ОГЛАВЛЕНИЕ

УДК 004.021 + 004.623 + 004.421

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ГЕНЕРАЦИИ ВОПРОСОВ ПО ЗАДАННЫМ СУБЪЕКТАМ ПРИ ПОМОЩИ СЕМАНТИЧЕСКОЙ СЕТИ

М. Д. Андреичев[1], А. А. Ференец[2]

[1-2] Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

[1] andreichev.m@mail.ru, [2]ist.kazan@gmail.com

Аннотация

Представлен подход к автоматическому построению вопросов для тестов или викторин при помощи графа знаний DBPedia. Выбранный граф знаний имеет около 5 млн. сущностей и дает возможность делать запросы к семантической сети при помощи языка SPARQL. В статье представлены алгоритм, основные запросы к графу знаний для построения вопросов и нестандартный подход к поиску сущностей.

Ключевые слова*: семантическая сеть, генерация вопросов, связанные данные, онтология, граф знаний, RDF, SPARQL, DBPedia*

Введение

В сфере образования и обучения большую роль играют тесты. Например, в 2019 году тест ЕГЭ по истории состоит из двух частей, первая из которых включает в себя 19 вопросов с вариантами ответов. Также существуют специальные викторины для облегчения запоминания материала. При этом для 2019 года утверждено 15 общеобразовательных предметов, по которым проводится ЕГЭ. Для каждого предмета создаются десятки вариантов, что соответствует сотням вопросов для каждого предмета. Как правило, каждый вопрос в тесте или викторине строится вокруг какого-то одного понятия и предполагает простую структуру, что может означать, что вопросы можно генерировать автоматически, если иметь базу данных понятий и их соотношений. Было выдвинуто предположение, что возможно строить вопросы с вариантами ответа при помощи семантических сетей. В то же время важно не

просто создавать вопросы, но и иметь возможность выбирать тематику этих вопросов, то есть необходимо иметь возможность выбирать субъекты, по которым будут предложены вопросы.

Таким образом, была поставлена цель – разработать программный комплекс для построения вопросов по субъекту, найденному или определенному при помощи семантической сети, с сопутствующим функционалом.

1. Графы знаний

Для решения текущей задачи (создания вопросов компьютером) требуется компьютерное представление информации (приведение к определенной структуре) в виде семантической сети (рис. 1). Одно из таких решений разработано и утверждено Консорциумом Всемирной паутины: это структура (модель представления данных) — Resource Description Framework (RDF) [1].

Основная концепция RDF — триплет. Это набор «субъект, предикат,

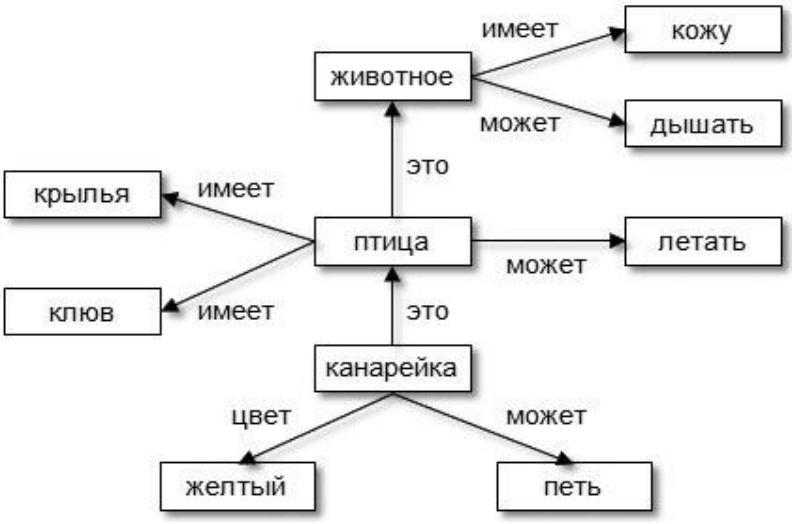


Рис. 1 — Семантическая сеть

объект» (например: дельфин – живет в – вода). В структуре RDF-ресурсом могут быть представлены любой объект реального мира или абстрактное понятие. Информацию в RDF можно просматривать в виде графа или утверждений и хранить в различных форматах: JSON, XML, N-Triples, Turtle, N3. Триплеты из RDF-документа объединяются в RDF-граф.

Связанные данные (linked data) играют все более важную роль в расширении интеллектуальных возможностей поиска в интернете. Связанные данные — это семантическая сеть, удовлетворяющая принципам, утвержденным Тимом Бёрнс-Ли [2]:

1) Использование URI в качестве идентификатора;
2) Предоставление URI, доступного по HTTP;
3) Предоставление доступа по стандартам SPARQL и RDF;
4) Предоставление по URI ссылки на другие URI.

В интернете имеются ресурсы связанных данных, в которых информация представлена в виде RDF. Одними из крупных ресурсов, в которых данные открыты в публичном доступе (linked open data или knowledge graph — граф знаний), являются Wikidata, DBPedia [3]. Сравнение Wikidata и DBPedia представлено в таблице 1.

DBpedia содержит множество ссылок на другие наборы данных в облаке LOD, таких, как Freebase, OpenCyc, GeoNames и другие. DBpedia широко использовалась в исследовательском сообществе Semantic Web, но также стала актуальной в коммерческих установках: например, такие компании, как BBC и New York Times, используют DBpedia для организации своего контента.

DBpedia берет свое начало в академических исследованиях, но также постоянно развивается благодаря сотрудничеству академических исследований и промышленности. Английская версия базы знаний DBpedia описывает 4,58 миллиона сущностей, из которых 4,22 миллиона классифицированы в последовательной онтологии, включая 1 445 000 человек, 735 000 мест (в том числе 478 000 населенных мест), 411 000 творческих работ (в том числе 123 000 музыкальных альбомов, 87 000 фильмов и 19 000 видеоигр), 241 000 организаций (включая 58 000 компаний и 49 000 учебных заведений), 251 000 видов и 6000 болезней. Это те данные, вокруг которых планируется строить запросы [4].

Таблица 1 — DBPedia и Wikidata

	Wikipedia	DBpedia	Wikidata
Сайт	wikipedia.org	dbpedia.org	www.wikidata.org
Метод генерации	Вручную/ заполнение сообществом	Автоматически или полу-автоматически	Полу-автоматически и заполнение сообществом (около 20000 редакторов)
Экземпляры, которые представляют объекты реального мира	——	4 298 433	18 697 897
Покрытие классов (сущностей >1)	——	88 %	41 %
Плюсы	Свободный текст /Легкость доступа и вклада	Большое количество интеграций в облаке LOD.	Качество (точность), редактирование сообществом, рост
Поддержка языков	294 (май 2019)	>125 (март 2019)	>358 (март 2019)

Предприятия, такие, как Apple (через Siri), Google (через Freebase и Google Knowledge Graph) и IBM (через Watson), и особенно соответствующие их проекты с высокой степенью видимости, связанные с искусственным интеллектом, получили огромную пользу от вклада DBpedia.

DBpedia также остается основой для академических занятий в областях проектирования онтологий, искусственного интеллекта, машинного обучения, обработки естественного языка и многого другого. Проект DBPedia оказался полезным для экспертов, ведущих проекты в Google, Microsoft, Facebook, Oracle, IBM, Apple и многих других.

2. Существующие решения

В рамках решения поставленной задачи был произведен обзор существующих решений, затрагивающих тематику подготовки тестов и викторин. Были найдены сервисы, которые способны создавать базу вопросов.

Многие из сервисов дают возможность проводить викторину онлайн, но вопросы в них не генерируются при помощи семантических сетей автоматически.

Также найдено приложение для мобильных устройств – программа-викторина clover quiz [5]. Она использует базу вопросов, сгенерированную с использованием семантических сетей. К сожалению, база вопросов фиксирована и не позволяет делать выгрузку. Создатель программы ссылается на программу linkeddata trivia [6]. Эта система способна сгенерировать вопрос. Но на это у приложения может уходить больше минуты. Тема вопроса не может быть задана, и вопрос строится вокруг неизвестного понятия, что делает использование этого инструмента крайне затруднительным для создания тестов и викторин на определённые темы. Для запуска этого приложения требуется запуск двух микросервисов, один на java, другой на javascript. Код этого приложения находится в открытом доступе.

При генерации вопроса в приложении linkeddata trivia к графу знаний выполняется большое количество SPARQL-запросов (около 30). Вопрос строится вокруг случайно взятой сущности. В базе более 4 миллионов реально существующих объектов, и получая случайную сущность, маловероятно сгенерировать вопрос, на который пользователь сможет ответить. Получившиеся вопросы представлены в таблице 2.

Таблица 2 – Сгенерированные вопросы

Вопрос	What is the location of 63 Building?			
Варианты ответов	Yeouido	Las Vegas	Nkawkaw	Eli, Mateh Binyamin
Время	20101,9 мс			
Вопрос	What is the location of San Bernardino Pass?			
Варианты ответов	Switzerland	Miholjsko	Svilići	Leira, Ørsta
Время	12273,5 мс			
Вопрос	What is the vein of Heart?			
Варианты ответов	Pulmonary vein	Angular vein	Deep dorsal vein of clitoris	Median antebrachial
Время	2334,2 мс			
Вопрос	What is the death place of William Warfield?			
Варианты ответов	Alumim	Chicago	Les Herbiers	Takoundou
Время	23447,1 мс			
Вопрос	What is the family of Pleuronodoceratidae?			
Варианты ответов	Xenodiscaceae	Cystoseira foeniculacea	Angaria rugosa	Scopula umbelaria
Время	23447,1 мс			

Приложение linkeddata trivia состоит из двух микросервисов. Один называется backend-core [7]. Он написан на java и запускается при помощи фреймворка Grizzly. Его роль состоит в том, чтобы получить наиболее часто встречающиеся предикаты объекта.

Другой микросервис запускается при помощи фреймворка Node js. Основной код построения вопроса находится на нем. Построение вопроса

приложение выполняет по несколько попыток (в среднем их требуется около 10). На каждую попытку приходится около 10 SPARQL-запросов. Попытка срывается в результате неподходящего ответа от DBPedia, и алгоритм запускается заново (вызывается catch метод в последовательности Promise javascript). Далее будет рассмотрена следующая последовательность построения вопроса микросервисом.

1. Подобрать любую случайную сущность из DBPedia. Для подбора случайной сущности требуется в запросе SELECT указать случайное значение offset. В программе имеется файл classes_sorted.json, в котором хранится количество сущностей на граф знаний для каждого класса. Из этого файла берутся значения, в каких пределах должно быть загадано случайное число для генерируемого объекта.

2. Сделать запрос и получить rdfs:label объекта. В нем хранится литерал с подписью объекта — текст на каком-то языке. Литерал состоит из текста литерала в кодировке Unicode и метки языка в формате RFC 3066.

3. Получить предикаты объекта (те, которые имеют литерал).

4. Сделать запрос на микросервис backend-core, получить наиболее часто встречающиеся предикаты объекта. Они хранятся в json-файле на сервере. Backend-core делает SPARQL-запрос для определения типа запрашиваемой сущности. В зависимости от типа из собственной базы возвращает значение.

5. Сделать запрос об информации о выбранном предикате, по которому будет строиться запрос. Выясняется литерал и rdfs:range (тип) объекта. На этом шаге часто прерывается алгоритм и запускается заново из-за того, что отсутствуют нужные предикаты у объекта.

6. Получить альтернативные варианты ответов. Если варианты числовые — запросов не делать и альтернативные варианты подобрать программно.

При необходимости создания вопроса на определённую тематику пользователю приходится несколько раз запускать генератор. Это происходит из-за случайного подбора предикатов, на которых требуется построить вопрос. Запросы можно усовершенствовать так, чтобы на этапе 3 заранее предусматривалось то, что будет нужно делать с субъектом на следующих этапах.

Были построены вопросы по классу dbo:Event (события) при помощи части кода системы linkeddata trivia. Вопросы получились по различным событиям: военные конфликты, спортивные события, парады и другие. Вопросы, построенные системой linkeddata trivia, представлены в таблице 3.

Таблица 3 — Вопросы, построенные linkeddata trivia

Вопрос	Ответ
Кто является командиром блокады Кисо-Фукусима?	Такэда Сингэн
Что является частью военного конфликта операции «Радуга»?	Второй Интифады
Что является частью военного конфликта битве минимойке?	Американская Революционная Война
Предыдущее событие Кубка 2013 финал Дель Рей?	2012 Копа Дель Рей финал
Каково место военного конфликта Битва Salsu?	Река Chongchon

3. Требования к программному комплексу

Получившаяся система должна удовлетворять следующим требованиям:

1) Возможность поиска сущности по онтологии (не более 10 секунд на поиск по онтологии);

2) Возможность извлечения случайной сущности, подходящей по заданному учебному предмету. Предмет определяет классы, по которым извлекать случайные сущности;

3) Возможность извлечений сущности для определенной зоны, отмеченной на карте (для поиска сущности по карте);

4) Построение вопросов по сущности;

5) Построение альтернативных (неверных) вариантов ответа.

