

Роберт Валерьевич Майер

Глазовский государственный инженерно-педагогический университет, доктор педагогических наук, доцент,
профессор кафедры физики и дидактики физики, Глазов, Россия
e-mail: robert_maier@mail.ru

О современных методах анализа учебных текстов

Аннотация. Учебник — это многомерный объект, который можно описать целой совокупностью различных показателей. Являясь моделью учебной дисциплины, учебник определяет методику обучения и содержание учебно-познавательной деятельности, поэтому развитие методов его анализа — актуальная проблема дидактики. Цель статьи состоит в изучении возможностей применения методов цифровой гуманитаристики для анализа учебных текстов. Рассмотрим конкретные примеры использования компьютерных методов анализа учебных текстов: 1) выявление ключевых понятий параграфа (раздела, курса); 2) изучение распределения терминов в тексте; 3) построение фазовой кривой изучения дисциплины; 4) установление семантической близости между разделами курса; 5) определение семантического сходства между научными понятиями, изучение семантического пространства; 6) кластеризация учебных понятий. Перечисленные методы позволяют исследовать содержание, структуру и эволюцию учебных курсов, оптимизировать учебники и методические материалы, повысить эффективность методики обучения.

Ключевые слова: гуманитаристика, дидактика, обучение, понятие, семантика, текст, учебник.

Robert V. Mayer

The Glazov Korolenko State Pedagogical University, Doctor of Pedagogical Sciences,
Associate Professor, Professor of the Department of Physics and Didactics of Physics, Glazov, Russia
e-mail: robert_maier@mail.ru

On Modern Methods of Analysing Educational Texts

Abstract. A textbook is a multidimensional object that can be described by a whole set of various indicators. As a model of an academic discipline, a textbook defines the teaching methods and the content of educational and cognitive activities. Therefore, the development of methods for its analysis is a pressing issue in didactics. The purpose of this article is to explore the potential of applying digital humanities methods to the analysis of academic texts. We will consider specific examples of the use of computer methods for analysing educational texts: 1) identifying key concepts in a paragraph (section, course); 2) studying the distribution of terms in the text; 3) constructing a phase curve for learning a subject; 4) establishing semantic proximity between course sections; 5) determining semantic similarities between scientific concepts, studying the semantic space; 6) clustering of academic concepts. These methods make it possible to study the content, structure, and evolution of academic courses, optimise textbooks and teaching materials, and improve the effectiveness of teaching methods.

Keywords: humanitaristics, didactics, teaching, concept, semantics, text, textbook.

Введение (Introduction)

Любой учебник — это многомерный объект, который можно описать целой совокупностью различных показателей. В общем случае он содержит большое количество понятий, принципов, законов, учебных заданий, формул, рисунков, которые связаны между собой логическими, ассоциативными и иными связями, образуют единую систему. Являясь моделью учебной дисциплины, учебник определяет методику обучения и содержание учебно-познавательной деятельности школьников и студентов.

При анализе различных текстов часто применяют следующие методы цифровой гуманитаристики: 1) частотный анализ терминов: построение частотных списков, упорядочивание терминов по частоте, выявление ключевых понятий; 2) количественный контент-анализ: автоматический подсчет ключевых терминов, понятий и категорий; 3) тематическое моделирование: выявление в тексте важных тем, которым соответствуют группы терминов-маркеров; 4) семантический анализ: выявление скрытых смыслов и связей между

© Майер Р. В., 2026

Для цитирования: Майер Р. В. О современных методах анализа учебных текстов // Вестник Омского государственного педагогического университета. Гуманитарные исследования. 2026. № 2 (51). С. 199–203. DOI: 10.36809/2309-9380-2026-51-199-203



Статья доступна на условиях лицензии

Creative Commons Attribution 4.0 International (CC BY 4.0)

понятиями, определение семантического сходства и различия концептов; 5) коллокационный анализ: выявление устойчивых словосочетаний; 6) определение семантического расстояния между терминами путем нахождения косинусной меры близости, изучение семантического пространства текста; 7) метод TF-IDF (Term Frequency-Inverse Document Frequency): оценка значимости терминов, которые часто встречаются в исследуемом тексте, но редко используются в данном корпусе текстов по той же теме, что позволяет выделить уникальные термины, присущие анализируемому документу. Перечисленные методы помогают изучать особенности учебных текстов для их сравнения и совершенствования.

