УДК 81+004.048

СЕМАНТИЧЕСКИЙ РЕКОМЕНДАТЕЛЬНЫЙ СЕРВИС ПРИСВОЕНИЯ КОДА УДК МАТЕМАТИЧЕСКИМ СТАТЬЯМ

О. А. Невзорова[1] [0000-0001-8116-9446], Д. А. Альмухаметов[2] [0000-0002-4888-7937]

[1,2]*Казанский (Приволжский) федеральный университет, ул. Кремлевская, 35, г. Казань, 420008*

[1]onevzoro@gmail.com, [2]dnlanik@gmail.com

Аннотация

Классификация документов с присвоением кодов-классификаторов является традиционным способом систематизации и поиска документов по определенной тематике. Универсальная десятичная классификация (УДК) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. В России УДК является обязательным реквизитом всей книжной продукции и информации по естественным и техническим наукам. Выбор классификационных кодов связан с анализом структуры дерева классификатора и традиционно выполняется автором научной статьи.

В настоящей работе предложено решение задачи автоматизации подбора классификационного кода УДК для математической статьи на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики, разработанной в Казанском федеральном университете. Подходом к решению задачи автоматизации является создание «кодовых карт» для каждого классифицирующего кода в дереве УДК в области математики. Под «кодовой картой» понимается взвешенный набор всех математических именованных сущностей, извлеченных с помощью онтологии OntoMath^{PRO} из коллекции статей с заданным кодом УДК. Создание «кодовых карт» основано на гипотезе о том, что выбор кода УДК обуславливается определённым набором классифицирующих признаков, которые можно представить классами из онтологии OntoMath^{PRO}. Предложенная гипотеза проверена и подтверждена: проверка гипотезы проведена на коллекции математических статей, опубликованных в журнале «Известия ВУЗов. Математика» в течение 1999–2009 гг.

***Ключевые слова**: Универсальная десятичная классификация, кодовая карта, онтология OntoMath^{PRO}, математическая статья*

ВВЕДЕНИЕ

В настоящее время рекомендательные системы используются в самых разных областях, выработаны основные подходы к их построению [1, 2]. Особый интерес представляют рекомендательные системы, ориентированные на издание и подготовку научных публикаций [3]. Такие системы формируют цифровую инфраструктуру электронных научных журналов, включающую программную платформу, реализующую основные рабочие процессы управления электронным журналом, и информационные системы, поддерживающие базовые и дополнительные сервисы с учетом, в частности, специфики предметной области этого журнала [4].

Классификация документов с присвоением кодов-классификаторов является традиционным способом систематизации и поиска знаний. Классификаторы – это тип метаданных в научных документах. Существуют различные национальные и международные универсальные системы классификации. В России широко используются такие классификационные системы, как Библиотечно-библиографическая классификация (ББК), Государственный рубрикатор научно-технической информации (ГРНТИ) и Универсальная десятичная классификация (УДК).

УДК (https://udcc.org) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. Эта классификация принята в качестве основной системы индексации научно-технической документации в большинстве стран мира. В России УДК является обязательным реквизитом для всей книжной продукции и информации по естественным и техническим наукам. В конце 2019 года данный классификатор содержал порядка 126 441 кодов. В настоящее время классификация переведена более чем на 50 языков.

Выбор классификационных кодов связан с анализом структуры дерева классификатора и занимает достаточно много времени. Ниже рассмотрена задача автоматизации подбора кода классификации УДК для математических статей из области УДК 51 «Математика» на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики.

Смежные работы

Классификация научных текстов в соответствии с УДК основывается на ключевых словах, содержащихся в тексте [5]. Точно так же библиографические метаданные, такие как заголовок, описание и тематические теги, могут использоваться для дополнения библиографических записей публикации десятичной классификацией Дью (Dewey Decimal Classification, DDC) [6]. Распространение цифровых ресурсов и их интеграции в традиционную библиотечную среду создали потребность в автоматизированном инструменте для определения тематики публикации в соответствии со схемами библиотечной классификации.

Обзор методов, таких как контентно-ориентированная и совместная фильтрация, графические и гибридные методы, можно найти в работе Bai et al. [7]. Анализ использования сервисов рекомендаций для научных кругов представлен в исследовании Bell et al. [8]. В [9] дан исчерпывающий обзор современных рекомендательный систем на основе глубокого машинного обучения. Методы машинного обучения используются в различных научных рекомендательных системах [10, 11]. В [10] авторы исследуют возможность автоматического назначения первичной классификации с использованием схемы математической предметной классификации (Mathematics Subject Classification, MSC), рассматривая проблему назначения классифицирующего кода как задачу мультиклассовой классификации машинного обучения. В [11] обсуждается модель на основе машинного обучения, предназначенная для автоматической классификации старых оцифрованных текстов из словенской цифровой библиотеки. Классификационные коды УДК новых научных работ, назначенные специалистами людьми, использовались для построения классификационной модели УДК старых оцифрованных текстов. В этой модели использовались различные алгоритмы кластеризации. Авторы названной статьи утверждают, что наиболее эффективным классификатором был SVM с использованием TF-IDF. В отличие от описанных ранее работ, в нашей работе рассмотрена задача автоматизации подбора кода классификации УДК для математических статей на основе специального ресурса – онтологии OntoMath^{PRO} профессиональной математики [12].

Онтология OntoMath^{PRO}

Онтология OntoMath^{PRO} – прикладная онтология для автоматической обработки профессиональных математических статей на русском и английском языках, разработанная в Казанском федеральном университете. Эта онтология охватывает широкий спектр областей математики, таких как теория чисел, теория множеств, алгебра, анализ, геометрия, теория вычислений, дифференциальные уравнения, численный анализ, теория вероятностей и статистика. Каждый концепт онтологии имеет аннотацию, имя на русском и английском языках, включая синонимы. Терминологическими источниками, использованными при разработке OntoMath^{PRO}, служили классические учебники, интернет-ресурсы, такие как Кембриджский математический тезаурус, статьи из научных журналов, например, журнала «Известия высших учебных заведений. Математика».

В онтологии можно выделить две таксономии по отношению ISA – иерархия областей математики и иерархия объектов математического знания. Первая иерархия близка к Универсальной десятичной классификации. Верхний уровень второй таксономии содержит понятия трех типов: 1) основные математические понятия (например, «Множество» и «Оператор»); 2) понятия, относящиеся к конкретным областям математики и заданные в соответствующих иерархиях (например, «Элемент теории вероятностей» или «Элемент численного анализа»); 3) общие научные понятия (например, «Задача», «Метод», «Утверждение», «Формула» и пр.).