4. Архитектура и API системы

Система реализована на языке Java с применением фреймворка Spring Web MVC, который обеспечивает архитектуру паттерна Model View Controller (MVC). Для создания http-запросов используется библиотека Apache Jena. Для генерации вопросов частично используется модифицированный код linkeddata trivia.

Большая часть системы весьма тривиальна в реализации (модель с аннотациями Spring Data Jpa, контроллеры), поэтому рассматриваться в данной работе не будет. Исходный код приложения выложен в открытый доступ: https://github.com/NGdev1/QuizEngine. Архитектура представлена на рисунке 2.

PredicatesRequestsService отвечает за запросы извлечения подходящих триплетов для построения вопроса и за построение альтернативных вариантов ответа.

ClassesRequestsService отвечает за извлечение случайных сущностей или поиск сущностей по заданным критериям, например, за поиск мест в базе, входящих в зону поиска для создания вопросов из области географии.

SparqlService объединяет эти сервисы и выполняет не только запросы, но и имеет логику их обработки. При построении альтернативных вариантов ответа SparqlService запрашивает у AlternativeAnswersHandler, требуется ли делать запрос к базе знаний для получения альтернативных (неверных) вариантов ответа. Если нет — сразу получает их.

PredicatesRequestsService и ClassesRequestsService, используя библиотеку Apache Jena, которая в свою очередь использует HttpClient, в контексте программы названный SparqlHttpClient.Jena, делает http запрос на DBPedia SPARQL точку доступа [8] или любую другую, указанную в конфигурации системы.

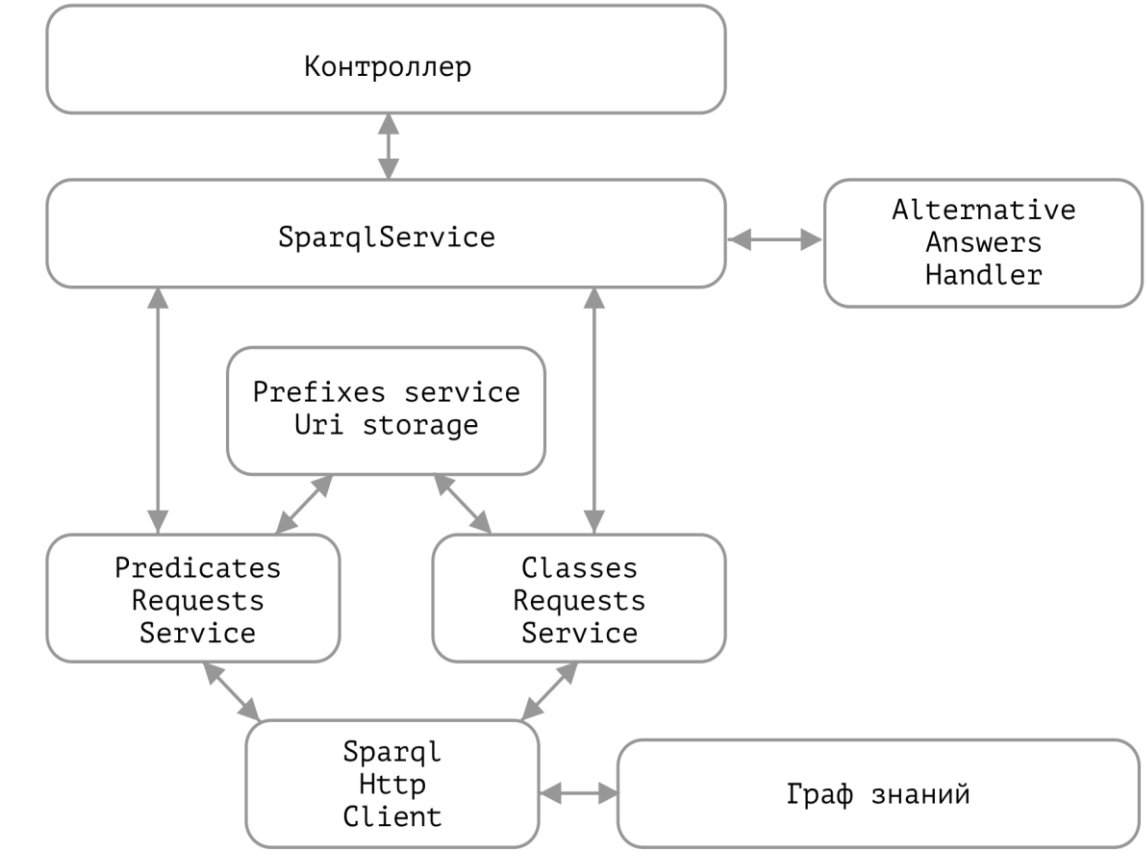


Рис. 2. — Архитектура

PrefixesService заменяет uri на сокращение или, наоборот, возвращает список префиксов с нужном формате для построения SPARQL-запроса.

Ниже приведены основные url для api системы, которые касаются исключительно генерации вопросов по заданным субъектам при помощи семантической сети.

1) Извлечение случайной сущности: /api/random-entity. В качестве параметра принимает класс сущности, по которой необходимо извлекать сущности. Для DBPedia список классов перечислен по ссылке DBPedia Ontology [9]. Если класс — dbo:Place, то необходимо извлечь из онтологии место. В таком случае рассматривается параметр запроса region. В нем передается список координат той области, внутри которой производится поиск мест по онтологии. В region передаются 4 значения координат, разделенных запятыми.

2) Запрос /api/find-entity возвращает то же, но в результате несколько вариантов, в предыдущем запросе возвращается только один. Запрос так же

принимает параметры класс, регион, но еще и строку, по которой происходит поиск.

3) Запрос /quiz/{quizId}/add-question-with-entity возвращает список доступных вопросов, которые могут быть добавлены в тест. В качестве параметра необходимо передать URI сущности, по которой будут сгенерированы вопросы.

4) Запрос /api/alternative-answers возвращает альтернативные варианты ответов для вопроса. Принимаемые параметры — класс предиката корректного варианта ответа. Например, если в ответе фигурирует страна, альтернативными вариантами должны быть другие страны. Еще запрос принимает параметр — корректный вариант ответа. Если корректный вариант — число или дата, никаких запросов к DBPedia не будет, и альтернативные варианты ответа (неверные) будут построены программно.

5) Редактирование вопроса, POST запрос: /quiz/{quizId}/question/{questionId}.

6) Удаление вопроса, DELETE запрос: /quiz/{quizId}/question/{questionId}.

7) Добавление вопроса в систему (финальный этап), POST запрос: /quiz/{quizId}/add-question/.

Для запросов к DBPedia используется библиотека Apache Jena, и приложение написано полностью на Java, опираясь на алгоритм linkeddata trivia.

Для извлечения литерала сущности (подписи на определенном языке) требуется задать язык в фильтре запроса. Многие ресурсы онтологии DBPedia, даже из тех, которые ссылаются на объекты реального мира и находятся в России, не имеют подписи на русском языке. Было решено строить каждый запрос с приоритетом языка. Если имеется подпись на русском языке — выводить подпись на русском, если нет — на английском. Это достигнуто тем, что фильтр подписи на русском ставится как OPTIONAL (не обязательный). Подпись на английском является обязательной, так как она имеется у каждого объекта. Далее переменные языков объединяются в одну методом COALESCE. Форма функции COALESCE возвращает значение RDF-члена первого выражения, которое оценивается без ошибок. В системе должны быть указаны два языка: язык с приоритетом 1 и язык с приоритетом 2. Язык с приоритетом 1 по умолчанию задан как русский (ru), а с приоритетом 2 — английский (en).

Пример использования функции COALESCE для извлечения литерала с приоритетом двух языков:

```
select ?entity, coalesce(?labelLang1, ?labelLang2) as ?label where {
 OPTIONAL {
  ?entity rdfs:label ?labelLang1 .
  FILTER(langMatches(lang(?labelLang1), "ru")) .
 }
 ?entity rdfs:label ?labelLang2 .
  FILTER(langMatches(lang(?labelLang2), "en")) .
}
```

5. Алгоритм

Чтобы пользователь имел возможность строить вопросы и по заданному субъекту, и по случайному субъекту, алгоритм должен предусматривать различные способы генерации. Основные способы построения вопросов: по поиску субъекта; случайному субъекту, определенному по классу; случайному субъекту, определенному по тематике вопросов. Алгоритм построения вопросов имеет три ответвления, основанные на способах генерации.

В ходе работы был создан и полностью реализовал алгоритм, способный быстрее делать то, что делает библиотека linkeddata trivia с более гибким применением (добавлена возможность поиска ключевой сущности для построения вопроса). Соответствующий алгоритм представлен на рисунке 3. Опорными пунктами являются подбор субъекта и построение вопросов по подходящим триплетам субъекта. При подборе субъекта по поиску в семантической сети DBPedia было решено использовать систему Google Custom Search для ускорения поиска.

6. Извлечение сущности

Извлечение сущности, по которой будет построен вопрос, — первичная задача системы при построении вопроса. Ниже приведен SPARQL-запрос, по которому извлекается сущность при поиске в базе DBPedia по запросу «Пушкин».

```
select distinct ?entity ?label where {
 ?entity a dbo:Person .
```

```
  ?entity rdfs:label ?label .
  FILTER contains(?label, "Пушкин") .
  FILTER(langMatches(lang(?label), "ru") || langMatches(lang(?label), "en"))
} limit 100
```

Префиксы — сокращения uri. В данном запросе использовались распространенные сокращения dbo, rdfs. На поиск ушло 54 секунды. Это очень долго, и такой запрос представляет нагрузку для системы DBPedia.

Альтернативный способ поиска по DBPedia. Ресурсы в DBPedia доступны как веб страницы по ссылке http://dbpedia.org/page/name, где name — имя ресурса. А URI ресурсов представлен так: http://dbpedia.org/resource/name. Поиск по веб-страницам любого сайта доступен при помощи поисковых систем. Было решено использовать поисковую систему Google для поиска необходимой страницы, после чего обращаться к необходимому ресурсу в базе знаний.

Система Custom Search JSON API позволяет настраивать программный поиск и отображения результатов из поиска Google. С помощью этого API была решена задача поиска по онтологии в виде поиска веб страниц отображаемых ресурсов онтологии. Результаты поиска обрабатывались в формате JSON.

Custom search engine ID — используется, чтобы указать пользовательскую поисковую систему, которая необходима для выполнения поиска. Пользовательская поисковая система (Custom Search Engine) должна быть создана с помощью панели управления (https://cse.google.com). В этой панели управления доступны настройки сайтов, на которых будет выполняться поиск, и сайты, исключенные из поиска. Пришлось исключить из поиска сайты http://eo.dbpedia.org, http://cs.dbpedia.org и другие для того, чтобы исключить из поиска некоторые нерабочие ссылки для системы генерации вопросов (те, которые не ссылаются на ресурсы онтологии DBPedia).

Ограничение использования бесплатной версии системы Google CSE -— 10000 запросов в день.