Методы (Methods)

Цель статьи состоит в изучении возможностей методов цифровой гуманитаристики при анализе учебных текстов (УТ). Методологической основой исследования являются работы следующих ученых: Н. К. Андриевская [1], С. Х. Бермудес [2], А. С. Ванюшкин и Л. А. Гращенко [3], Т. Н. Гнитецкая [4], Л. А. Кузнецов и В. Ф. Кузнецова [5], Р. В. Майер [6], С. И. Монахов, В. В. Турчаненко, Е. А. Федюкова и Д. Н. Чердаков [7], Ю. И. Морозова [8], С. В. Ракитина [9], С. Д. Мэннинг, П. Рагхаван и Х. Шютце [10]. При этом применяются следующие методы: контент-анализ учебной литературы, определение семантической близости понятий и УТ, кластерный анализ, программирование, обоснование путем перечисления примеров, элементы системного подхода. Для обработки и форматирования текстов, автоматического выявления и подсчета ключевых понятий, построения графиков использовались восемь компьютерных программ (КП), написанных в ABCPascal, а также ресурсы Интернета. Визуализация полученных результатов осуществлялась в MS Excel.

Результаты и обсуждение (Results and Discussion)

Изучение вхождения различных терминов и особенностей их дистрибуции в учебных текстах позволяет установить их структуру и смысловые взаимосвязи. Рассмотрим конкретные примеры использования компьютерных методов анализа УТ.

1. Выявление ключевых концептов параграфа (раздела, курса). Ключевыми называются концепты, усвоение которых имеет определяющее значение для понимания текста в целом. Они, как правило, чаще других упоминаются в учебнике, используются при рассмотрении различных вопросов и тем, имеют большое количество однокоренных слов и связей. Например, часто встречающиеся в учебнике физики слова «электрический», «электрон», «электризация», «электродвигатель», «электромагнит», «наэлектризовать» относятся к одному концепту «электричество».

Чтобы выделить ключевые концепты, УТ обрабатывают на интернет-ресурсе advego.com. На выходе получается список всех используемых слов с указанием частот употребления. Ресурс обрабатывает тексты объемом не более 100 000 символов, поэтому тексты учебников приходится анализировать по частям. Получающиеся файлы с помо-

щью специальной КП-1 объединяют в один так, что количества употреблений одинаковых слов при этом суммируются. КП-1 на выходе выдает общий список слов, который с помощью Word упорядочивают по алфавиту. Из него выделяют часто используемые однокоренные понятия, выявляют ключевые концепты и вычисляют суммарное количество их употреблений.

В качестве примера рассмотрим ключевые понятия физики микромира, выявленные в итоге контент-анализа текстового файла, полученного в результате слияния следующих фрагментов: 1) вузовский учебник Т. И. Трофимовой [11] (параграфы 202–275, за исключением параграфов 206, 216–222, 229–250, 260); 2) вузовский учебник С. И. Кузнецова [12, с. 114–264]; 3) статьи из «Википедии»: «атом», «молекула», «электрон», «протон», «нейтрон», «ион», «кварк» [13]; 4) ответы нейросети *gpt* на запросы типа «перечисли ключевые понятия (идеи) физики микромира» и т. д. Получившееся облако слов представлено на рисунке 1.1; размер шрифта пропорционален частоте термина. Для 38 часто употребляемых концептов определим отношение числа употреблений n_i i -го понятия ($i = 1, 2, \dots, 38$) к суммарному числу употреблений всех понятий и построим круговую диаграмму (рис. 1.2). Видно, что более 50 % словоупотреблений приходится на электрон — 0,126 (1246 раз), частицу — 0,108 (1064 раза), ядро — 0,099 (981), энергию — 0,094 (925), атом — 0,092 (912).