Онтология OntoMath^{PRO} разработана на языке OWL-DL/RDFS и содержит в настоящий момент времени 3 450 классов, 5 свойств объектов, 3 630 экземпляров подклассов свойств и 1 140 экземпляров других свойств. Постоянно происходит дальнейшее наполнение этой онтологии.

Описание подхода

Исследование проведено на коллекции статей из выпусков, опубликованных журналом «Известия высших учебных заведений. Математика» за 10 лет (с 1999 по 2009 годы). Коллекция содержит 1 356 математических статей в формате XML. Каждая статья имеет как минимум один код УДК. В рассмотренных выпусках

наибольшее количество статей пришлось на классифицирующий код УДК 517 «Анализ», всего в коллекции оказалось 883 статьи с данным кодом.

Предлагаемый нами подход к автоматическому назначению классифицирующего кода УДК математическим статьям основан на использовании онтологии OntoMath^{PRO}. Как отмечено выше, онтология содержит базовые понятия, такие как задача, система, теория, уравнение, формула и т. д. Ключевая идея предлагаемого подхода состоит в том, что выбор классифицирующего кода УДК базируется на определенных наборах классифицирующих признаков, которые использует автор статьи. Эти признаки представлены в онтологии базовыми математическими понятиями. Задачей исследования было выделение наиболее релевантных признаков среди онтологических понятий, определяющих выбор классифицирующего кода УДК.

Нами был проведен опрос экспертов-математиков с целью выяснения, какие признаки являются для них определяющими при выборе классифицирующего кода УДК для научной статьи. В результате был сделан вывод, что наиболее значимыми признаками являются метод, задача и уравнение, что составляет содержание принятой рабочей гипотезы.

Для проверки этой гипотезы был проведен ряд экспериментов на наиболее репрезентативной подколлекции с кодом УДК 517 («Анализ») из имеющейся коллекции математических статей. В экспериментах попарно сравнивались подколлекции с разными кодами УДК. Выбор кодов был основан на их положении в иерархии дерева УДК (разные поддеревья первого уровня в кодовом дереве с корневой вершиной под номером 517), родстве (потомки одного предка) и размере подколлекций.

В экспериментах использовалась подсистема семантической аннотации, которая обеспечивала функциональные возможности для аннотирования статей с точки зрения фиксированного набора предметных областей онтологии OntoMath^{PRO}. Из текста статьи извлекались все математические именованные сущности (Mathematical Named Entity, MNE), распознаваемые онтологией, и на основе словаря онтологии составлялся вектор документа.

Для процесса оценки релевантности классификационных признаков использовался модуль фильтрации математических именованных сущностей, который получал на вход два набора подколлекций статей с разными кодами УДК и список классифицирующих признаков. Результатом работы модуля являлся набор именованных математических сущностей, отобранных на основе выбранных классифицирующих признаков, для определенных кодов УДК. Модуль оценки сравнивал два полученных набора, определяя общие и специфичные признаки для каждого кода УДК. В результате модуль определял актуальность каждого классифицирующего признака для соответствующего кода УДК.

Обозначим S(f_i, c_j) – набор выделенных именованных сущностей для статей с кодом УДК c_j, отобранных по признаку f_i. Для оценки релевантности классифицирующего признака для определенного кода УДК использовалась следующая формула

$$REL_{c_j c_k}^{f_i} = \frac{S(f_i,c_j) \cap S(f_i,c_k)}{S(f_i,c_j) \cup S(f_i,c_k)}.$$

Оценка релевантности классифицирующего признака f_i представляет собой нечеткую лингвистическую переменную со значениями «слабый», «умеренный», «сильный». Были предложены следующие экспертные правила для выявления различий/сходства в паре подколлекций.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0..0.3], то можно говорить о сильном различии в паре подколлекций УДК по данному признаку.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0.3..0.7], то пара подколлекций УДК является умеренно различимой по данному признаку.

Если значение функции оценки $REL_{c_j c_k}^{f_i}$ находится в диапазоне [0.7..1], то пара подколлекций УДК слабо различима по данному признаку.

Результаты нескольких экспериментов представлены ниже. На диаграмме показано количество общих и специфичных терминов классифицирующих признаков для пары подколлекций с выбранными кодами УДК.

В первом эксперименте были рассмотрены подколлекции с кодами УДК одного уровня и сопоставимые по размерам: УДК 517.51 «Функции действительных

переменных. Действительные функции» (89 статей), УДК 517.54 «Конформное отображение и геометрические вопросы теорий комплексного переменного. Аналитические функции и их обобщение» (87 статей), УДК 517.97 «Вариационное исчисление и математическая теория оптимального управления» (75 статей).

Результаты эксперимента показаны на рис. 1, а интерпретация этих результатов в терминах введенной нечеткой лингвистической переменной представлена на таблице 1. Синим цветом на рис. 1 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 51, серым – для коллекции с подкодом 54, а желтым – для коллекции с подкодом 97.

Рассмотрим сравнение коллекций с подкодами 51 и 54, оранжевый и серый цвета. На графике видно достаточно представительное ядро методов у этих коллекций, что можно объяснить их родством в дереве УДК. Но при этом коллекция с подкодом 54 располагает большим числом уникальных задач и уравнений, и на основе данных «экспертных классов» мы можем различать эту пару УДК.