Пользовательская поисковая система реализована на стороне клиента. По запросу за 0.34 секунды были найдены подходящие результаты. Первый результат по запросу «Пушкин»: http://dbpedia.org/page/Alexander_Pushkin. Полученный url требуется преобразовать, заменив /page на /resource:

```
entity = entity.replace('/page', '/resource').
```

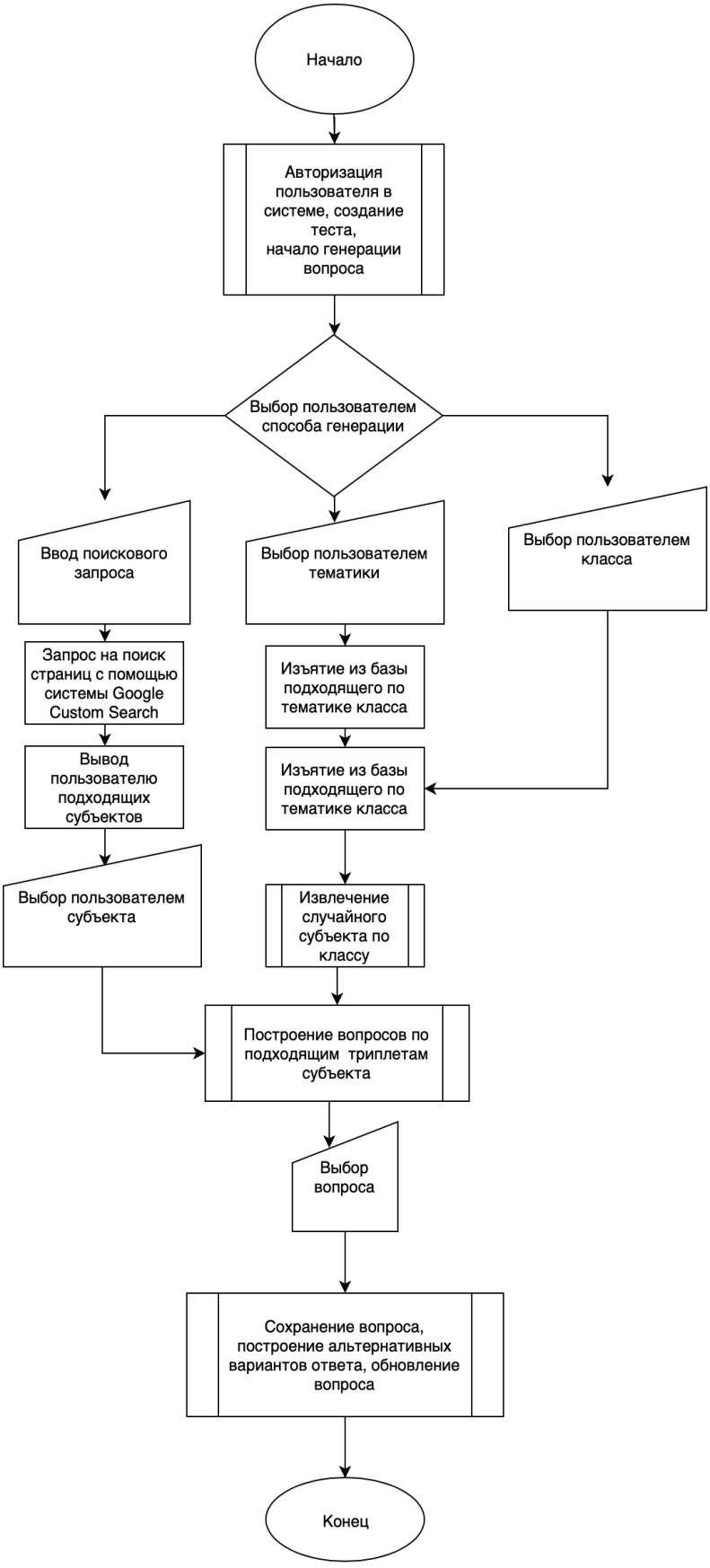


Рис. 3. — Алгоритм построения вопросов

Система получает Uri ресурсов семантической сети DBPedia за необходимое время по поисковому запросу.

Далее будет рассмотрена ситуация, в которой требуется извлечь сущность для региона, определенного по карте.

В первую очередь необходимо получить координаты, выбрав их на карте. В системе был выбран сервис Яндекс-карт. В параметрах Ajax-запроса используется следующий код: *myMap.getBounds().toString()*, где myMap — Яндекс-карта на веб-странице.

Если был выбран следующий регион: город Казань, SPARQL запрос, выполняемый на стороне сервиса, выглядит следующим образом:

```
PREFIX geopos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX georss: <http://www.georss.org/georss/>
select ?place coalesce(?labelLang1, ?labelLang2) as ?label where {
 ?place a dbo:Place .
 ?place geopos:lat ?lat .
 ?place geopos:long ?long .
 OPTIONAL {
  ?place rdfs:label ?labelLang1 .
  FILTER(langMatches(lang(?labelLang1), "ru")) .
 }
 ?place rdfs:label ?labelLang2 .
 FILTER (
  ?lat > 55.75866589150474 &&
  ?lat < 55.83602591889409 &&
  ?long > 48.98691801306621 &&
  ?long < 49.24441007849591 &&
  lang(?labelLang2) = "en"
 )
} LIMIT 1000
```

Результаты выполнения запроса следующие: Мост Миллениум (Казань), Казанский кремль, Кул-Шариф, Казань, Центральный (стадион, Казань), Баскет-холл (Казань), Горки (станция метро), Мечеть аль-Марджани, Храм Воздвижения Святого Креста (Казань), Суконная слобода (станция метро), Площадь Габдуллы

Тукая (станция метро) и другие. После того, как пользователь выберет подходящий субъект, программа построит по нему вопросы.

7. Построение вопросов

Имеется сущность, по которой системе нужно построить вопрос. Для построения вопроса требуется извлечь подходящие триплеты для этой сущности. Многие предикаты повторяются, и построение вопроса по этим предикатам не имеет смысла. Например, id ресурса онтологии на сайте Википедия. Чтобы исключить те предикаты, по которым не требуется строить вопрос, был реализован черный список предикатов, который хранится в PrefixesService.

Часть тех URI, которые вошли в черный список:

- "http://www.w3.org/1999/02/22-rdf-syntax-ns#type"
- "http://www.w3.org/2002/07/owl#sameAs"
- "http://www.w3.org/ns/prov#wasDerivedFrom"
- "http://dbpedia.org/ontology/wikiPageRevisionID"
- "http://dbpedia.org/ontology/wikiPageID"
- "http://dbpedia.org/ontology/wikiPageWikiLink"
- "http://dbpedia.org/ontology/wikiPageDisambiguates"
- "http://dbpedia.org/ontology/soundRecording"
- "http://www.w3.org/2000/01/rdf-schema#seeAlso"
- "http://www.w3.org/2000/01/rdf-schema#label" и другие.

Далее будет рассмотрены SPARQL-запросы, которые строятся сервисом PredicatesRequestsService. Первый запрос извлекает триплеты (субъект, предикат, объект), в которых субъектом является заданная сущность. Например, если вопросы строятся про Репина Илью Ефимовича, то в первом случае один из вопросов будет выглядеть так: Кто был под влиянием Репина Ильи Ефимовича? Ответ: Суриков, Василий Иванович. То есть основой вопроса в этом случае является заданная сущность. Из графа знаний извлекаются только те узлы, которые направлены от заданного узла (который ссылается на заданную сущность).

Ниже приведен сам первый SPARQL-запрос для извлечения триплетов для построения вопросов по заданному субъекту:

```
select DISTINCT coalesce(?subjectLabelLang1, ?subjectLabelLang2) as ?subjectLabel, ?predicate, coalesce(?predicateLabelLang1, ?predicateLabelLang2) as ?predicateLabel, ?object, coalesce(coalesce(?objectLabelLang1, ?objectLabelLang2), ?object) as ?objectLabel where {
  <http://dbpedia.org/resource/Ilya_Repin> ?predicate ?object.
  OPTIONAL {
   <http://dbpedia.org/resource/Ilya_Repin> rdfs:label ?subjectLabelLang1.
   FILTER(langMatches(lang(?subjectLabelLang1), "ru")).
  }
  <http://dbpedia.org/resource/Ilya_Repin> rdfs:label ?subjectLabelLang2.
  FILTER(langMatches(lang(?subjectLabelLang2), "en")).
  OPTIONAL {
   ?object rdfs:label ?objectLabelLang1.
   FILTER(langMatches(lang(?objectLabelLang1), "ru")).
  }
  OPTIONAL {
   ?object rdfs:label ?objectLabelLang2.
   FILTER(langMatches(lang(?objectLabelLang2), "en")).
  }
  OPTIONAL {
   ?predicate rdfs:label ?predicateLabelLang1.
   FILTER(langMatches(lang(?predicateLabelLang1), "ru")).
  }
  ?predicate rdfs:label ?predicateLabelLang2 .
  FILTER(langMatches(lang(?predicateLabelLang2), "en")).}
 limit 3000
```

В списке возвращаемых значений субъект (URI ресурса), его литерал, предикат (URI), его литерал, объект (URI), литерал объекта. Предикат нужен для построения в дальнейшем альтернативных вариантов ответа. Литерал объекта может отсутствовать, тогда URI объекта подставляется на место литерала. Это добавляет некоторые вопросы относительно заданного субъекта, когда

некоторые числовые значения для субъекта не имеют литерала ни на русском, ни на английском.

Второй запрос извлекает те триплеты, в которых объектом триплета является заданная сущность. В этом случает один из вопросов будет выглядеть так: Кто находился под влиянием Врубеля Михаила Александровича? Ответ: Репин, Илья Ефимович. Или такой вопрос: Кто является автором «Запорожцы» (картина)? Ответ: Репин, Илья Ефимович. То есть основой вопроса в этом случае являются те ресурсы, которые ссылаются на заданный ресурс. Из графа знаний извлекаются только те узлы, которые направлены к заданному узлу.

```
select DISTINCT ?subject, coalesce(?subjectLabelLang1, ?subjectLabelLang2) as ?subjectLabel, ?predicate, coalesce(?predicateLabelLang1, ?predicateLabelLang2) as ?predicateLabel, coalesce(?objectLabelLang1, ?objectLabelLang2) as ?objectLabel where {
  ?subject ?predicate <http://dbpedia.org/resource/Ilya_Repin> .
  OPTIONAL {
   <http://dbpedia.org/resource/Ilya_Repin> rdfs:label ?objectLabelLang1 .
   FILTER(langMatches(lang(?objectLabelLang1), "ru")) .
  }
  <http://dbpedia.org/resource/Ilya_Repin> rdfs:label ?objectLabelLang2 .
  FILTER(langMatches(lang(?objectLabelLang2), "en")) .
  OPTIONAL {
   ?subject rdfs:label ?subjectLabelLang1 .
   FILTER(langMatches(lang(?subjectLabelLang1), "ru")) .
  }
  ?subject rdfs:label ?subjectLabelLang2 .
  FILTER(langMatches(lang(?subjectLabelLang2), "en")) .
  OPTIONAL {
   ?predicate rdfs:label ?predicateLabelLang1 .
   FILTER(langMatches(lang(?predicateLabelLang1), "ru")) .
  }
  ?predicate rdfs:label ?predicateLabelLang2 .
  FILTER(langMatches(lang(?predicateLabelLang2), "en")) .}
 limit 3000
```

Distint означает, что возвращаемые после выполнения запроса значения не должны повторяться.

Резюмируя представленные запросы, выражаясь в терминологии триплетов, можно сказать, что первый запрос возвращает все предикаты и объекты по заданному субъекту, а второй запрос возвращает субъекты и предикаты по заданному объекту.

Результаты каждого запроса фильтруются через черный список предикатов, все результаты собирает сервис SparqlService и возвращает в виде массива TripleDto.

Построение шаблона для построения вопроса на английском достаточно просто. Из списка получившихся выше описанных триплетов построить список предлагаемых пользователю вопросов удается по шаблону:

What is the {triple.**predicate**Label} of {triple.**subject**Label}? ({triple.**object**Label}).

Было решено выбрать следующий шаблон для русского языка:

Кто, или что, или какой {triple.**predicate**Label} {triple.**subject**Label}? ({triple.**object**Label}).

Учитывая, что в базе более 400 млн триплетов, система способна сгенерировать разные вопросы, и около 300 млн из них могут не повторяться.

Получившийся список некоторых вопросов для заданного субъекта Казань (http://dbpedia.org/resource/Kazan) представлен в таблице 4. Всего для этого субъекта получилось 605 вопросов. Выбрав подходящий вопрос из списка, пользователь может добавить его в создаваемый тест.

8. Построение альтернативных вариантов ответа

Если необходимо предоставить альтернативный вариант ответа или если необходимо построить вопрос по случайному субъекту (как в linkeddata trivia), система выбирает случайную сущность по заданному классу. Для извлечения случайной сущности сперва необходимо получить случайное смещение (random offset). Смещение может быть указано от нуля до количества сущностей заданного класса. Чтобы извлечь количество сущностей для заданного класса, система выполняет следующий SPARQL-запрос:

Таблица 4 — Некоторые вопросы по ресурсу dbr:Kazan

Вопрос	Ответ
Кто, или что, или какой country Казань?	Россия
Кто, или что, или какой federal subject Казань?	Republic of Tatarstan
Кто, или что, или какой leader name Казань?	Метшин, Ильсур Раисович
Кто, или что, или какой May record low C Казань?	-6.5
Кто, или что, или какой May record high C Казань?	33.5
Кто, или что, или какой Jun record low C Казань?	-1.4
Кто, или что, или какой death place Кул Шариф (религиозный деятель)?	Казань
Кто, или что, или какой death place Девятаев, Михаил Петрович?	Казань
Кто, или что, или какой death place Вахитов, Мулланур Муллазянович?	Казань
Кто, или что, или какой birth place Никифоров, Николай Анатольевич?	Казань
Кто, или что, или какой birth place Шаляпин, Фёдор Иванович?	Казань
Кто, или что, или какой birth place Попов, Валерий Георгиевич?	Казань
Кто, или что, или какой stadium Чемпионат мира по футболу 2018?	Казань
Кто, или что, или какой city Чемпионат Европы по тяжёлой атлетике 2011?	Казань

```
select count(?obj) as ?count
where {
 ?obj a ?type .
 ?obj rdfs:label ?label
 FILTER(langMatches(lang(?label), LANGUAGE2)) .
}
```

Получившееся количество задается как верхняя граница случайного смещения. В следующем запросе извлекается случайная сущность:

```
select coalesce(?labelLang1, ?labelLang2) as ?label where {
 ?subject a ?type.
 OPTIONAL {
  ?subject rdfs:label ?labelLang1.
  FILTER(langMatches(lang(?labelLang1), LANGUAGE1)).
 }
 ?subject rdfs:label ?labelLang2.
 FILTER(langMatches(lang(?labelLang2), LANGUAGE2))
} OFFSET ?offset
LIMIT 1
```

При построении альтернативных вариантов ответа запрос к графу знаний не обязателен. Если значение верного ответа числовое или значение является датой или годом, альтернативные варианты ответа получаются программно.