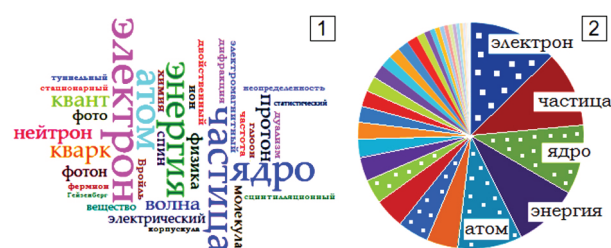


Рис. 1. Ключевые понятия темы «Физика микромира»

2. Анализ дистрибуции терминов в тексте. Изучение распределения конкретных терминов в тексте учебника позволяет судить о тематике и содержании отдельных разделов, ключевых темах, структуре учебника, связях между понятиями, их синергии и общей сложности УТ. Для анализа распределения понятий в тексте учебника астрономии [14] используется следующий метод:

1. Из pdf-файла учебника получают текстовый файл. С помощью специальной КП-2 преобразуют все буквы в прописные, удаляют все символы, за исключением букв русского алфавита (включая цифры, знаки препинания), а также слова длиной 1 или 2 буквы и двойные пробелы.

2. С помощью КП-3 определяют координаты x_i изучаемого термина (например, концепта «Солнце»), т. е. номерá его словоформ, отсчитываемые от начала УТ; результаты при этом записываются в файл *Vihod.txt*. Для этого следует отбросить конечную изменяемую часть слова: программа должна искать не «планета», а «планет». На выходе получается список из нескольких сотен координат $x_i \in [1; M]$. В нашем случае объем текстового файла $N = 32\,500$ слов.

3. Найденные координаты термина x_i вставляют в КП-4, рисуящую интегральную кривую распределения данного термина вдоль текста, а также двумерные и трехмерные фазовые кривые (рис. 2).

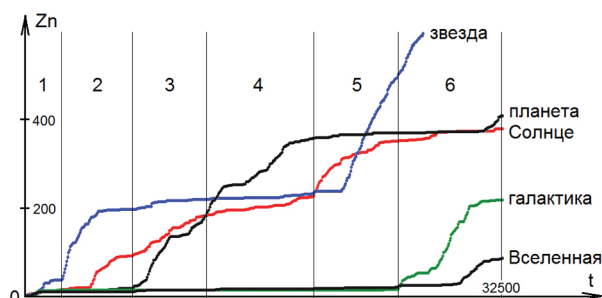


Рис. 2. Распределение понятий в учебнике астрономии [14]

На рисунке 2 представлены графики распределения ключевых понятий «планета», «Солнце», «звезда», «галактика», «Вселенная» в учебнике астрономии [14]. Вертикальные линии показывают границы разделов: 1. Введение. 2. Практические основы астрономии. 3. Строение Солнечной системы. 4. Природа тел Солнечной системы. 5. Солнце и звезды. 6. Строение и эволюция Вселенной. Видно, что звезды подробно изучаются в 1-й половине 2-го раздела, 2-й половине 5-го и 1-й половине 6-го раздела; планеты — в 3-м, 4-м разделах и в конце курса, а Солнце — во 2-й половине 2-го раздела, 3-м и 5-м разделах. В тексте учебника термин «звезда» упоминается 718 раз, «Солнце» — 379 раз, «планета» — 408 раз, «галактика» — 217 раз, «Вселенная» — 85 раз. Их синергия приводит к более глубокому пониманию материала.

3. Построение фазовой кривой текста. Рассмотрим фазовое пространство, образованное осями, которые соответствуют ключевым концептам (их множествам или темам), упоминаемым в УТ. Например, для текста по астрономии вдоль одной оси откладывают количество упоминаний понятия «планета», вдоль другой — «Солнце», вдоль третьей — «звезда». По мере продвижения вдоль текста количества употреблений этих понятий изменяются, и в фазовом пространстве точка вырисовывает некоторую кривую. Разные астрономические тексты будут иметь различные фазовые кривые. Если фазовое пространство нормировано, то на осях откладывается доля упоминаний понятия (П) от общего числа упоминаний П в тексте (изменяется от 0 до 1).

Для построения фазовой кривой получают одномерные матрицы координат ключевых x_i терминов рассмотренным выше методом. С помощью КП-5 нами получены дву- и трехмерные фазовые кривые учебника [14] в нормированных пространствах «планета — Солнце» (рис. 3.1) и «Солнце — звезда — планета» (рис. 3.2). На них отмечены точки, показывающие начало и конец 1-го (0–a), 2-го (a–b), 3-го (b–c) и последующих разделов. Получившиеся графики (рис. 3) позволяют сделать вывод об особенностях

изложения учебного материала, синергии изучаемых понятий и сопоставить различные тексты.

