

$$\cos\theta = \frac{a \times b}{\|a\| \times \|b\|}, \quad (4)$$

где $a \times b$ – скалярное произведение векторов a и b,

$\|a\| \times \|b\|$ – произведение их длин

Перемножив попарно векторы из таблицы на рисунке 2, получим матрицу корреляции [7] между терминами (рисунок 3). Значениями данной матрицы являются

значения косинусов угла между векторами терминов, записанных как ее строки и столбцы.

	алгоритм	анализ	байт	порт	слой	тип	функция	цель	энергия	язык
алгоритм	1.000	0.746	0.368	0.368	0.368	0.655	0.746	1.000	0.368	0.368
анализ	0.746	1.000	0.368	0.607	0.746	0.832	0.607	0.746	0.746	0.368
байт	0.368	0.368	1.000	0.746	0.368	0.655	0.746	0.368	0.368	1.000
порт	0.368	0.607	0.746	1.000	0.746	0.832	0.607	0.368	0.746	0.746
слой	0.368	0.746	0.368	0.746	1.000	0.655	0.368	0.368	1.000	0.368
тип	0.655	0.832	0.655	0.832	0.655	1.000	0.832	0.655	0.655	0.655
функция	0.746	0.607	0.746	0.607	0.368	0.832	1.000	0.746	0.368	0.746
цель	1.000	0.746	0.368	0.368	0.368	0.655	0.746	1.000	0.368	0.368
энергия	0.368	0.746	0.368	0.746	1.000	0.655	0.368	0.368	1.000	0.368
язык	0.368	0.368	1.000	0.746	0.368	0.655	0.746	0.368	0.368	1.000

Рисунок 3. Матрица корреляции терминов.

Используя данные матрицы корреляции, можно определять, насколько взаимосвязаны (коррелированы) термины в рамках коллекции корпусов *D* (наибольшей корреляции соответствует значение 1, наименьшей - 0). А задача нахождения совместной корреляции нескольких терминов сводится к построчному перемножению значений их косинусов. Полученные перемноженные значения можно нормализовать по методу «минимакс» [8] (формула 5).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (5)$$

где X – текущее значение,

X_{min} и X_{max} – минимальное и максимальное значение из диапазона соответственно.

Вернемся к исходной задаче персонализации адаптивно-обучающей методики. На примере терминов «*цель*», «*энергия*» и «*слой*» необходимо построить кластерный вектор $[q_1^k, q_2^k, \dots, q_N^k]$, такой, чтобы подстановка его значений в критерий качества обучения (формула 4) формировала порции обучающей информации из терминов, входящих с исходными в один кластер.

Используя матрицу корреляции терминов, вычислим нормализованное произведение значений косинусов для векторов терминов «*цель*», «*энергия*» и «*слой*» и получим их кластерный вектор, изображенный на рисунке 4.

Рисунок 4. Кластерный вектор для терминов «слой», «цель» и «энергия».

Каждое *i*-е значение данного вектора – это кластерный показатель $q_i^к$ из формулы 4, которому также соответствует *i*-е значение относительной частоты $q_i^о$.

Используя помимо частотной составляющей терминов их кластерную составляющую, адаптивный алгоритм меняет свое поведение и выдает к обучению термины с учетом их актуальности для ученика. Мы получаем к выдаче термины в порядке, отличном от того, который мы имели при классическом подходе (таблица 2).

4. Полученные результаты

В таблице 4 приведены термины в порядке убывания показателя $q_i^о + q_i^к$. Видно, что термины «*цель*», «*энергия*» и «*слой*» подвинулись вверх в списке терминов для выдачи к обучению.

Таблица 4. Выдача терминов к обучению. Кластерный подход.