Если необходимо извлечь альтернативный вариант ответа из графа знаний, система определяет rdfs:range для ресурса корректного варианта ответа. Rdfs:range объявляет класс или тип данных объекта. Выполняется следующий запрос:

```
select ?range where {
 < predicateUri> rdfs:range ?range.
}
```

После того, как система получает данные о классе сущности верного ответа, она получает случайную сущность по этому классу так, как это описано вышев данном разделе.

Система сохраняет класс верного ответа в базе данных вместе с составленным вопросом для того, чтобы при необходимости составитель теста

или викторины мог снова сгенерировать и поменять альтернативные варианты ответов.

9. Апробация. Генерация вопросов

Системой были построены вопросы для тестирования по теме «Живопись». Выбранные пользователем субъекты: Леонардо да Винчи, Репин, Илья Ефимович, Врубель, Михаил Александрович, Куинджи, Архип Иванович, Шишкин, Иван Иванович. Некоторые из построенных вопросов были выбраны для теста. Все вопросы были проверены вручную. На поиск субъектов уходило не более секунды. Компьютер, на котором происходило построение, имеет следующую конфигурацию: 4 гб озу, процессор 2,7 GHz Intel Core i5. Скорость соединения с интернетом: 30 мбит/с. На построение вопросов по заданному субъекту затрачено около пяти секунд. Построенные вопросы представлены в таблице 5.

Получившиеся вопросы переведены не полностью. Приоритет извлечения языковых литералов был установлен сначала на русский язык, и при его отсутствии извлекался литерал на английском языке. Те предикаты и субъекты семантической сети DBPedia, которые не имеют литерала на русском языке, остались на английском.

Таблица 5 — Построенные вопросы

Вопрос	Ответ
Кто, или что, или какой автор Мона Лиза?	Леонардо да Винчи
Кто, или что, или какой movement Репин, Илья Ефимович?	Реализм
Кто, или что, или какой movement Врубель, Михаил Александрович?	Символизм
Кто, или что, или какой training Врубель, Михаил Александрович?	Императорская Академия художеств
Кто, или что, или какой influenced by Куинджи, Архип Иванович?	Шишкин, Иван Иванович
Кто, или что, или какой автор Запорожцы (картина)?	Репин, Илья Ефимович
Кто, или что, или какой training Шишкин, Иван Иванович?	Московское училище живописи, ваяния и зодчества
Кто, или что, или какой death place Девятаев, Михаил Петрович?	Казань
Кто, или что, или какой death place Врубель, Михаил Александрович?	Санкт-Петербург
Кто, или что, или какой автор Демон сидящий?	Врубель, Михаил Александрович
Кто, или что, или какой works Шишкин, Иван Иванович?	Morning in a Pine Forest
Кто, или что, или какой автор Тайная вечеря?	Леонардо да Винчи
Кто, или что, или какой автор Портрет Джиневры де Бенчи?	Леонардо да Винчи

ЗАКЛЮЧЕНИЕ

Разработана система генерации вопросов для тестов и викторин по заданному субъекту, найденному в поиске или при помощи географической карты. С помощью разработанной системы были построены тесты, которые были предложены для прохождения десяткам студентов ВШ ИТИС КФУ. Ими были пройдены все тесты и отмечено, что все вопросы были понятны.

В созданной системе есть функционал, реализующий поиск субъектов для построения вопросов при помощи географической карты. Это дает возможность построения викторин и тестов по краеведению.

Клиентская часть системы может быть разработана для мобильных устройств так, что у пользователей будет возможность создавать обучающие тесты для себя и друг для друга.

Пользователи (школьники, студенты, преподаватели, другие), пользуясь функционалом системы, могут отправлять друг другу созданные тесты и таким образом повышать собственную эрудицию.

Созданную систему можно доработать, добавив в нее доступные графы знаний Wikidata. Это увеличит количество вопросов, которые способна предоставить система.

СПИСОК ЛИТЕРАТУРЫ

1. RDF Schema 1.1 // W3C. URL: https://www.w3.org/TR/rdf-schema/ (дата обращения: 25.03.2019).

2. *Tim Berners-Lee.* Linked Data. URL: https://www.w3.org/DesignIssues/LinkedData.html (дата обращения: 11.04.2019).

3. *Amrapali Zaveri*, University of Leipzig. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. URL: http://www.aifb.kit.edu/images/c/ca/KG-Comparison-SWJ-Article.pdf (дата обращения: 05.04.2019).

4. Learn about DBPedia // About | DBPedia. URL: https://wiki.dbpedia.org/about (дата обращения: 10.04.2019).

5. *G. Vega-Gorgojo, Don Naipe*. Clover quiz: A mobile trivia game based on DBpedia data. URL: http://ceur-ws.org/Vol-1963/paper474.pdf (дата обращения: 11.04.2019).

6. Auto-generated trivia questions based on DBPedia data // n1try/ linkeddata-trivia. URL: https://github.com/n1try/linkeddata-trivia (дата обращения: 01.02.2019).

7. Linked Open Data Seminar 2016 – Knowledge Panel. // n1try/kit-lod16-knowledge-panel. URL: https://github.com/n1try/kit-lod16-knowledge-panel (дата обращения: 01.02.2019).

8. Virtuoso SPARQL Query Editor // DBpedia. URL: https://dbpedia. rg/ sparql (дата обращения: 20.03.2019).

9. Ontology Classes // DBPedia mappings. URL: http://mappings.dbpedia. org/server/ontology/classes/ (дата обращения: 27.03.2019).

DEVELOPMENT OF A SOFTWARE PACKAGE FOR GENERATING QUESTIONS FOR SPECIFIED SUBJECTS USING A SEMANTIC NETWORK

M. D. Andreichev [1], A. A. Ferenetz [2]

[1-2] Higher School for Information Technologies and Intelligent Systems of Kazan (Volga Region) Federal University

[1] andreichev.m@mail.ru, [2] ist.kazan@gmail.com

Abstract

An approach to automatically generating questions for tests or quizzes using the DBPedia knowledge graph is presented here. The selected knowledge graph has about 5 million entities. DBpedia SPARQL endpoint the ability to make queries to the semantic network using the SPARQL language. The algorithm, the basic queries to the knowledge graph for constructing questions, a non-standard approach to the search for entities are presented in this article.

Keywords*: semantic network, generation of questions, linked data, ontology, knowledge graph, RDF, SPARQL, DBPedia*

REFERENCES

1. RDF Schema 1.1 // W3C. URL: https://www.w3.org/TR/rdf-schema

2. *Tim Berners-Lee*. Linked Data. URL: https://www.w3.org/DesignIssues/LinkedData.html

3. *Amrapali Zaveri*, University of Leipzig. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. URL: http://www.aifb.kit.edu/ images/c/ca/KG-Comparison-SWJ-Article.pdf

4. Learn about DBPedia. URL: https://wiki.dbpedia.org/about

5. Auto-generated trivia questions based on DBPedia data. URL: https://github.com/n1try/linkeddata-trivia

6. Linked Open Data Seminar 2016 – Knowledge Panel. URL: https://github.com/n1try/kit-lod16-knowledge-panel

7. *G. Vega-Gorgojo, Don Naipe*. Clover quiz: A mobile trivia game based on DBpedia data. URL: http://ceur-ws.org/Vol-1963/paper474.pdf

8. Virtuoso SPARQL Query Editor // DBpedia. URL: https://dbpedia.org/sparql

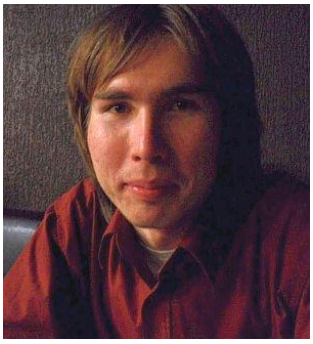
9. Ontology Classes // DBPedia mappings. URL: http://mappings.dbpedia.org/server/ontology/classes/

СВЕДЕНИЯ ОБ АВТОРАХ

АНДРЕИЧЕВ Михаил Дмитриевич – бакалавр Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Michail Dmitrievich ANDREICHEV – bachelor of Higher School of ITIS KFU.

email: andreichev.m@mail.ru



ФЕРЕНЕЦ Александр Андреевич – ассистент, преподаватель кафедры программной инженерии Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Alexander Andreevich FERENETZ – assistant at Department of Software Engineering of Higher School of ITIS KFU

email: ist.kazan@gmail.com

Материал поступил в редакцию 28 июня 2019 года

УДК 004.85

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ВЗАИМОСВЯЗИ АКАДЕМИЧЕСКОЙ УСПЕВАЕМОСТИ И ДАННЫХ ПРОФИЛЯ СОЦИАЛЬНОЙ СЕТИ

И. Р. Ихсанов[1], И. С. Шахова[2]

Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

[1]ilias.ihsanov@gmail.com, [2]is@it.kfu.ru

Аннотация

Предложена модель машинного обучения для выявления взаимосвязи между данными профиля социальной сети и академической успеваемости учащегося, а также прогнозирования среднего балла успеваемости по данным параметрам.

Ключевые слова*: машинное обучение, социальные сети, психометрия, академическая успеваемость, образование, абитуриент*

ВВЕДЕНИЕ

Исследование, проведенное в Тинто в 1987 году, показало, что примерно 57% студентов выбирают учебное заведение, не обращая внимания на факультет обучения, а 43% студентов вуза бросают учебу, так и не получив диплом о высшем образовании. Особое внимание в исследовании уделялось факторам, влияющим на способность студента успешно закончить высшее учебное заведение. Был изучен ряд академических факторов для выявления студентов, которые с наибольшей вероятностью достигнут успеха. Исследователями была выявлена зависимость, что студенты, обладающие высокой уверенностью в себе, самообладанием, устремленностью в достижении целей связаны с более высокой успеваемостью. Кроме того, студенты, которые являются адаптивными перфекционистами, с большей вероятностью успешно завершают обучение. Таким образом, было выявлено, что личностные параметры пригодны для определения будущей успеваемости и вероятности отчисления студента из вуза [1].

Однако сбор и анализ данных о личностных характеристиках представляют собой трудозатратный процесс, так как включают в себя целый набор задач: от составления вопросов анкетирования до анализа проведенного тестирования для выявления персональных характеристик респондента.

Таким образом, данная работа нацелена на разработку модели машинного обучения для выявления взаимосвязи индивидуальных характеристик учащихся и их академической успеваемости, а также прогноза среднего балла успеваемости по данным характеристикам.

ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Психометрия – это изучение психологических изменений личности: способностей, взглядов и качеств. В рамках психометрии спроектирована модель личности человека, состоящая из пяти черт: экстраверсии (черта характеризуется склонностью к широким социальным контактам), доброжелательности, добросовестности, невротизма (эмоциональной нестабильности) и открытости к новому опыту [2]. Личностная модель позволяет с научной точки зрения прогнозировать действия человека, делать выводы о его профессиональной пригодности, перспективах профессионального роста, возможности работы в коллективе и многое другое.

В частности, в Израиле существует единый психометрический экзамен для поступающих в вузы. Он официально рассматривается в Израильском центре экзаменов и оценок как средство прогнозирования шансов на успех в занятиях в высших учебных заведениях [3].

Подходы к автоматизации данного тестирования приведены в научной работе доктора Б. Шуотан из Китайской Национальной академии наук, где путем анализа активности в социальной сети вычисляется каждая из пяти диспозиций личности [4]. Основные параметры, собранные из социальных сетей во время исследования: возраст, родной город, частота использования социальной сети, частота загрузки материалов и многие другие. Для решения задачи классификации в данном исследовании было проведено тестирование набора данных на многих алгоритмах классификации, таких, как наивный байесовский классификатор (NB), метод опорных векторов (SVM), дерево решений и так далее. Ученые выяснили, что дерево решений C4.5 может дать лучшие

результаты [5].

Помимо указанных выше параметров, ученые собирали информацию о демографических переменных (пол, этническая принадлежность и уровень образования родителей), частоте использования Facebook (FBTime или FBCheck) и частоте действий в Facebook.

Кроме того, в рамках данного исследования удалось выявить зависимость между затраченным временем на учебную работу и временем, проведенным в социальной сети. Эти результаты согласуются с другими исследованиями, которые обнаружили, что использование интернета и, в частности, Facebook, определенным образом приводит к улучшению психосоциальных результатов, а использование Twitter определенным образом приводит к лучшим академическим результатам [6].