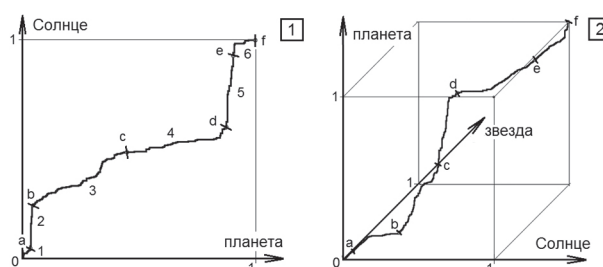


Рис. 3. Получающиеся фазовые кривые в 2D- и 3D-пространствах

4. Установление семантической близости между УТ.

Для выявления силы меж- и внутрипредметных связей может быть использован метод ключевых слов. Для каждого текста T_A и T_B составляют списки ключевых слов A и B с указанием числа их употреблений. Если УТ содержат формулы, то выписывают понятия, соответствующие входящим в формулы величинам. Путем объединения списков A и B получают тезаурус объемом $N = N(A \cup B)$. В N -мерном пространстве, образованном частотами употребления ключевых слов, списки A и B (а значит, и тексты T_A , T_B) изображаются векторами $\vec{a}(a_1, a_2, \dots, a_N)$ и $\vec{b}(b_1, b_2, \dots, b_N)$. Для вычисления косинусной меры семантического сходства текстов T_A и T_B применяется известная формула, рассмотренная в работах [1; 2].

Для учета логических связей между понятиями в используемой программе КП-6 создают два массива $w1[i]$ и $w2[i]$, содержащие M пар связанных терминов, один из которых входит в определение другого. Например: нуклон — нейтрон, электрон — заряд, мощность — энергия и т. д. Если в файле $t1.txt$ понятие $w1[i]$ имеет частоту a_i , а в файле $t2.txt$ логически связанное с ним понятие $w2[i]$ встречается с частотой b_i , то программа находит произведение $0,75a_i b_i$, которое прибавляется к числителю и знаменателю дроби в формуле для $K(A, B)$. Получается так:

$$K(A, B) = \frac{a_1 b_1 + a_2 b_2 + \dots + a_N b_N + 0,75(a_1' b_1' + a_2' b_2' + \dots + a_M' b_M')}{|a| \cdot |b| + 0,75(a_1' b_1' + a_2' b_2' + \dots + a_M' b_M')}$$

Если $A \cap B = \emptyset$ и нет связанных терминов, то $K(A, B) = 0$, т. е. векторы $\vec{a}$ и $\vec{b}$ ортогональны. Когда T_A и T_B очень близки по содержанию, векторы $\vec{a}$ и $\vec{b}$ почти сонаправлены и $K(A, B)$ примерно равен 1.

В результате анализа учебника астрономии [15] получена матрица семантических расстояний $L(A, B) = 1 / K(A, B) - 1$ между восемью его разделами. Это позволило построить граф (рис. 4), на котором короткие (т. е. сильные) внутрипредметные связи показаны прямыми отрезками, а длинные (слабые) — изогнутыми линиями. Из него следует, что ядро курса образуют разделы 3, 5 и 7, которые близки друг к другу в содержательном плане; от них наиболее удален раздел 1 и в меньшей степени — раздел 4. Аналогичным образом можно оценить силу межпредметных связей различных дисциплин.

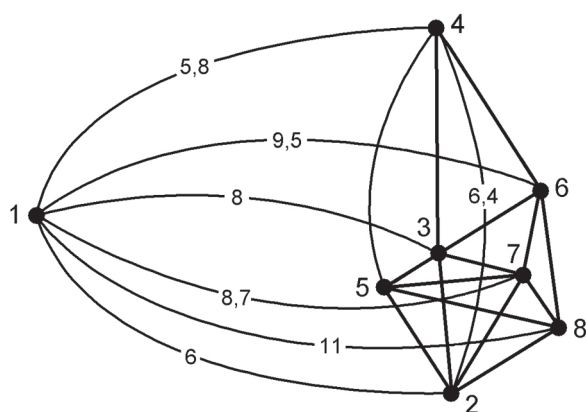


Рис. 4. Связи между разделами учебника [15]

5. Построение семантического пространства УТ.