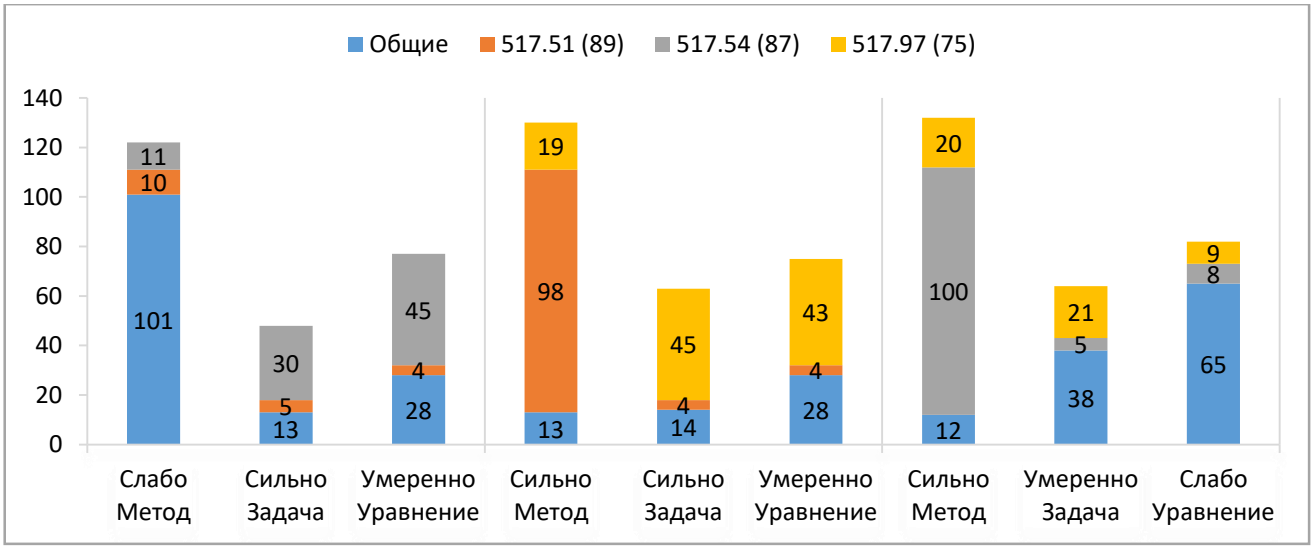


Рис. 1. Результаты эксперимента 1

	Метод	**Задача**	**Уравнение**
517.51 & 517.54	Слабо	Сильно	Умеренно
517.51 & 517.97	Сильно	Сильно	Умеренно
517.54 & 517.97	Сильно	Умеренно	Слабо

Таблица 1. Оценка релевантности классифицирующих признаков эксперимента 1

Рассмотрим коллекции с подкодами 51 и 97 (оранжевый и желтый цвета на рисунке 1). В коллекции с подкодом 97 не используется такое большое число методов, как в коллекции с подкодом 51. Но при этом в коллекции с подкодом 97, по сравнению с коллекцией с подкодом 51, преобладают задачи и уравнения. Данную пару мы можем различать по всем трем «экспертным классам».

Рассматривая пару коллекций с подкодами 54 и 97 (серый и желтый цвета на рисунке 1), мы видим такую же тенденцию по методам, как и в предыдущем сравнении. В столбце уравнений видно, что эти коллекции используют общий набор уравнений, и, следовательно, мы не можем различать данную пару по этому признаку. В задачах же преобладает коллекция с подкодом 97. Таким образом, для классификации этих коллекций можно использовать методы и задачи.

Второй эксперимент проводился между одноуровневыми подклассами одного класса УДК 517.9 «Дифференциальные, интегральные и другие функциональные уравнения. Вариационное исчисление и конечные разности», имеющими наибольшее количество статей среди подклассов (рис. 2). В эксперименте участвовали следующие коллекции: УДК 517.92 «Методы решения различных типов уравнений и систем уравнений» (192 статьи), УДК 517.95 «Дифференциальные уравнения с частными производными» (156 статей) и УДК 517.98 «Функциональный анализ и теория операторов» (133 статьи). Синим цветом на рисунке 2 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 92, серым – для коллекции с подкодом 95, а желтым – для коллекции с подкодом 98. Интерпретация результатов эксперимента согласно формуле оценки релевантности представлена в таблице 2.

Рассмотрим коллекции с подкодами 92 и 95 (оранжевый и серый цвета на рисунке 2). Пара коллекций использует общий набор уравнений и не может классифицироваться по этому признаку. В коллекции с подкодом 95 преобладают методы и задачи, и по этим «экспертным классам» мы можем различать данную пару.

Пару коллекций с подкодами 92 и 98 (оранжевый и желтый цвета на рисунке 2) мы можем уверенно различать только по методам, поскольку они используют общие наборы задач и уравнений.

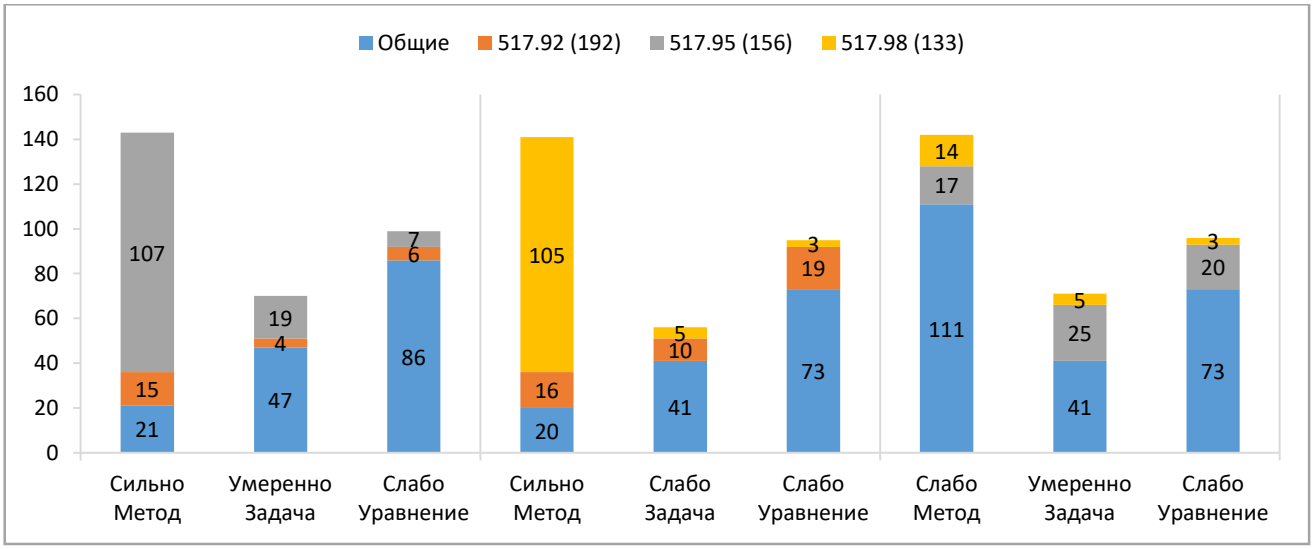


Рис. 2. Результаты эксперимента 2

	Метод	**Задача**	**Уравнение**
517.92 & 517.95	Сильно	Умеренно	Слабо
517.92 & 517.98	Сильно	Слабо	Слабо
517.95 & 517.98	Слабо	Умеренно	Слабо

Таблица 2. Оценка релевантности классифицирующих признаков эксперимента 2

Перейдем к паре коллекций с подкодами 95 и 98 (серый и желтый цвета на рисунке 2). Статьи с такими УДК используют общие наборы методов и уравнений и могут классифицироваться только по задачам.