Ученые из Калифорнии показали, что легкодоступные цифровые записи поведения, такие, как Facebook Likes, могут использоваться для автоматического и точного прогнозирования ряда очень чувствительных личных качеств, включая этническую принадлежность, религиозные и политические взгляды, личностные качества, интеллект, счастье, факт развода родителей, возраст и пол [7]. Данный анализ основан на наборе данных о более чем 58 000 добровольцах, которые предоставили свои «лайки» в Facebook, подробные демографические профили и результаты нескольких психометрических тестов. Предложенная модель использует уменьшение размерности для предварительной обработки данных Likes, которые затем вводятся в линейную регрессию для прогнозирования отдельных психо-демографических профилей из оценок «Мне нравится» [8].

Также в рамках текущего исследования была проанализирована работа Национального исследовательского Томского государственного университета «Методы и инструменты выявления перспективных абитуриентов в социальных сетях» [9]. Результаты исследования показывают, что методика предсказания образовательных интересов и признаков одаренности по подпискам пользователей дала лучший результат. По этим параметрам есть возможность конструирования прогностической модели выявления перспективных абитуриентов через проекцию целевой модели выпускника Томского государственного университета.

ВХОДНЫЕ ДАННЫЕ

По данным немецкого аналитического агентства Statista [10], в России проникновение социальных сетей оценивается в 47%, аккаунты в них имеют 67,8 млн россиян. На рисунке 1 видно, что активнее всего в РФ используют YouTube (63% опрошенных), второе место занимает ВКонтакте — 61%. Глобальный лидер Facebook лишь на четвертой строчке с показателем в 35%. Среди мессенджеров доминируют Skype и WhatsApp по 38%.

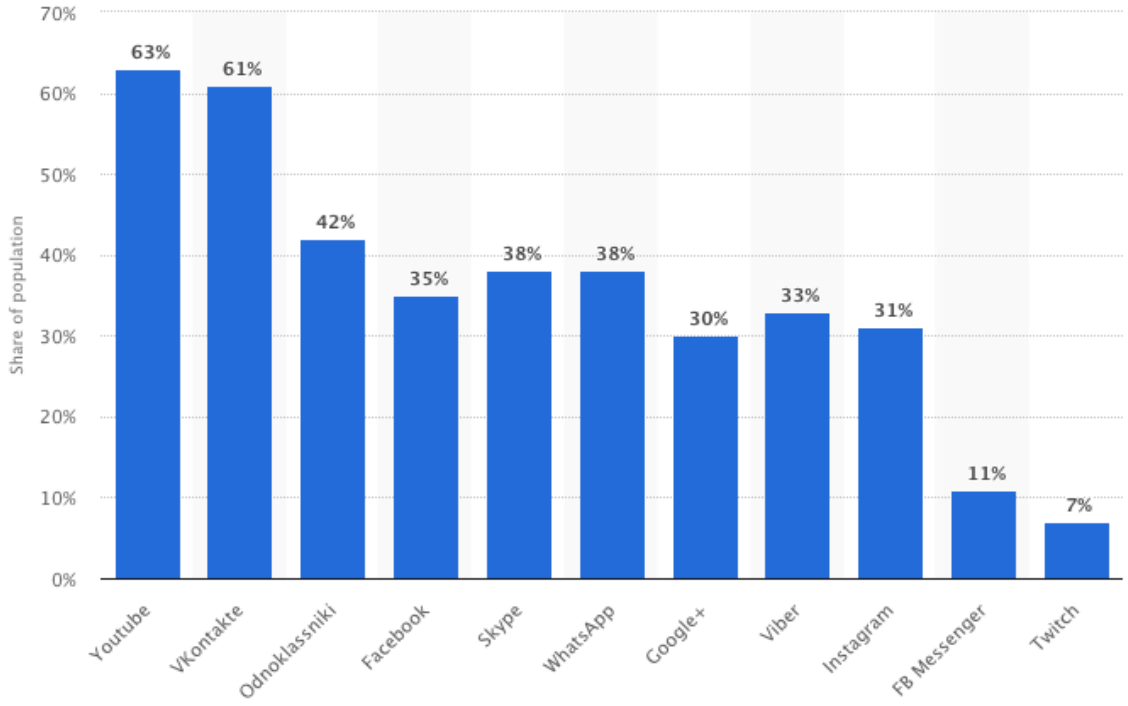


Рис. 1 – Активность использования социальных сетей в России

В исследовании [11] было выявлено, что 97,4% опрошенных студентов зарегистрировано в социальных сетях и являются активными пользователями интернета. Лидирующей социальной сетью была определена ВКонтакте [12], на нее указали 95,5% респондентов. С 94,7%, использующих социальную сеть в возрасте 15–20 лет, и 91,3% в возрасте 18–21 год. Таким образом, в качестве источника данных для текущего исследования была выбрана социальная сеть ВКонтакте.

Политика предоставления данных социальной сетью ВКонтакте дает возможность получить только те данные, которые пользователь разрешил показывать остальным пользователям ВКонтакте. Также политика

предоставления данных ВКонтакте ограничивает предоставление информации о пользователе без регистрации в системе [13]. Информация, предоставляемая ВКонтакте, которую можно использовать для первичного анализа в рамках текущего исследования, включает в себя следующие данные: пол, дата рождения, количество видео, количество альбомов, количество аудио, количество заметок, количество фотографий на странице, количество групп, количество друзей, общее количество доступных фотографий, количество подписчиков, количество интересных страниц, семейное положение, количество подписок, количество фотографий профиля, интересные страницы с тематикой страниц в порядке популярности, подарки, количество подписок на популярных личностей.

СБОР И ОБРАБОТКА ДАННЫХ

Чтобы получить сочетание ФИО, рейтинга и идентификатора профиля в ВКонтакте, вручную были проведены поиск пользователей в социальной сети и сопоставление их идентификатора с местом в рейтинге. В результате входные данные представляют собой модель, состоящую из набора: name – имя пользователя, id – уникальный идентификатор пользователя в социальной сети ВКонтакте, range – средний балл студента за год по сданным предметам в первый и второй семестры.

Таким образом, из 394 человек 354 человека были со страницами в открытом доступе и 40 человек – с приватными профилями и минимально возможными данными (количество друзей, пол, дата рождения).

Следующим этапом обработки данных стала их очистка. В полученных данных были преобразованы поля приватных аккаунтов, изменены признаки, имеющие тип объекта, в числовой тип, а также пустые значения были заменены на соответствующий тип.

На этом этапе была проведена очистка от выбросов (результат измерения, отличающийся от общей выборки). Удаление значений признаков было произведено, руководствуясь правилом экстремальных аномалий [16] по формулам:

$$IQ = (Q3 – Q1), \quad (1)$$

$$Q_н = Q1 – 3 * IQ, \quad (2)$$

$$Q_в = Q3 + 3 * IQ, \quad (3)$$

где $Q_н$ – внешний нижний забор; $Q_в$ – внешний верхний забор; IQ – интерквартильный размах; Q1 – первый квартиль, определяемый как 25-й процентиль данных; Q3 – третий квартиль, определяемый как 75-й процентиль данных.

Значения ниже нижнего внешнего забора и выше внешнего верхнего в признаках «подписки» и «количество фотографий профиля» были удалены.

После обработки аномальных значений была удалена одна колонка – количество групп, в которых состоит пользователь, так как в нем отсутствовало больше половины значений.

Следующим этапом стал разведочный анализ данных. В процессе анализа были выявлены аномальные значения – подписки студента на популярные личности. Также в процессе анализа были выявлены сильные зависимости баллов студента и его подписок на сообщества. Например, была выявлена сильная по сравнению с другими параметрами корреляция с тематикой групп, на которые подписаны студенты, и их средний балл. То есть, исходя из корреляций на рисунке 2, можно сказать, что порядковый номер страницы в списке интересных страниц имеет связь с итоговым баллом студента. Для уточнения были построены графики плотности, представляющие собой сглаженные гистограммы для демонстрации взаимосвязи тематики групп и баллов студента [17]. Фактическое влияние на итоговый средний балл студента изображено на рисунках 3–7.

Если посмотреть на графики плотности баллов студента и группы с конкретной тематикой в приоритете страниц пользователя с 1 по 5, то заметно, что студенты с интересной страницей на первом месте по приоритету с тематикой «программирование» и «соседи» учатся в диапазоне с 80 до 95

баллов. Студенты, подписанные на страницы с тематикой «образование», «видеоигры» и «креативная работа» учатся с 60 до 80 баллов. Таким образом, этот признак был в первую очередь включен в модель.

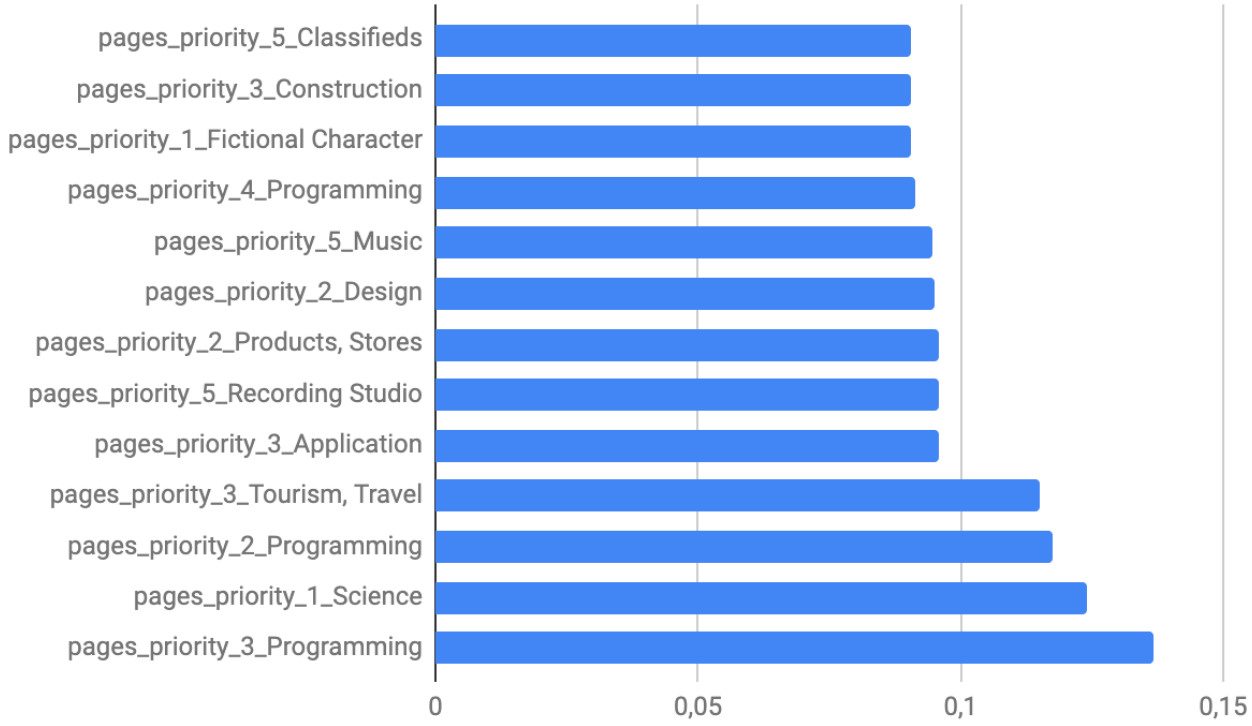


Рис. 2 – Корреляция баллов и тематик групп студента

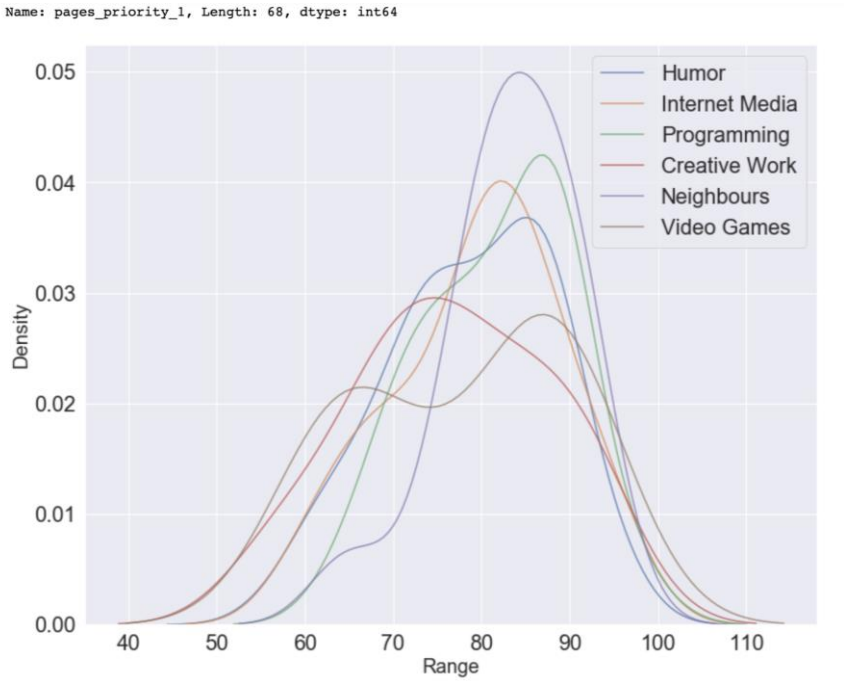


Рис. 3 – Плотность баллов студента и интересных страниц на первом месте по приоритету

