Сервис назначения кода УДК математическим статьям на основе семантических технологий

Д.А. Альмухаметов[1], О.А. Невзорова[1]

[1] Казанский федеральный университет

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 решается автором научной статьи. В статье предложено решение задачи автоматизации подбора классификационного кода УДК для математической статьи на основе специального ресурса – онтологии OntoMathPro для 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 извлеченных, с помощью онтологии OntoMathPro, математических именованных сущностей 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 OntoMathPro. Предложенная гипотеза проверена и подтверждена в статье. Проверка гипотезы проводилась на коллекции математических статей, опубликованных в журнале «Известия ВУЗов. Математика» в течение 1999-2009 гг.

Ключевые слова: Универсальная десятичная классификация, кодовая карта, онтология OntoMathPro, математическая статья

Service for assigning a UDC code to mathematical articles based on semantic technologies

D.A. Almukhametov[1], O.A. Nevzorova[1]

[1] *Kazan federal university*

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 OntoMathPro 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 OntoMathPro 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 OntoMathPro 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 OntoMathPro 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 OntoMathPro 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 OntoMathPro ontology, mathematical article

1. 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), Универсальная десятичная классификация (УДК).

Универсальная десятичная классификация (УДК) (https://udcc.org) лежит в основе систематизации знаний, представленных в библиотеках, базах данных и других хранилищах информации. Классификация УДК 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. В статье рассматривается задача автоматизации подбора кода классификации УДК для математических статей из области УДК 51 «Математика» на основе специального ресурса – онтологии OntoMathPro для профессиональной математики.

2. 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 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 утверждают, что наиболее эффективным классификатором был *SVM* с использованием *Tf-idf* (CA 5 0,963). В отличие от описанных ранее работ, в данной работе рассматривается задача автоматизации подбора кода классификации УДК для математических статей на основе специального ресурса – онтологии OntoMathPro для профессиональной математики [12].

3. Онтология OntoMathPro

Онтология OntoMathPro – прикладная онтология для автоматической обработки профессиональных математических статей на русском и английском языках, разработанная в Казанском федеральном университете. Онтология OntoMathPro 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, служили классические учебники, интернет ресурсы, такие как Кембриджский математический тезаурус, научные 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, относящихся к конкретным областям математики, заданные в соответствующих иерархиях (например, Элемент теории вероятностей или Элемент численного анализа); 3) общие научные понятия (например, Задача, Метод, Утверждение, Формула и пр.).

Онтология OntoMathPro разработана на языке *OWL-DL/RDFS*. Онтология содержит 3 450 классов, 5 свойств объектов, 3 630 экземпляров подклассов свойств и 1 140 экземпляров других свойств.

4. Описание подхода

Исследование проводилось на коллекции статей из сборников, опубликованных журналом «Известия высших учебных заведений. Математика» за 10 лет (с 1999 по 2009 года). Коллекция содержит 1 356 математических статей в формате *XML*. Каждая статья имеет как минимум один код УДК. В рассматриваемых сборниках наибольшее количество статей приходить на классифицирующий код УДК 517 «Анализ», всего в коллекции 883 статьи с данным кодом.

Предлагаемый в работе подход к автоматическому назначению классифицирующего кода УДК математическим статьям основан на использовании онтологии OntoMathPro. Как отмечалось 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 выделения 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 рабочей 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 основывался на их положении в иерархии дерева УДК (разные поддеревья первого уровня в кодовом дереве с корневой вершиной с 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 OntoMathPro. 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 различие 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.

Результат одного из экспериментов представлен ниже. На диаграмме показано количество общих и специфичных терминов классифицирующих признаков для пары подколлекций с выбранными кодами УДК.

В 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).