Коллекции статей с кодами УДК одного предка обладают большим количеством общих представителей экспертных классов, что объясняется их родством в дереве УДК, тем не менее, мы все еще можем их различать.

В третьем эксперименте были рассмотрены классифицирующие коды УДК узкоспециализированной направленности: УДК 517.956 «Линейные и квазилинейные уравнения и системы» (57 статей), УДК 517.958 «Дифференциальные и интегральные уравнения математической физики» (59 статей), УДК 517.982 «Линейные пространства, снабженные топологией, порядком и другими структурами» (21 статья) и УДК 517.983 «Линейные операторы и операторные уравнения» (36 статей). Светло-синим цветом на рисунке 3 отображено число общих математических именованных сущностей для пары коллекций в сравнении, оранжевым – число уникальных сущностей для коллекции с подкодом 956, серым – для коллекции с подкодом 958, желтым – для коллекции с подкодом 982, а темно-синим – для коллекции с подкодом 983. Интерпретация результатов представлена в таблице 3.

По данным эксперимента видно, что все группы статей могут в той или иной степени классифицироваться по «экспертным классам». Малое количество извлеченных концептов «экспертных классов» в коллекциях с кодами 982 и 983 можно связать с недостаточным их представительством в нашей коллекции статей.

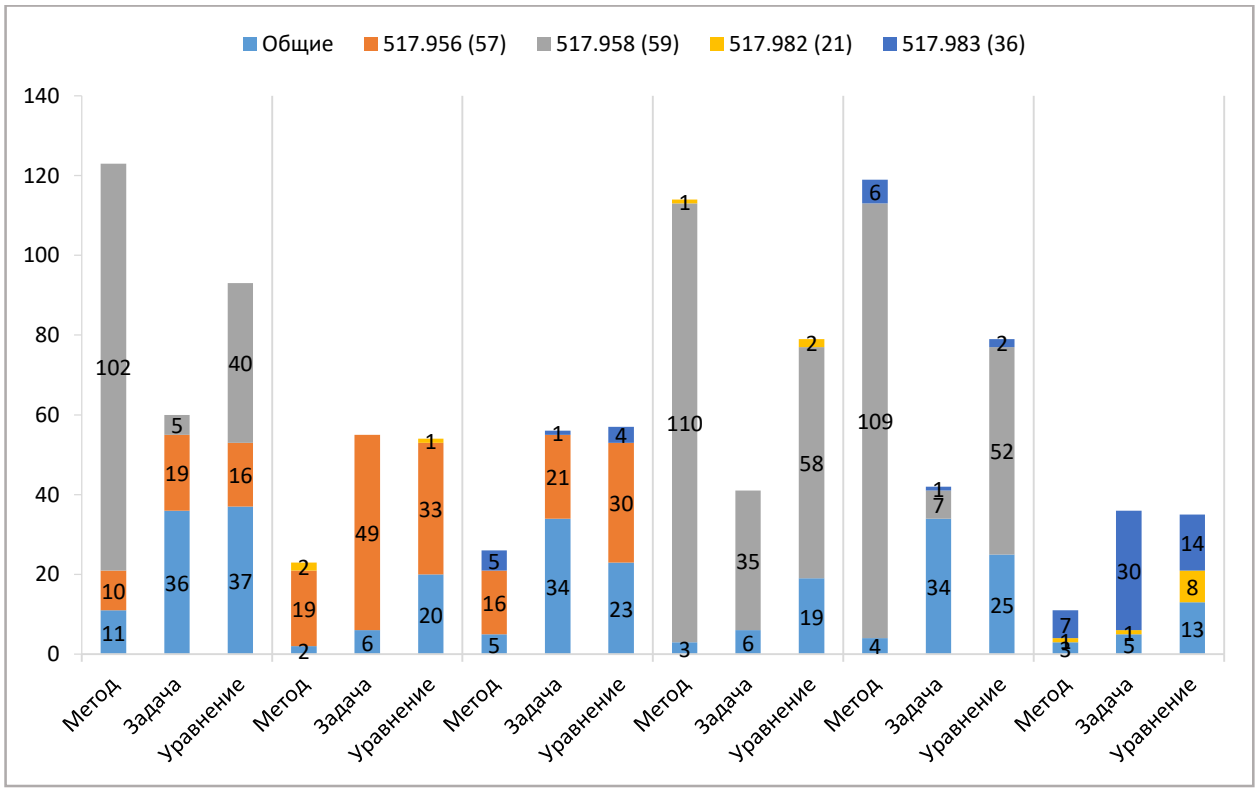


Рис. 3. Результаты эксперимента 3

	Метод	**Задача**	**Уравнение**
517.956 & 517.958	Сильно	Умеренно	Умеренно
517.956 & 517.982	Сильно	Сильно	Умеренно
517.956 & 517.983	Сильно	Умеренно	Умеренно
517.958 & 517.982	Сильно	Сильно	Сильно
517.958 & 517.983	Сильно	Слабо	Сильно
517.982 & 517.983	Умеренно	Сильно	Умеренно

Таблица 3. Оценка релевантности классифицирующих признаков эксперимента 3

Проведенное нами исследование подтверждает предложенную гипотезу о том, что группу математических кодов УДК можно классифицировать по таким признакам, как «метод», «задача» и «уравнение».

Основываясь на результатах проверки гипотезы, представляется перспективным создание «кодовых карт» для каждого кода УДК в области «Математика». Под кодовой картой мы подразумеваем взвешенный набор всех извлеченных именованных математических сущностей из подколлекции статей с определенным кодом УДК.

Кодовая карта

Кодовая карта строится на основе словаря онтологии OntoMath^{PRO}. На рисунке 4 представлена иерархия онтологии «Элемент математического знания», которая включает такие общие концепты, как *величина, геометрический объект, гипотеза, задача, метод, множество, неравенство, оператор, операция, отображение, оценка, преобразование, равенство, тензор, теорема, уравнение, утверждение, формула, характеристика* и др.