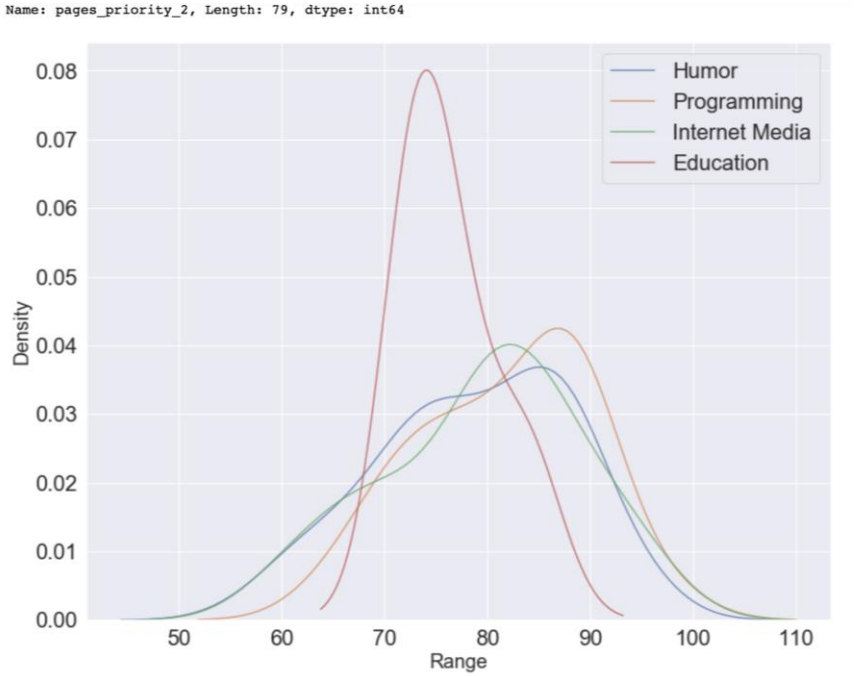


Рис. 4 – Плотность баллов студента и интересных страниц на втором месте по приоритету

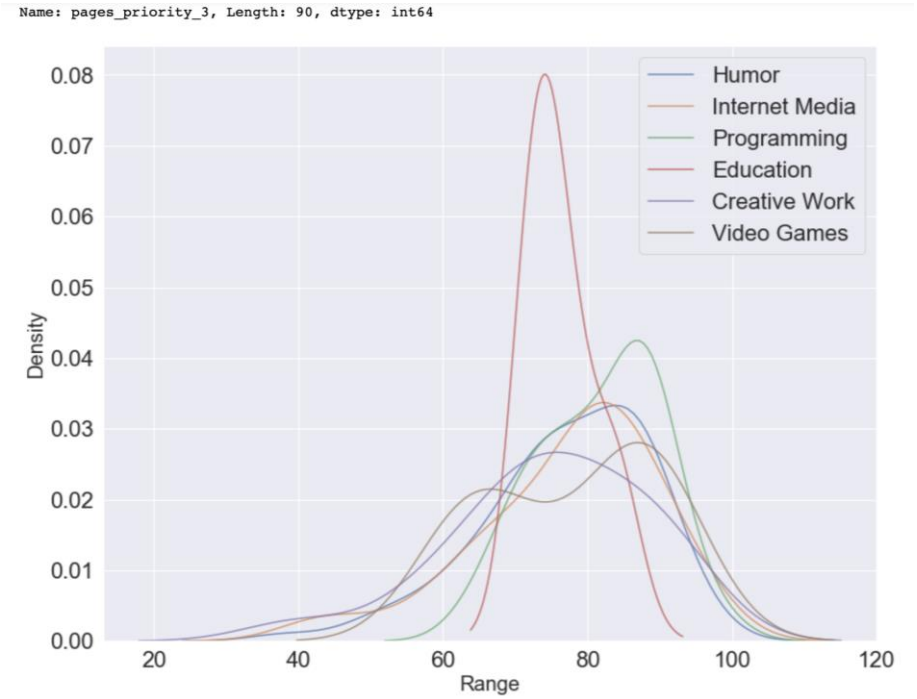


Рис. 5 – Плотность баллов студента и интересных страниц на третьем месте по приоритету

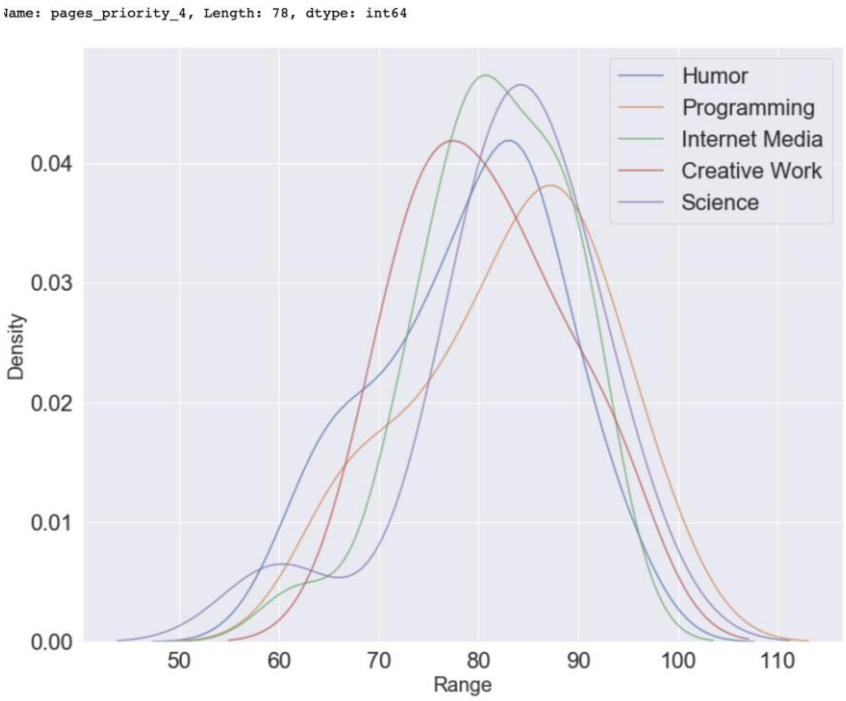


Рис. 6 – Плотность баллов студента и интересных страниц на четвертом месте по приоритету

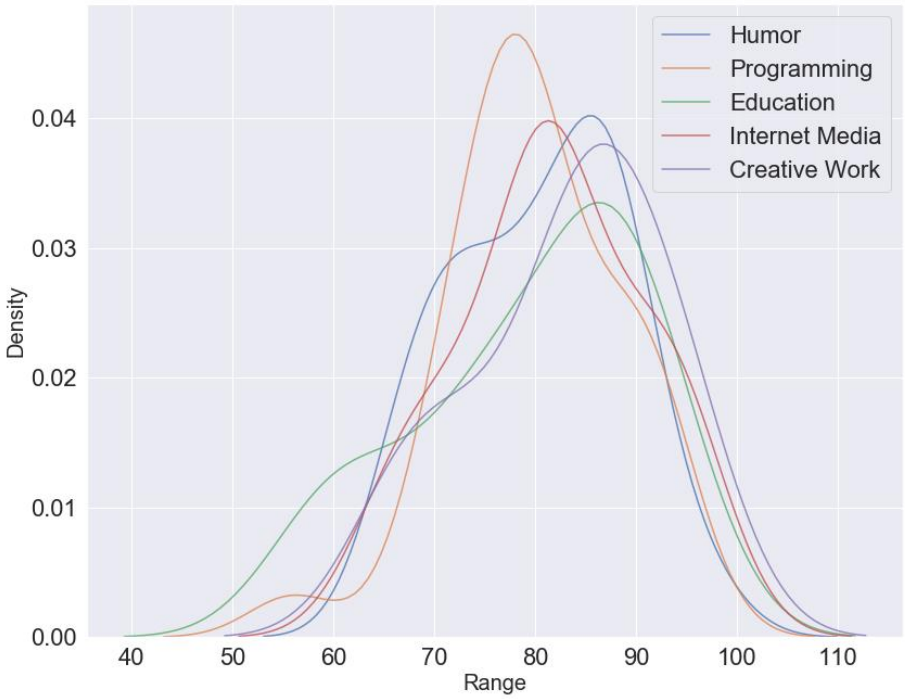


Рис. 7 – Плотность баллов студента и интересных страниц на пятом месте по приоритету

За счет конструирования (создания новых признаков из имеющихся данных) и выбора признаков (удаление лишних признаков и сохранение

коррелирующих) уменьшаются временные затраты на машинное обучение. Конструирование представляет собой получение и создание новых признаков из полученных данных. К операциям конструирования относятся извлечение натурального логарифма, применение кодирования к категориальным переменным (алгоритмы машинного обучения не работают с строковыми типами напрямую), извлечение квадратного корня. Выбор признаков заключается в процессе выбора из данных подходящих признаков. В данном процессе удаляется часть несущественных признаков, и остаются те, которые оказывают влияние на модель машинного обучения.

В процессе преобразований были выполнены кодирование категориальных переменных и извлечение натурального логарифма от числовых переменных.

В процессе обучения было замечено, что пользователи, имеющие менее 56 баллов (отчисленные), ухудшают процесс обучения модели, в результате чего средняя абсолютная ошибка для модели достигает 10,3. Было принято решение исключить этих пользователей из выборки. На начало преобразований было 15 признаков. В результате проведенных операций осталось 368 пользователей и 27 признаков: пол, количество видео, количество альбомов, количество аудио, количество заметок, количество фото, количество друзей, количество подписчиков, количество популярных личностей, количество интересных страниц, семейное положение, интересные страницы в приоритете от 1 до 5 по категориям, а также признаки, полученные при преобразовании: извлечение квадратного корня и логарифма.

Было проверено, нет ли избыточных признаков (признаки, коррелирующие друг с другом), так как удаление одного из коллинеарных признаков помогает модели быть более интерпретируемой.

Коллинеарность признаков была вычислена популярным методом фактора увеличения дисперсии. Для удаления и поиска коллинеарных признаков был использован коэффициент B-корреляции, если коэффициент корреляции между признаками был больше 0,6, то один из признаков был исключен.

Последним шагом в преобразовании данных стало масштабирование количественных признаков. Необходимость этого шага обусловлена тем, что

алгоритм градиентного бустинга чувствителен к масштабированию данных, а также значения признаков находятся в разных диапазонах. Количественные признаки: количество видео, фото, аудио, подписок, групп, подписчиков, друзей, интересных страниц, популярных личностей были нормализованы, то есть диапазон значений был приведен к формату от 0 до 1. Больше всего нормализация требуется для алгоритмов k-ближайших соседей и метода опорных векторов, так как они в своих алгоритмах учитывают евклидово расстояние между наблюдениями.

Существует два способа масштабирования объектов – это стандартизация и нормализация. Для решения данной задачи был выбран способ нормализации, потому что было замечено, что при обучении с разными способами преобразования нормализация показала лучший результат. Нормализация заключается в том, что выбирается минимальное значение признака и делится на разницу между максимальным и минимальным. Таким образом гарантируется диапазон значений от 0 до 1. На выходе были получены количественные признаки с нормальным форматом данных для модели с диапазоном от 0 до 1.

ВЫБОР И НАСТРОЙКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

На рисунке 8 представлена связь между точностью и интерпретируемостью нескольких алгоритмов [19]. Интерпретируемость означает, что мы можем понять причины, по которым алгоритм дал конкретный ответ. Также интерпретируемость гарантирует, что модель является правильной и неправильной по определенным причинам. Понимание модели помогает улучшить ее и укрепить уверенность в том, что выбранная модель будет работать с меньшей ошибкой или с большим процентом точности.

В качестве методов машинного обучения были выбраны:

- Линейная регрессия,
- Метод k-ближайших соседей,
- «Случайный лес»,
- Градиентный бустинг,
- Метод опорных векторов.