	RU	$q^о$	$q^к$	$q^о+q^к$
1	анализ, 327	0.327	1.000	1.327
2	энергия, 15	0.015	0.886	0.901
3	слой, 2	0.002	0.886	0.888
4	тип, 5	0.005	0.677	0.682
5	порт, 4	0.004	0.493	0.497
6	алгоритм, 126	0.126	0.326	0.452
7	цель, 94	0.094	0.326	0.420
8	функция, 113	0.113	0.243	0.356
9	язык, 101	0.101	0.120	0.221
10	байт, 47	0.047	0.120	0.167

Проведя анализ результатов работы кластерного подхода, можно отметить следующее:

- Термины «*энергия*» и «*слой*» существенно поднялись вверх в списке выдачи ОИ, позиция термина «*цель*» изменилась, но менее заметно
- Термины «*тип*» и «*порт*» показывают большую степень корреляции с терминами «*энергия*» и «*слой*», т.к. они используются с ними в связке в одном корпусе «Энергетика». Поэтому они вытесняют термин «*цель*» по своей значимости для всего корпуса
- Термин «*алгоритм*» стоит на позицию выше, чем «*цель*» за счет равных значений $q_i^к$ и большего значения q_i^o
- Термины «*язык*» и «*байт*» имеют одинаковые $q_i^к = 0.120$. Эти термины имеют нулевую корреляцию с терминами «*цель*», «*энергия*» и «*слой*», поэтому ранжирование при равных $q_i^к$ осуществляется только за счет разницы значений относительной частоты q_i^o. Следует сказать, что кластерный подход в целом предполагает подобное поведение: первые N (в зависимости от распределения значимости) наиболее связанных терминов (попавших в один кластер с исходными) поступают для выдачи в ОИ в соответствии со значением $q_i^к$, остальные термины с индексами больше N ранжируются по значению q_i^o (соответствует поведению при классическом подходе).

5. Заключение

По результатам проведенного исследования можно сделать следующие выводы:

- Рассмотрен алгоритм частотного взвешивания TF-IDF, метод векторного сравнения терминов на базе косинусного расстояния.
- Составлена корреляционная матрица, на основе которой предложено прогнозировать принадлежность терминов тому или иному корпусу текстов (кластеру).
- Выполнено сведение алгоритмов адаптивного обучения и кластеризации терминов в единую концепцию.
- Произведена персонализация адаптивно-обучающего алгоритма Л.А. Растригина на базе частотного мультилингвистического словаря с использованием

кластерного анализа корпуса текстов. Модифицирован критерий качества обучения данного алгоритма.

- Представлено сравнение релевантности выдачи информации относительно лексических целей обучаемого при классическом и кластерном подходах. Последний показывает лучшие результаты в контексте выдачи наиболее связанных с запросами ученика терминов.

Список литературы

1. Ковалев И.В. Мультилингвистический частотный словарь по системному анализу и информационным технологиям / И.В. Ковалев, М.В. Карасева. – Красноярск: СибГУ им. М.Ф. Решетнева, 2021. – 82 с.
2. Ковалев И.В. Системные аспекты организации и применения мультилингвистической адаптивно-обучающей технологии / И.В. Ковалев, М.В. Карасева, Е.А. Суздалева // Educational Technology & Society. – 2002. – № 5. – С. 207-208.
3. Карасева М.В. Модификация алгоритма обучения иностранной терминологии / М.В. Карасева, А.А. Ступина, М.И. Мельдер // Современные проблемы науки и образования. – 2014. – № 3. – С. 6.
4. Jurafsky Daniel. Question Answering, Information Retrieval, and Retrieval Augmented Generation / Daniel Jurafsky, James H. Martin // Speech and Language Processing, Standford. – 2024. – № 3-14. – С. 4.
5. Scikit Learn: сайт. – 2007. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращения: 01.11.2024).
6. Muhammad Umaruddin Syam. Implementation of Cosine Similarity Algorithm on Omnibus Law Drafting / Aristoteles*, Muhammad Umaruddin Syam, Tristiyanto, Bambang Hermanto // (IJACSA) International Journal of Advanced Computer Science and Applications. – 2024. – № 4-15. – С. 202.
7. Alexandria Ree Hadd. Correlation Matrices in Cosine Space.: диссертация ... master of science in psychology / Alexandria Ree Hadd. – Nashville, Tennessee, 2016. – 4 с.
8. Старовойтов В.В. Нормализация данных в машинном обучении / В.В. Старовойтов, Ю.И. Голуб // Информатика – 2021. – Т. 18, № 3. – С. 86.