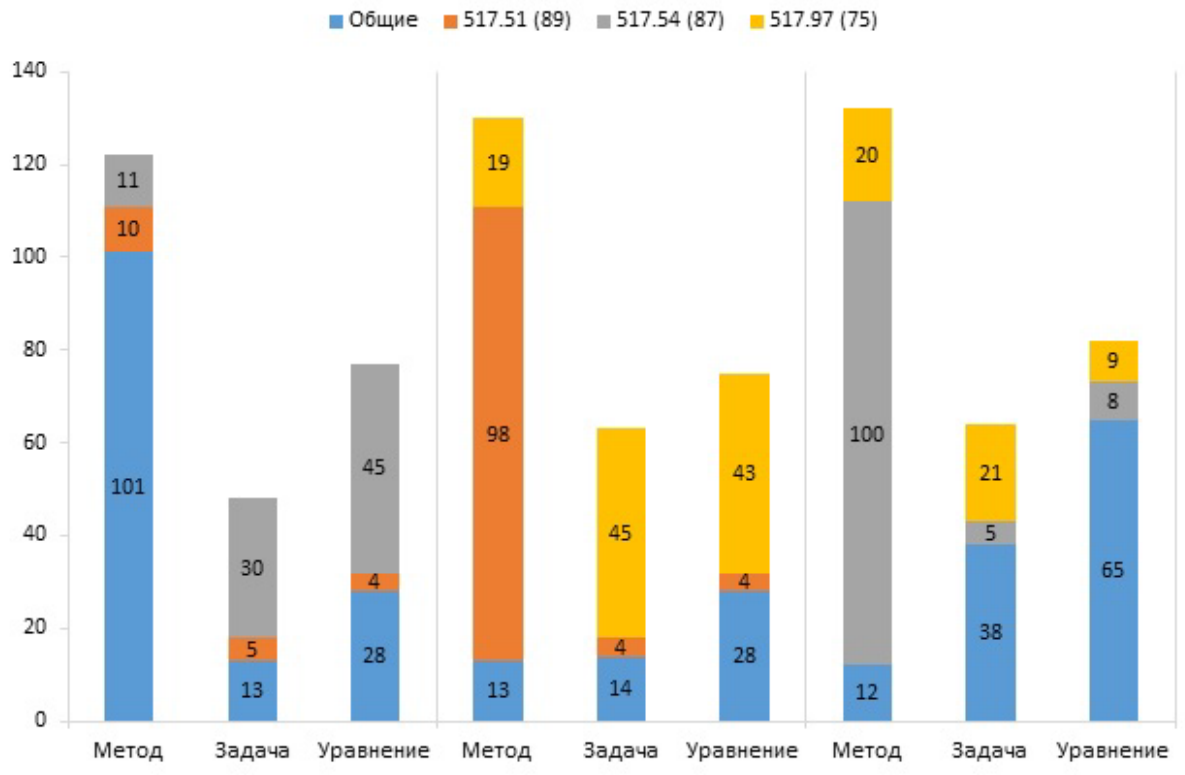

Рис. 1. Результаты эксперимента

Результаты эксперимента показаны на рис. 1, а интерпретация этих результатов в терминах введенной нечеткой лингвистической переменной представлена на рис. 2. По результатам можно сделать вывод, что для УДК 517.51 и 517.54 наиболее актуальным признаком будут методы, а для УДК 517.97 – уравнения.

	Метод	*Задача*	*Уравнение*
517.51 & 517.54	Слабый	Сильный	Умеренный
517.51 & 517.97	Сильный	Сильный	Умеренный
517.54 & 517.97	Сильный	Умеренный	Слабый

Рис. 2. Оценка релевантности классифицирующих признаков

Проведенное 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 понимается взвешенный набор всех извлеченных именованных математических сущностей из подколлекции статей с определенным кодом УДК.

5. Кодовая карта

Кодовая карта строится на основе словаря онтологии OntoMathPro. На рис. 3 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.