Сущность этого метода заключается в определении семантического расстояния между ключевыми понятиями текста [1; 2] и в построении графа, визуализирующего связи между ними (рис. 5). Из анализируемого текста удаляют все символы, за исключением русских букв, стоп-слова, двойные пробелы и все слова длиной 1 или 2 буквы. Запускают КП-7, которая разбивает оставшуюся последовательность слов на N перекрывающихся контекстных окон по 20 слов со смещением 5 слов и считает количество вхождений в них ключевых понятий. Для каждого понятия находят контекстный вектор — одномерную матрицу $\vec{a}(a_1, a_2, \dots, a_N)$, элементы которой a_k показывают число вхождений термина в k -е окно ($k = 1, 2, \dots, N$). Косинусная мера близости $K(A, B)$ понятий A и B вычисляется как косинус угла между векторами $\vec{a}(a_1, a_2, \dots, a_N)$ и $\vec{b}(b_1, b_2, \dots, b_N)$ в N -мерном контекстном пространстве [1; 2]. Результаты записываются в файл Vihod.txt.

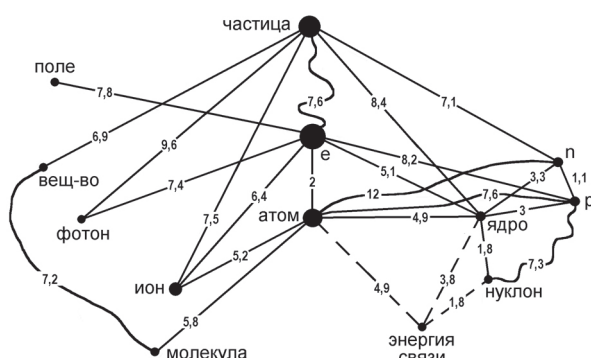


Рис. 5. Семантическое пространство понятий

С помощью Excel находят семантическое расстояние между понятиями $L(A, B) = 1 / K(A, B) - 1$, получают матрицу расстояний и на ее основе рисуют граф (рис. 5). Он получен в результате анализа текстов по физике микромира общим объемом 60 тыс. слов (речь идет об упомянутых выше фрагментах учебников [11; 12] и т. д.). Определяют центральное понятие, к которому близки все остальные.

Короткие ребра изображают прямыми отрезками, а длинные — кривыми линиями. Радиус кругов (вершин графа) пропорционален частоте использования понятия в УТ. Видно, что семантические расстояния отличаются в десятки и сотни раз. Расстояние между редко используемыми понятиями определяется неточно, соответствующие ребра нарисованы пунктиром.

6. Кластеризация понятий. Матрица близости понятий позволяет осуществить их кластеризацию. На рисунках 6.1 и 6.2 показаны кластеры, образованные ключевыми понятиями физики микромира в учебных физических текстах. Они получены в результате объединения упомянутых выше фрагментов учебников [11; 12], статей из «Википедии» [13] и ответов нейросети.

Кластеризация осуществляется в Excel в соответствии со следующим алгоритмом: 1) сначала считают, что каждый i -й объект является отдельным кластером с весом m_i , показывающим число его употреблений в УТ; 2) находят два наиболее близких кластера K_i и K_j ; 3) объединяют их в один новый кластер K_{ij} с весом $m_{ij} = m_i + m_j$; 4) вычисляют расстояние от нового кластера K_{ij} до остальных кластеров, учитывая вес K_i и K_j :

$$L(K_{ij}, K_r) = \frac{m_i L(K_i, K_r) + m_j L(K_j, K_r)}{m_i + m_j}, r = 1, \dots, N,$$

кроме i и j ; $L(K_{ij}, K_{ij}) = 0$; 5) если число кластеров больше, чем требуется, то возвращаются к операции 2, а иначе — заканчивают кластеризацию.

Из рисунка 6.1 видно, что на первом шаге объединились понятия «протон» и «нейтрон», затем «атом» и «электрон» и т. д. Результаты кластеризации ключевых понятий у разных текстов, вообще говоря, различны, но имеют схожие черты. Последнее обусловлено объективно существующими связями между понятиями и обозначаемыми объектами.