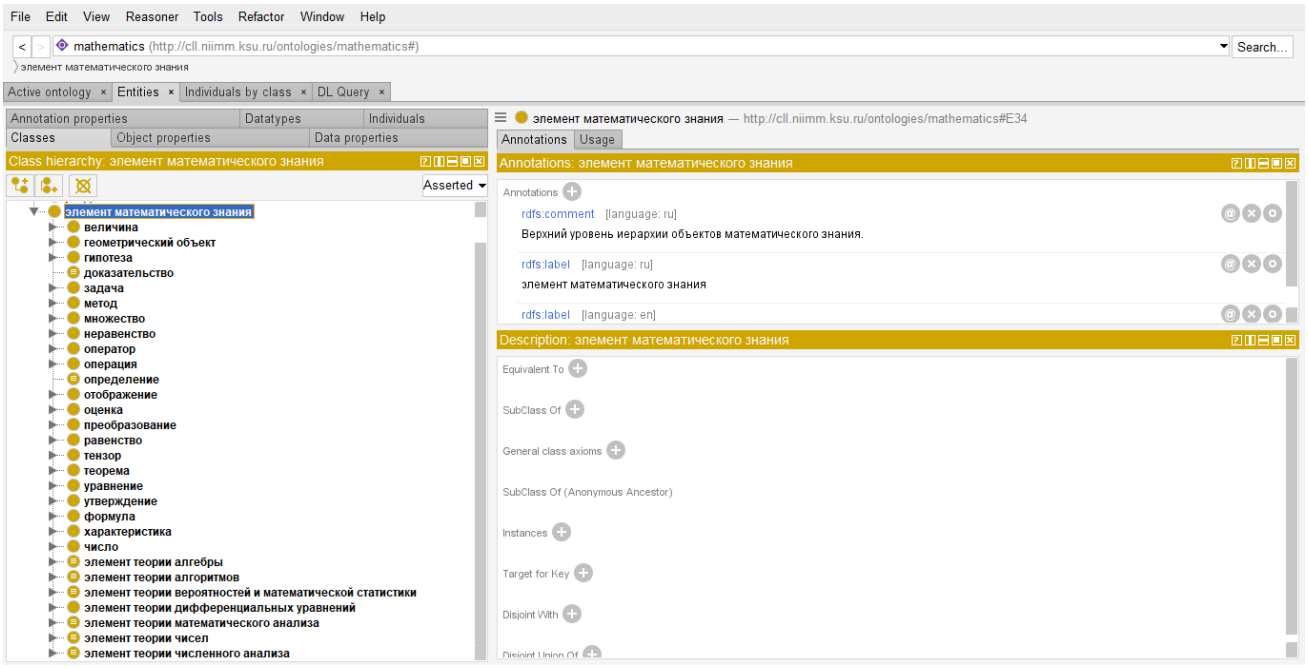
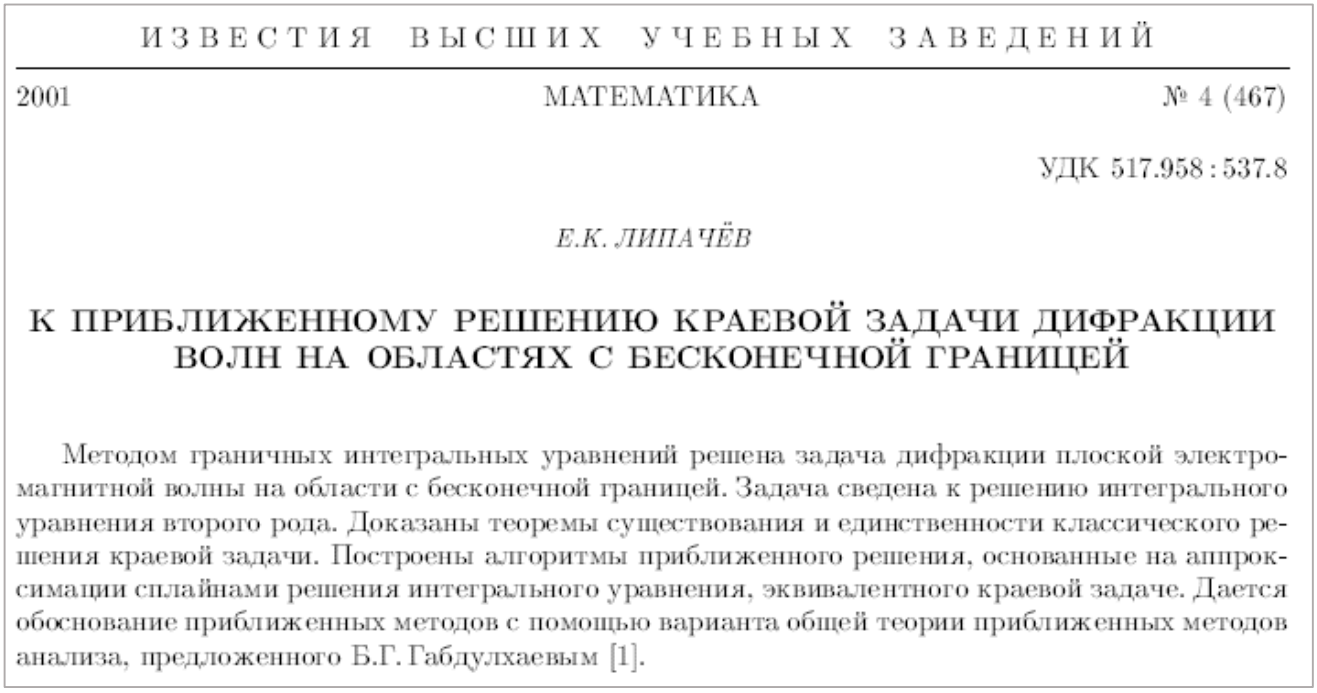


Рис. 4. Иерархия «Элементы математического знания» в онтологии OntoMath^{PRO}

Всего онтология OntoMath^{PRO} содержит сегодня 3 450 классов, 5 свойств объектов, 3 630 экземпляров подклассов свойств и 1 140 экземпляров других свойств. Например, класс *геометрический объект* содержит 333 подкласса, класс *задача* – 125 подклассов, а класс *метод* – 500 подклассов.

В рекомендательной системе предлагается использовать общий шаблон для формирования кодовых карт кодов УДК и карт статей. Шаблон содержит 2739 термов из 22 основных класса из иерархии элементов математического знания. Оценка близости статьи к определенному коду УДК происходит посредством нормировки карты статьи и сравнивания её с кодовой картой УДК.



ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

2001 МАТЕМАТИКА № 4 (467)

УДК 517.958 : 537.8

Е.К. ЛИПАЧЁВ

К ПРИБЛИЖЕННОМУ РЕШЕНИЮ КРАЕВОЙ ЗАДАЧИ ДИФРАКЦИИ ВОЛН НА ОБЛАСТЯХ С БЕСКОНЕЧНОЙ ГРАНИЦЕЙ

Методом граничных интегральных уравнений решена задача дифракции плоской электромагнитной волны на области с бесконечной границей. Задача сведена к решению интегрального уравнения второго рода. Доказаны теоремы существования и единственности классического решения краевой задачи. Построены алгоритмы приближенного решения, основанные на аппроксимации сплайнами решения интегрального уравнения, эквивалентного краевой задаче. Дается обоснование приближенных методов с помощью варианта общей теории приближенных методов анализа, предложенного Б.Г. Габдулхаевым [1].

Рис. 5. Фрагмент статьи из журнала «Известия высших учебных заведений. Математика».

Для классификации в качестве примера приведем статью Е.К. Липачёва «К приближенному решению краевой задачи дифракции волн на областях с бесконечной границей» [13]. Автор в своей статье указал два кода УДК: УДК 517.958 «Дифференциальные и интегральные уравнения математической физики» и УДК 537.8 «Электромагнетизм. Электромагнитное поле. Электродинамика. Теория Максвелла» (рис. 5).

Рекомендательная система также назначила статье код УДК 517.958. Выделим классификационные признаки, послужившие основанием для отнесения к конкретному классу УДК, и сравним две близкие кодовые карты в дереве классификатора.

Рассмотрим кодовые карты для кода УДК 517.956 «Линейные и квазилинейные уравнения и системы» и кода УДК 517.958, которые являются наследниками кода УДК 517.95 «Дифференциальные уравнения с частными производными».

Статистика по экспертным классам «методы», «задачи» и «уравнения», термы которых содержатся в кодовых картах, а также данные о пересечении списков термов из статьи и кодовых карт по этим классам приведена в таблице 4.

	Метод	**Задача**	**Уравнение**
517.956	59	51	37
517.958	75	51	29
Статья ∩ 517.956	5	1	5
Статья ∩ 517.958	9	3	5

Таблица 4. Статистика по классификационным термам из экспертных классов

В данном примере в пересечении списков термов из статьи и кодовой карты УДК 517.958 содержится 5 термов, которые входят в набор термов пересечения из статьи и кодовой карты УДК 517.956. Множество термов пересечения включает такие общие термы, как «вычислительная схема», «метод», «анализ», «спектральный метод» и «метод интегральных уравнений». Дополнительными классифицирующими термами для УДК 517.958 служат еще 4 терма: «метод граничных интегральных уравнений», «метод обобщенных потенциалов», «метод коллокаций» и «метод сплайн-коллокаций». Термы для класса уравнений совпадают для указанных кодов УДК, выделены термы «уравнение», «уравнение Фредгольма», «уравнение Фредгольма первого рода», «уравнение Фредгольма второго рода» и «уравнение Гельмгольца». По классу «задача» выделены общий терм «задача» и дополнительные термы по пересечению статьи и кодовой карты УДК 517.958 «задача численного решения интегральных уравнений» и «задача численного решения интегральных уравнений Фредгольма второго рода».

Реализация рекомендательной системы

Для реализации рекомендательной системы были выбраны высокоуровневый язык программирования общего назначения *Python* и свободный фреймворк для веб-приложений *Django*. В качестве СУБД использовалась *SQLite*. Для обработки загружаемых в систему научных статей в реальном времени применен менеджер задач с открытым исходным кодом *Celery*. В качестве брокера сообщений выбран *Redis*. В данный момент рекомендательная система работает с файлами в формате *PDF*. Для извлечения текста из статьи использован инструмент с открытым исходным кодом для оптического распознавания символов на основе нейронной сети *Tesseract OCR*.

На рис. 6 приведен интерфейс личного кабинета, в котором пользователь может вводить свои данные, а также увидеть статус обработки статьи и рекомендации по выбору классифицирующего кода УДК для загруженной в систему статьи. Система способна давать рекомендации по уточнению кода УДК (пример 1), правильно классифицировать код УДК статьи (примеры 2 и 3) или корректно определять общую тематическую направленность работы (пример 4).