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

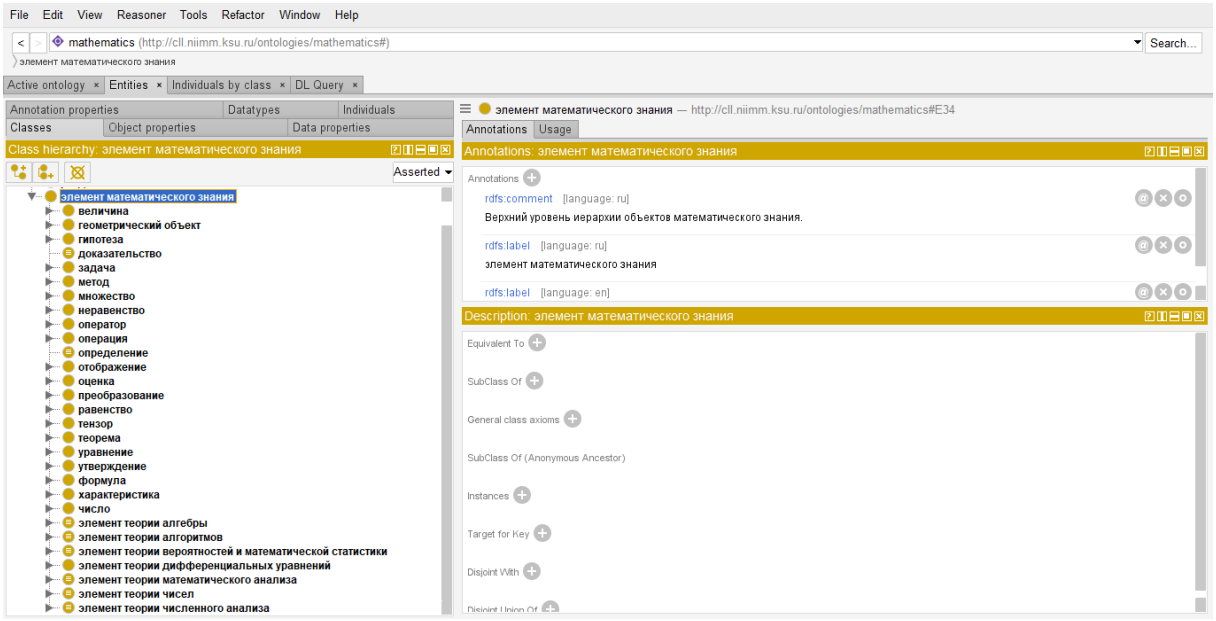

Рис. 3. Иерархия «Элементы математического знания» в онтологии OntoMathPro

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

В качестве примера классификации приведем статью Е.К. Липачёва «К приближенному решению краевой задачи дифракции волн на областях с бесконечной границей». Автор указал в статье два кода УДК: УДК

517.958 «Дифференциальные и интегральные уравнения математической физики» и УДК 537.8 «Электромагнетизм. Электромагнитное поле. Электродинамика. Теория Максвелла».

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 на рис. 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

Рис. 4. Статистика по классификационным термам из экспертных классов

В 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» выделен 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».

6. 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 используется *SQLite*. Для обработки загружаемых в систему научных статей в реальном времени применяется 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 используется инструмент с открытым исходным кодом для оптического распознавания символов на основе нейронной сети *Tesseract OCR*.

На рис. 5 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.