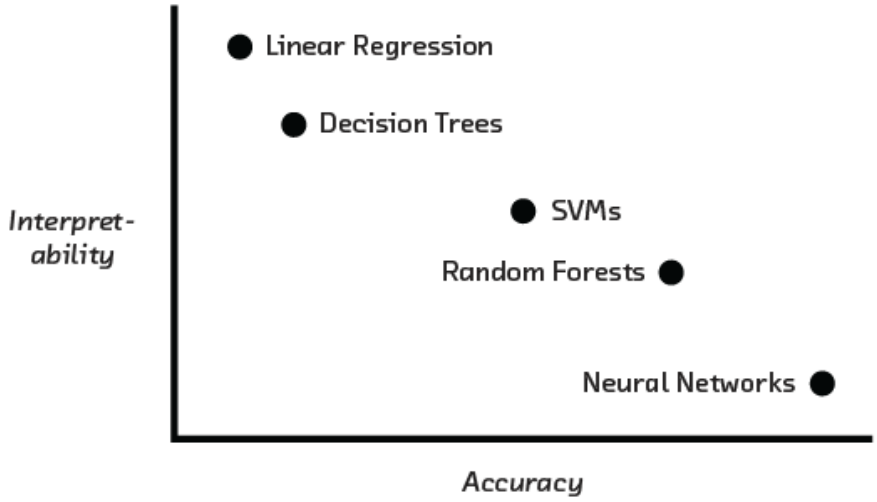


Рис. 8 – График интерпретируемости и точности алгоритмов

При обработке данных были отброшены признаки, в которых не хватает больше половины значений, однако остались признаки с меньшим количеством отсутствующих значений. Каждое пустое значение было заполнено методом медианного заполнения, который заменяет пустые значения средним значением соответствующего признака.

В результате обработанные данные были использованы для обучения модели. С помощью метрики средней абсолютной ошибки была произведена оценка качества каждой модели, а затем выбрана наиболее эффективная для дальнейшей оптимизации. Самый простой алгоритм линейной регрессии был исключен из-за значительно большего показателя средней абсолютной ошибки. Результат сравнения алгоритмов изображен на рисунке 9.

Так как базовый уровень ошибки составил 8.8829, а полученный результат оказался лучше, то можно сказать, что поставленная задача решается с помощью машинного обучения.

Лучшими моделями в сравнении оказались:

- Градиентный бустинг – 8.2285;
- К-ближайших соседей – 8.2318.

Хотя два алгоритма имеют небольшую разницу в результатах, был выбран алгоритм градиентного бустинга для его дальнейшей оптимизации.

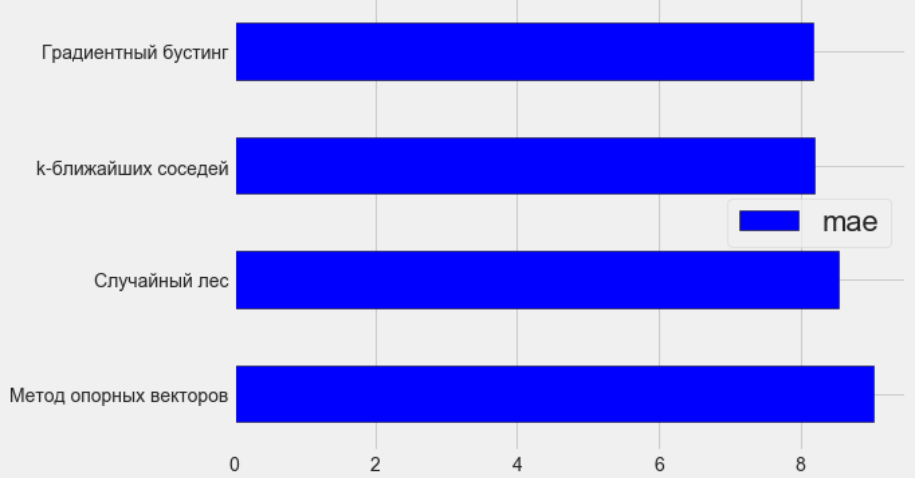


Рис. 9 – Сравнение моделей по средней абсолютной ошибке

Гиперпараметры модели могут быть рассмотрены как настройки алгоритма машинного обучения, настраиваемые перед обучением. Примерами могут служить число деревьев в случайном лесу или число соседей, используемых в регрессии k-ближайших соседей. Изменяя параметры, изменяем баланс между переобучением и недообучением. На рисунке 10 видно, как недообучение и переобучение модели влияют на прогнозы [20].

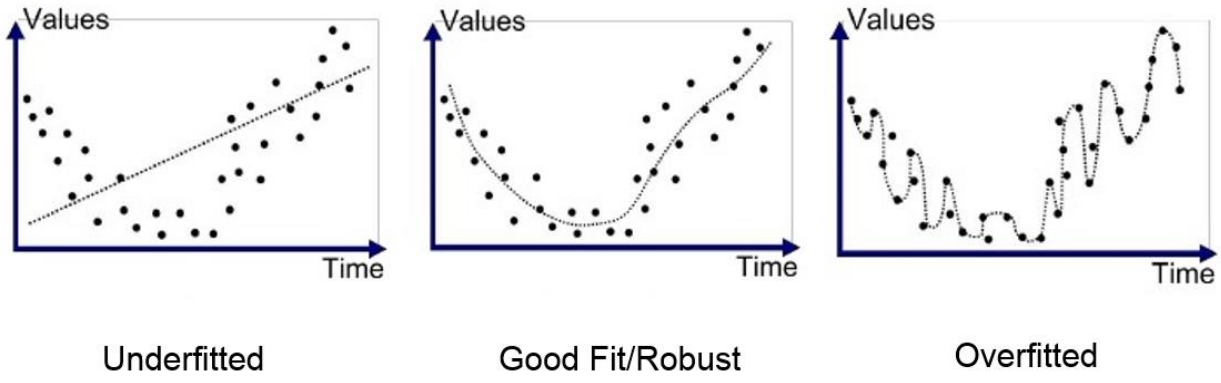


Рис. 10 – График недообучения (слева) и график переобучения (справа)

Метод настройки гиперпараметров был реализован с помощью случайного поиска с перекрестной проверкой. Перекрестная проверка – это метод оценки модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз. Таким образом, каждая из k частей данных используется для тестирования [21].

Весь процесс перекрестной проверки был выполнен следующим образом:

- задана сетка гиперпараметров (внешняя конфигурация по отношению к модели, значения которой невозможно оценить по данным) [22];
- случайно была выбрана комбинация гиперпараметров;
- создана модель с использованием этой комбинации;
- оценивается полученная средняя ошибка работы модели с помощью перекрестной проверки;
- принято решение, какие гиперпараметры дают лучший результат.

В данном случае используется регрессионная модель на основе градиентного бустинга. Метод является сборным (состоит из множества учеников), в данном случае из отдельных деревьев решений. Ученики в градиентном бустинге обучаются последовательно, то есть каждый последующий ученик учитывает ошибки предыдущего.

В выбранной модели были настроены: способ минимизации функции потерь, количество используемых слабых деревьев решений, минимальное количество примеров в узле дерева решений, минимальное количество для разделения узла дерева решений и максимальное количество признаков для разделения узлов. Для определения лучшей настройки были использованы 29 различных гиперпараметров.

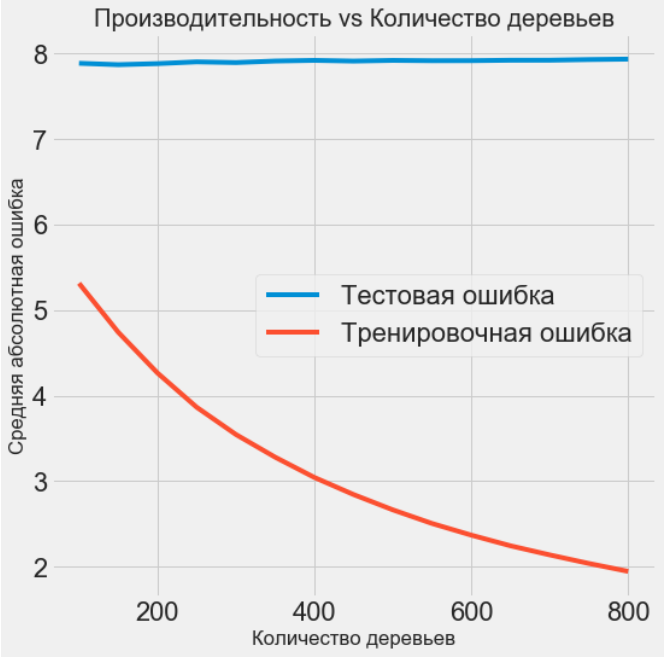


Рис. 11 – Зависимость количества деревьев и MAE

На рисунке 11 показано, как изменение количества деревьев (оценщиков) с сохранением параметров других настроек влияет на абсолютную ошибку.

Из рисунка 11 видно, что ошибка обучения значительно ниже, чем ошибка тестирования, что показывает применимость данной модели для решения поставленной задачи. Более того, по мере увеличения количества деревьев количество подгонки увеличивается. Как ошибка на тестовой выборке, так и ошибка на обучающей выборке уменьшаются по мере увеличения числа деревьев, но ошибка на обучающей выборке уменьшается быстрее.

Основываясь на результатах перекрестной проверки, лучшая модель использует 100 деревьев и достигает ошибки перекрестной проверки под 8. Это указывает на то, что средняя оценка перекрестной проверки оценки студента находится в пределах 8 баллов от истинного ответа.

Ошибка базовой модели на тестовой выборке: MAE=8.2285, ошибка финальной модели на тестовой выборке: MAE=7.9807. Гиперпараметрическая настройка помогла улучшить точность модели на 3%.

РЕЗУЛЬТАТЫ РАБОТЫ

На рисунках 12 и 13 показано, как были спрогнозированы баллы студентов по сравнению с реальными значениями. Как видно, спрогнозированные данные на финальной модели были более приближены к реальным значениям в отличие от базовой (ненастроенной модели). Также на рисунках 14 и 15 можно заметить, что количество предсказаний с большой ошибкой стало меньше по сравнению с базовой моделью. На рисунках 16 и 17 продемонстрирована плотность прогнозируемых и реальных значений. По рисункам видно, что плотность баллов в районе 80 баллов стала меньше.

Самым значимым признаком при определении оценки были: количество друзей, интересных страниц, фотографий профиля, подписчиков. На рисунке 18 видно, что среди интересных страниц большее влияние произвели группы с тематикой: программирование, юмор, креативная работа, «соседство».