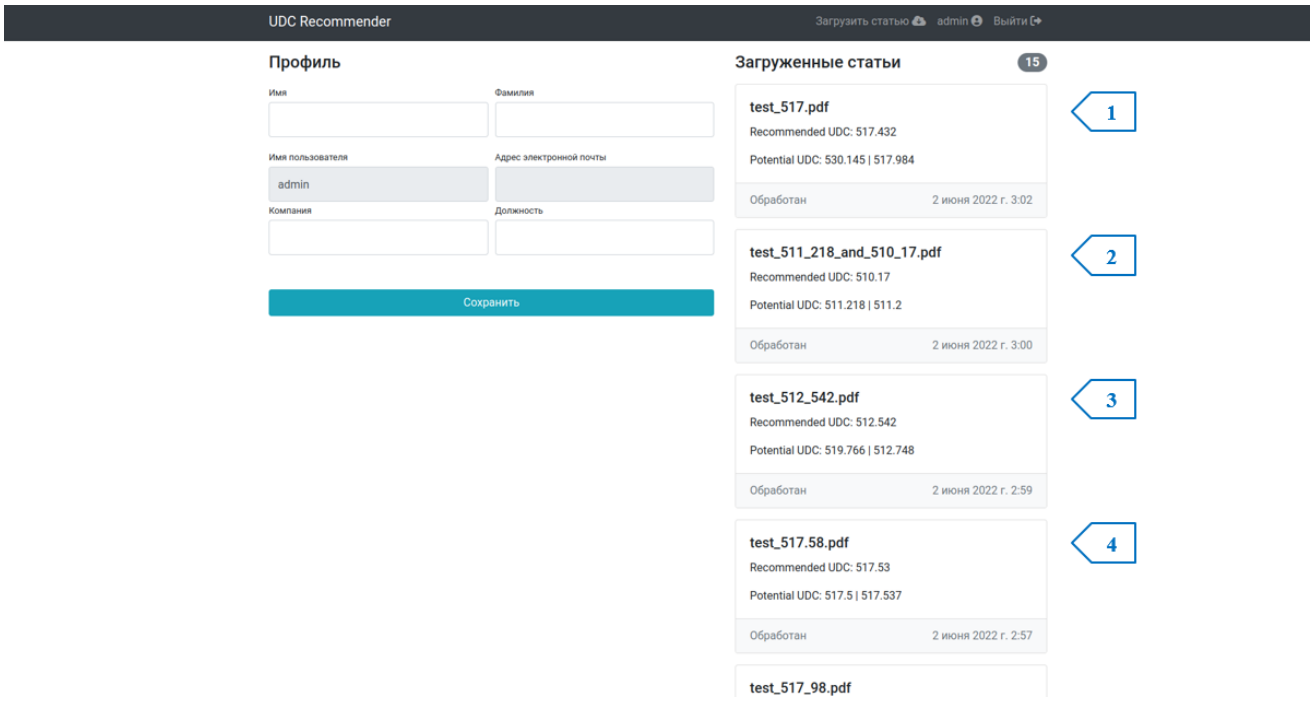


Рис. 6. Личный кабинет пользователя

Для разработки текущей версии рекомендательной системы использовалась коллекция статей журнала «Известия высших учебных заведений. Математика» за 50 лет (1968–2018 г.г.). Коллекция содержит более 6000 статей с назначенными классифицирующими кодами УДК. Статьям присвоены 622 различных кода УДК, из них 564 приходятся на раздел 51 «Математика», в котором имеется 1660 классификационных кодов. Процент успеха классификации варьируется от 30 до 80 процентов в зависимости от размера обучающей подколлекции и узкоспециализированной направленности кода УДК. Предполагается, что в дальнейшем к системе будут подключены внешние источники загрузки статей для увеличения размера обучающей коллекции и повышения качества классификации.

В таблице 5 приведены данные о качестве классификации для различных кодов УДК, находящихся на разных уровнях в иерархии дерева УДК. Здесь указаны код УДК, количество статей, содержащихся в подколлекциях с данным кодом УДК, и процент успешных классификаций наборов тестовых статей с данным кодом. Классификация считалась успешной, если хотя бы один из трех рекомендованных кодов совпадал с кодом статьи или уточнял его.

УДК	**Кол-во статей**	**Процент успеха**
510	48	84 %
511	67	87 %
512	472	74 %
514	443	64 %
515	54	76 %
517.51	540	47 %
517.54	372	58 %
517.97	150	42 %
517.98	470	53 %
517.512	156	52%
517.518	213	36%
517.544	212	47%
517.929	122	37%
517.956	183	64%
517.968	126	26%
517.983	121	47%

Таблица 5. Оценка качества классификации тестовых наборов

В подколлекциях с высоким уровнем в иерархии УДК – 510 «Фундаментальные и общие проблемы математики», 511 «Теория чисел», 512 «Алгебра», 514

«Геометрия» и 515 «Топология» – процент успешной классификации высок, поскольку названные области математики сильно отличаются по терминологии, и их легко отличить на основе словаря онтологии OntoMath^{PRO}.

При переходе на более низкий уровень иерархии УДК, в частности, на примере коллекции 517 «Анализ», процент успеха классификации снижается, поскольку тексты имеют более близкую направленность, и сложность классификации возрастает.

При переходе ниже в иерархии УДК, как ожидалось, процент успеха также снижается, несмотря на достаточно представительный размер подколлекций. Снижение не наблюдается у подколлекции 517.956 «Линейные и квазилинейные уравнения и системы», это, скорее всего, связано со спецификой тематики подколлекции. Самое большое снижение наблюдается у подколлекции 517.968 «Интегральные уравнения», поскольку термы из данной тематики широко применяются в смежных коллекциях.

В настоящее время проводятся дополнительные исследования для определения наиболее подходящей рекомендательной модели и выбора соответствующих весов для классифицирующих признаков.

Заключение

В статье представлены результаты разработки рекомендательной системы, ориентированной на автоматическое присвоение кодов УДК научным статьям в области УДК 51 «Математика». Решение задачи автоматизации подбора кода УДК для математической статьи основано на специальном ресурсе – онтологии OntoMath^{PRO} профессиональной математики. Подходом к решению задачи автоматизации является создание кодовых карт для каждого кода в дереве УДК в области математики. Под кодовой картой подразумевается взвешенный набор всех математических именованных сущностей, извлеченных с помощью онтологии OntoMath^{PRO} из коллекции статей с заданным кодом УДК. Создание кодовых карт основано на гипотезе о том, что выбор кода УДК обусловлен определённым набором классифицирующих признаков, в качестве которых могут выступать классы математических именованных сущностей, выбранных из онтологии.

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда, проект № 21-11-00105.

СПИСОК ЛИТЕРАТУРЫ

1. *Lu J., Wu D., Mao M., Wang W., Zhang G.* Recommender system application developments: A survey // Decision Support Systems. 2015. V. 74. P. 12–32.

2. *Ricci F.* Recommender Systems: Models and Techniques // Encyclopedia of Social Network Analysis and Mining. Springer: 2014, P. 1511–1522. https://doi.org/10.1007/978-1-4614-6170-8_88

3. *Elizarov A.M., Lipachev E.K.* Methods of processing large collections of scientific documents and the formation of digital mathematical library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

4. *Elizarov A.M., Lipachev E.K.* Big Math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

5. *Romanov A.Y., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L.* Research of neural networks application efficiency in automatic scientific articles classification according to UDC // 2016 International Siberian Conference on Control and Communications (SIBCON), Moscow, Russia, 12–14 May, 2016. IEEE: 2016, P. 7–11. https://doi.org/10.1109/SIBCON.2016.7491783

6. *Khoo M.J., Ahn J.W., Binding C., Jones H.J., Lin X., Massam D., Tudhope D.* Augmenting Dublin core digital library metadata with Dewey decimal classification // Journal of Documentation. 2015. V. 71. No. 5. P. 976–998.

7. *Bai X., Wang M., Lee I., Yang Z., Kong X., Xia F.* Scientific paper Recommendation: a survey // IEEE Access. IEEE. 2019. V. 7. P. 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

8. *Beel J., Aizawa A., Breitinger C., Gipp B.* Mr. DLib: recommendations-as-a-service (RaaS) for academia // Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2017.

9. *Zhang S., Yao L., Sun A., Tay Y.* Deep Learning Based Recommender System: A Survey and New Perspectives // ACM Computing Surveys. 2019. V. 52(1). P. 1–38.

10. *Schubotz M.* et al. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020.