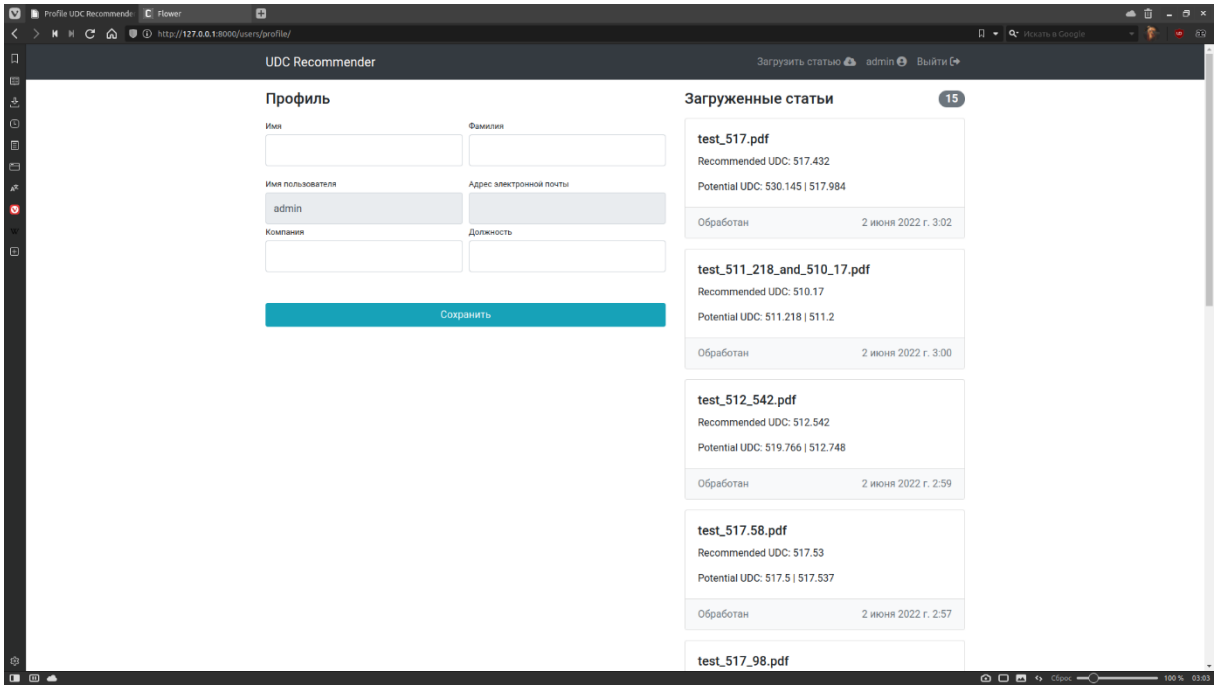

Рис. 5. Личный кабинет пользователя

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 присвоено более 600 различных кодов УДК. Процент успеха классификации варьируется от 30 до 80 процентов, в зависимости от размера обучающей подколлекции и узкой направленности кода УДК.

Предполагается, что в дальнейшем к системе будут подключены внешние источники загрузки статей, такие как электронные библиотеки,

для увеличения размера обучающей коллекции и повышения качества классификации.

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

7. Заключение

В статье представлены результаты по разработке рекомендательной системы, ориентированной на автоматическое присвоение кодов УДК научным статьям в области УДК 51 «Математика». Решение задачи автоматизации подбора кода УДК для математической статьи основан на специальном ресурсе – онтологии OntoMathPro для 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 извлеченных, с помощью онтологии OntoMathPro, математических именованных сущностей из коллекции статей с заданным кодом УДК. Создание кодовых карт основано на гипотезе о том, что выбор кода УДК обуславливается определённым набором классифицирующих признаков, в качестве которых могут выступать классы математических именованных сущностей, выбранных из онтологии.

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-11-00105).

Литература

1. Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, Guangquan Zhang: Recommender system application developments: A survey. In: Decision Support Systems, vol. 74, pp.12-32 (2015).

2. Ricci F. (2014) Recommender Systems: Models and Techniques. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY. 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. In: CEUR Workshop Proceedings, vol. 2543, pp.354–360 (2020).

4. Elizarov A, Lipachev E.: Big math methods in Lobachevskii-DML digital library. In: CEUR Workshop Proceedings, vol.2523, pp.59-72 (2019).

5. Romanov, A.Y., Lomotin, K.E., Kozlova, E.S. and Kolesnichenko, A.L.: Research of neural networks application efficiency in automatic scientific articles classification according to UDC. In: International Siberian Conference on Control and Communications, SIBCON 2016 – Proceedings, pp. 7–11 (2016).

6. Khoo, M.J., Ahn, J.W., Binding, C., Jones, H.J., Lin, X., Massam, D. and Tudhope, D.: Augmenting Dublin core digital library metadata with Dewey decimal classification. In: Journal of Documentation, vol. 71, No. 5, pp. 976–998 (2015).

7. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X. and Xia, F.: Scientific paper Recommendation: a survey. In: IEEE Access, IEEE, vol. 7, pp. 9324–9339, doi: 10.1109/ACCESS.2018.2890388 (2019).

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

9. Shuai Zhang, Lina Yao, Aixin Sun, Yi Tay: Deep Learning Based Recommender System: A Survey and New Perspectives. In:ACM Computing Surveys, 52(1), pp. 1-38 (2019).

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

11. Matjaž Kragelj, Mirjana Kljajić Borštnar: Automatic classification of older electronic texts into the Universal Decimal Classification–UDC. In: Journal of Documentation, vol. 77, no. 3 (2021).

12. Olga A. Nevzorova, Nikita Zhiltsov, Alexander Kirillovich and Evgeny Lipachev: OntoMathPro Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Proceedings. Series: Communications in Computer and Information Science, vol. 468, pp. 105–119. Springer (2014).