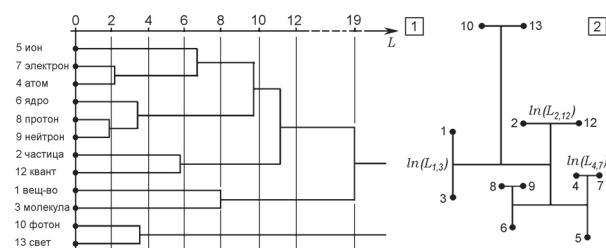


Рис. 6. Кластеризация ключевых понятий физики микромира

Представленные на рисунке 6 результаты отражают важные связи физики микромира, которые проявляются в перечисленных текстах. Например, из протонов и нейтронов состоит атомное ядро. Атом, кроме ядра, содержит электроны. Если от атома оторвать или добавить электроны, то он превратится в ион. Молекула — предел химической делимости вещества. Свет — поток фотонов и т. д. Длины отрезков, соединяющих кластеры на рисунке 6.2, пропорциональны логарифмам $\ln(L_{ij})$, где $L_{ij} = L(K_i, K_j)$.

Заключение (Conclusion)

Рассмотрены следующие методы анализа учебных текстов: 1) выявление ключевых понятий параграфа (раздела, курса); 2) изучение распределения терминов в тексте; 3) построение фазовой кривой изучения дисциплины; 4) установление семантической близости между разделами курса; 5) определение семантического сходства между научными понятиями, изучение семантического пространства; 6) кластеризация учебных понятий. Их использование позволяет выявить ключевые темы и основные идеи учебного

текста, установить ключевые понятия и связи между ними, оценить уровень сложности курса, разнообразие терминологии, выявить пробелы и перекосы в обучении, связанные с недостаточным освещением важных вопросов, идентифицировать меж- и внутрипредметные связи, диагностировать стиль изложения и ясность материала. Всё это помогает сформировать более глубокий взгляд на содержание, структуру и эволюцию учебных курсов, оптимизировать учебники и учебно-методические пособия, повысить эффективность методики обучения.

1. Андриевская Н. К. Гибридная интеллектуальная мера оценки семантической близости // Проблемы искусственного интеллекта. 2021. № 1. С. 4–17.
2. Бермудес С. Х. Метод измерения семантического сходства текстовых документов // Изв. Юж. федер. ун-та. Технические науки. 2017. № 3. С. 17–29. DOI: 10.23683/2311-3103-2017-3-17-29
3. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. № 19. С. 85–93.
4. Гнитецкая Т. Н. Основы теории внутрипредметных связей // Физическое образование в вузах. 1999. № 2. С. 23–39.
5. Кузнецов Л. А., Кузнецова В. Ф. Технология автоматизированной оценки содержательной близости текстов // Программные продукты и системы. 2013. № 1. С. 34–42.
6. Майер Р. В. Сложность учебных понятий и текстов : моногр. Глазов : Глазов. гос. инженер.-пед. ун-т, 2024. 132 с.
7. Анализ терминологии в современных школьных учебниках методами компьютерной лингвистики / С. И. Монахов, В. В. Турчаненко, Е. А. Федюкова, Д. Н. Чердаков // Информатизация образования и методика электронного обучения: цифровые технологии в образовании : материалы V Междунар. науч. конф. : в 2 ч. Красноярск : Сиб. федер. ун-т, 2021. Ч. 2. С. 209–215.
8. Морозова Ю. И. Построение семантических векторных пространств различных предметных областей // Информатика и ее применения. 2013. Т. 7, вып. 1. С. 90–93.
9. Ракитина С. В. Концептосфера и семантическое пространство научного текста // Альманах современной науки и образования. 2009. № 8–1. С. 125–126.
10. Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. New York : Cambridge University Press, 2008. 482 p.
11. Трофимова Т. И. Курс физики : учеб. пособие для вузов. М. : Высш. шк., 2001. 542 с.
12. Кузнецов С. И. Курс физики с примерами решения задач : учеб. пособие : в 3 ч. Ч. III : Геометрическая и волновая оптика. Элементы атомной и ядерной физики. Основы физики элементарных частиц. Томск : Изд-во Том. политех. ун-та, 2015. 302 с.
13. Википедия: свободная энциклопедия : [сайт]. URL: ru.wikipedia.org (дата обращения: 06.09.2025).
14. Воронцов-Вельяминов Б. А., Страут Е. К. Астрономия 11 кл. : учеб. для общеобразоват. учеб. заведений. М. : Дрофа, 2003. 224 с.
15. Галузо И. В., Голубев В. А., Шимбалёв А. А. Астрономия : учеб. для 11-го кл. учреждений общ. сред. образования с рус. яз. обучения. Минск : Адукацыя і выхаванне, 2015. 224 с.