11. *Kragelj M., Kljajić Borštnar M.* Automatic classification of older electronic texts into the Universal Decimal Classification–UDC // Journal of Documentation. 2021. V. 77. No. 3.

12. *Nevzorova O.A., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* OntoMath^{PRO} Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Springer: 2014. V. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4

13. *Липачёв Е.К.* К приближенному решению краевой задачи дифракции волн на областях с бесконечной границей // Изв. Вузов. Математика. 2001. № 4 (467). С. 69–72.

SEMANTIC RECOMMENDATION SERVICE FOR ASSIGNING UDC CODE TO MATHEMATICAL ARTICLES

O. A. Nevzorova[1] [0000-0001-8116-9446], D. A. Almukhametov[2] [0000-0002-4888-7937]

[1,2]*Kazan (Volga Region) Federal University, 35 Kremlyovskaya str., Kazan, 42008*
[1]onevzoro@gmail.com, [2]dnlanik@gmail.com

Abstract

Classification of documents with the assignment of classifier codes is a traditional way of systematizing and searching for documents on a specific topic. The Universal Decimal Classification (UDC) underlies the systematization of knowledge presented in libraries, databases and other information repositories. In Russia, UDC is an obligatory attribute of all book production and information on natural and technical sciences. The choice of classification codes is associated with the analysis of the structure of the classifier tree and is traditionally decided by the author of a scientific article. This article proposes a solution for automating the assigning the UDC classification code for a mathematical article based on a special resource – the OntoMath^{PRO} ontology for professional mathematics, developed at Kazan Federal University. An approach

to solving the problem is to create "code maps" for each classifying code in the UDC tree in the field of mathematics. Under the "code map" is meant a weighted set of all extracted, with the help of OntoMath^{PRO} ontology, mathematical named entities from the collection of articles with a given UDC code. The creation of "code maps" is based on the hypothesis that the choice of the UDC code is determined by a certain set of classifying features that can be represented by classes from the OntoMath^{PRO} ontology. The proposed hypothesis was tested and confirmed in the paper. The hypothesis was tested on a collection of mathematical articles An approach to solving the problem is to create "code maps" for each classifying code in the UDC tree in the field of mathematics. Under the "code map" is meant a weighted set of all extracted, with the help of OntoMath^{PRO} ontology, mathematical named entities from the collection of articles with a given UDC code. The creation of "code maps" is based on the hypothesis that the choice of the UDC code is determined by a certain set of classifying features that can be represented by classes from the OntoMath^{PRO} ontology. The proposed hypothesis was tested and confirmed in the paper. The hypothesis was tested on a collection of mathematical articles published during 1999-2009 in the "Izvestiya VUZov. Mathematics" journal.

Keywords*: the Universal Decimal Classification, code map, the OntoMath^{PRO} ontology, mathematical article*

REFERENCES

1. *Lu J., Wu D., Mao M., Wang W., Zhang G.* Recommender system application developments: A survey // Decision Support Systems. 2015. V. 74. P. 12–32.

2. *Ricci F.* Recommender Systems: Models and Techniques // Encyclopedia of Social Network Analysis and Mining. Springer: 2014, P. 1511–1522. https://doi.org/10.1007/978-1-4614-6170-8_88

3. *Elizarov A.M., Lipachev E.K.* Methods of processing large collections of scientific documents and the formation of digital mathematical library // CEUR Workshop Proceedings. 2020. V. 2543. P. 354–360.

4. *Elizarov A.M., Lipachev E.K.* Big Math methods in Lobachevskii-DML digital library // CEUR Workshop Proceedings. 2019. V. 2523. P. 59–72.

5. *Romanov A.Y., Lomotin K.E., Kozlova E.S., Kolesnichenko A.L.* Research of neural networks application efficiency in automatic scientific articles classification according to UDC // 2016 International Siberian Conference on Control and Communications (SIBCON), Moscow, Russia, 12–14 May, 2016. IEEE: 2016, P. 7–11. https://doi.org/10.1109/SIBCON.2016.7491783

6. *Khoo M.J., Ahn J.W., Binding C., Jones H.J., Lin X., Massam D., Tudhope D.* Augmenting Dublin core digital library metadata with Dewey decimal classification // Journal of Documentation. 2015. V. 71. No. 5. P. 976–998.

7. *Bai X., Wang M., Lee I., Yang Z., Kong X., Xia F.* Scientific paper Recommendation: a survey // IEEE Access. IEEE. 2019. V. 7. P. 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

8. *Beel J., Aizawa A., Breitinger C., Gipp B.* Mr. DLib: recommendations-as-a-service (RaaS) for academia // Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2017.

9. *Zhang S., Yao L., Sun A., Tay Y.* Deep Learning Based Recommender System: A Survey and New Perspectives // ACM Computing Surveys. 2019. V. 52(1). P. 1–38.

10. *Schubotz M.* et al. AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels // Proceedings of the 13th Conference on Intelligent Computer Mathematics. 2020.

11. *Kragelj M., Kljajić Borštnar M.* Automatic classification of older electronic texts into the Universal Decimal Classification–UDC // Journal of Documentation. 2021. V. 77. No. 3.

12. *Nevzorova O.A., Zhiltsov N.G., Kirillovich A.V., Lipachev E.K.* OntoMath^{PRO} Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Springer: 2014. V. 468. P. 105–119. https://doi.org/10.1007/978-3-319-11716-4

13. *Lipachev E.K.* Approximation solution of the boundary value problem of wave diffraction on domain with infinite boundary // Izv. VUZ. Mathematics. 2001. No. 4 (467). P. 69–72.

СВЕДЕНИЯ ОБ АВТОРАХ

НЕВЗОРОВА Ольга Авенировна – доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского федерального университета, к. т. н. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Olga Avenirovna NEVZOROVA – Kazan Federal University, Institute of Computational Mathematics and Information Technologies, Associated Professor of the Department of Information System, PhD. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: onevzoro@gmail.com
ORCID: 0000-0001-8116-9446

АЛЬМУХАМЕТОВ Дамир Альбертович – инженер кафедры программной инженерии Института информационных технологий и интеллектуальных систем Казанского федерального университета. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Damir Albertovich ALMUKHAMETOV – Engineer of the Department of Software Engineering of the Institute of Information Technology and Information Systems of Kazan Federal University. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: dnlanik@gmail.com
ORCID: 0000-0002-4888-7937

Материал поступил в редакцию 6 февраля 2023 года