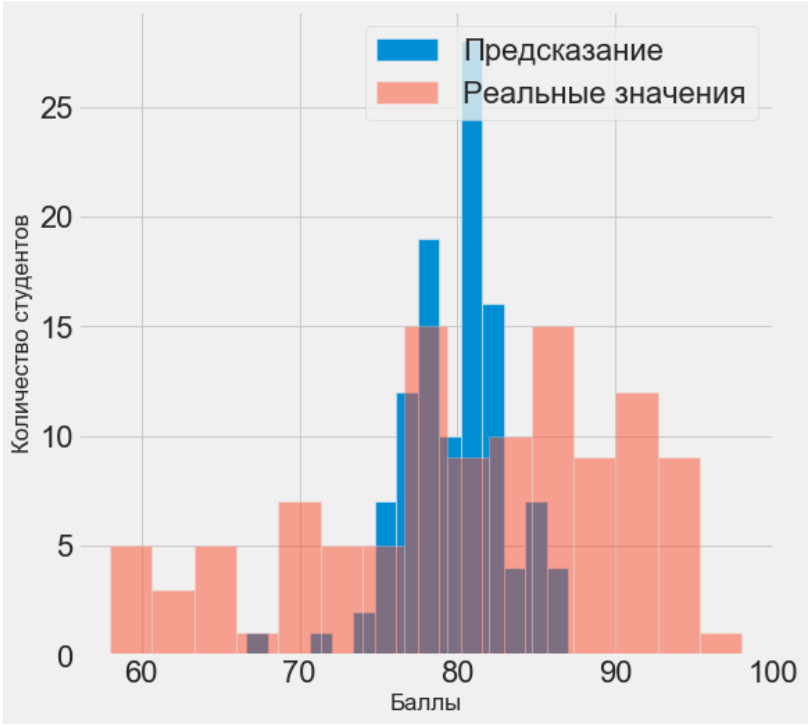


Рис. 12 – Распределение баллов на базовой модели

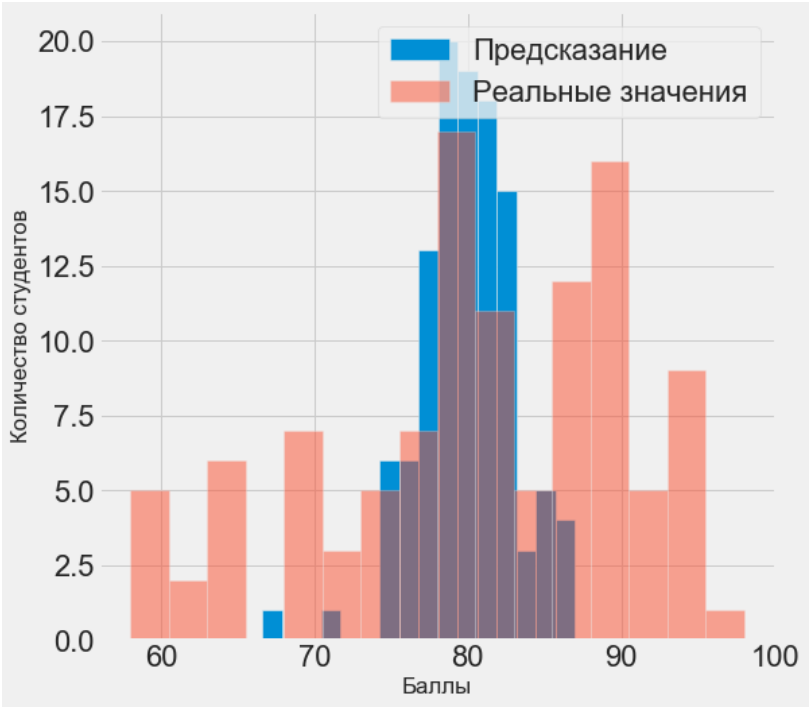


Рис. 13 – Распределение баллов финальной модели

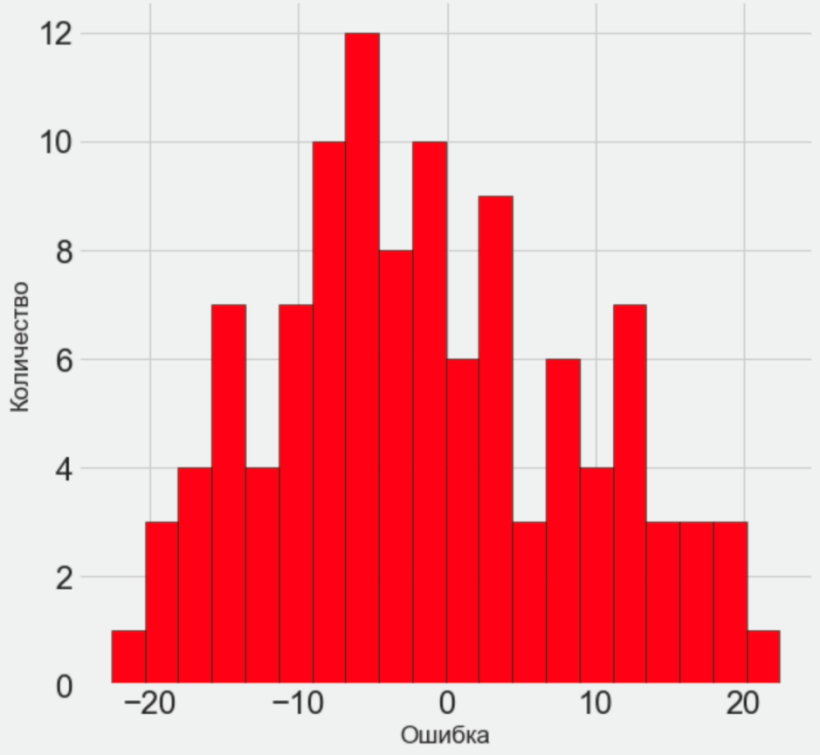


Рис. 14 – Гистограмма погрешности для базовой модели

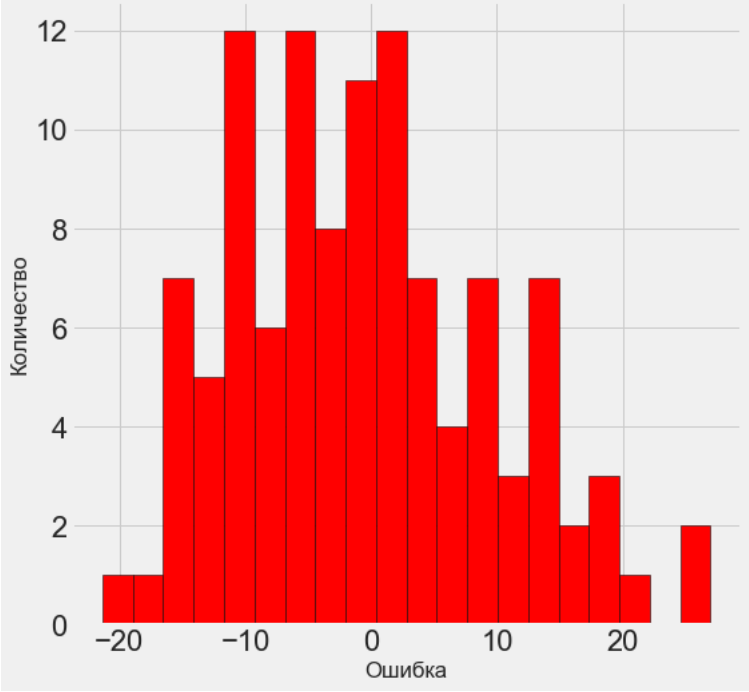


Рис. 15 – Гистограмма погрешности для финальной модели

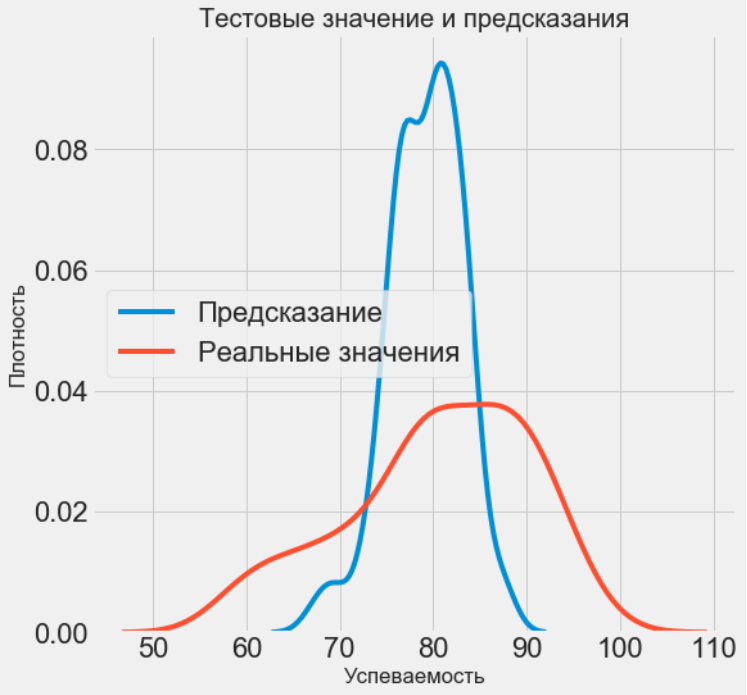


Рис. 16 – Плотность прогнозных и реальных значений базовой модели

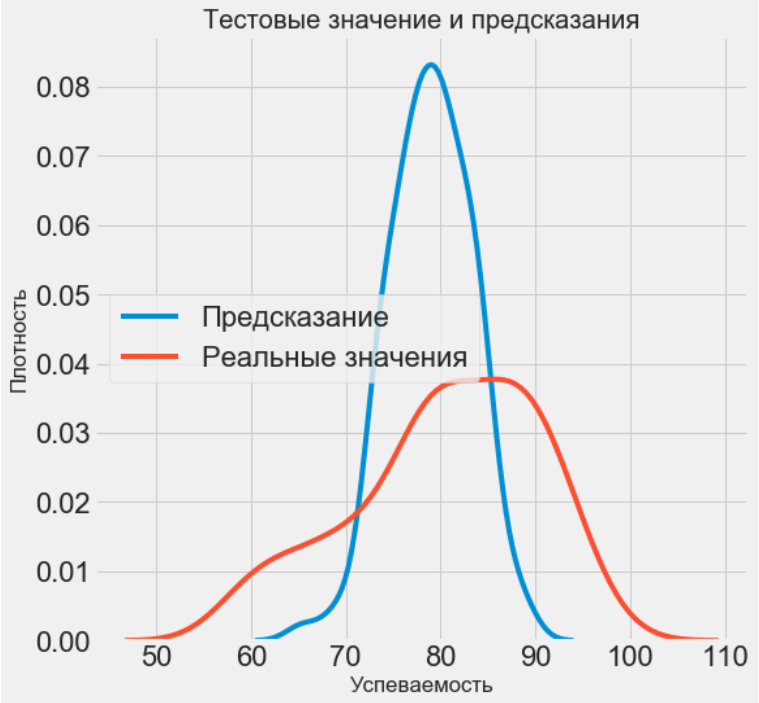


Рис. 17 – Плотность прогнозных и реальных значений финальной модели

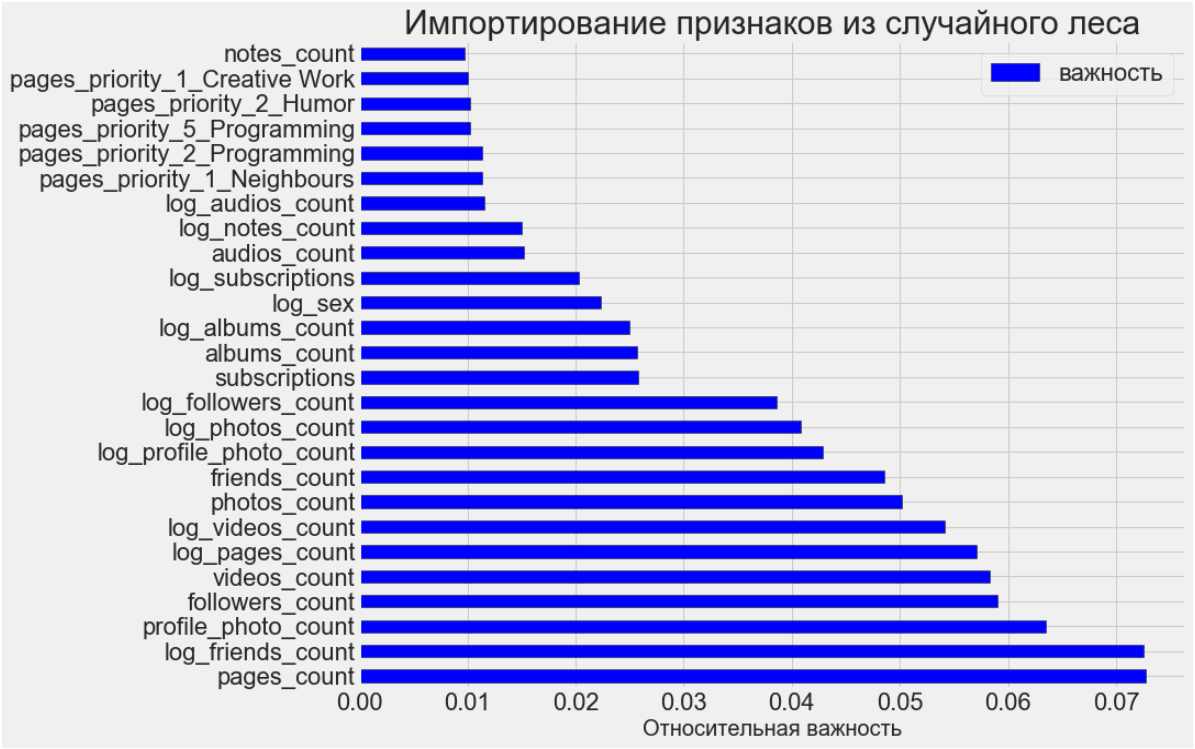


Рис. 18 – Относительная важность признаков

ЗАКЛЮЧЕНИЕ

В статье рассмотрена разработанная модель машинного обучения, которая выявляет взаимосвязь индивидуальных характеристик учащихся и их академической успеваемости, а также прогнозирует средний балл успеваемости по данным характеристикам.

В результате обучения модели были выявлены признаки, больше всего повлиявшие на составление вывода по полученным данным. К этим признакам относятся количественные признаки: подписки на популярные личности, друзья, фото профиля, а также категориальные признаки: подписки на интересные страницы.

Для улучшения качества модели в дальнейшем могут учитываться дополнительные параметры, такие, как активность в группах, время, проводимое в сети, и др. Кроме того, для получения дополнительной информации могут быть рассмотрены другие социальные сети.

Благодарности

Авторы выражают глубокую признательность младшему научному сотруднику НИЛ «Медицинская информатика» Высшей школы информационных технологий и интеллектуальных систем Алимовой Ильсеяр Салимовне за оказанную помощь в проведении данного исследования.

СПИСОК ЛИТЕРАТУРЫ

1. *Pritchard M.* Using Emotional and Social Factors to Predict Student Success // Journal of College Student Development. 2003. V. 44, No 1. P. 18–28.

2. *What your Facebook likes say about you*. URL: https://www.cbc.ca/ news/technology/facebook-likes-like-a-gift-1.3893298.

3. *Психометрический вступительный экзамен в Израиле* // Официальный сайт путеводителя по Израилю. URL: https://guide-israel.ru/country/ 37376-psixometricheskij- vstupitelnyj- ekzamen/.

4. *Shuotian B., Tingshao Z., Li C.* Big-Five Personality Prediction Based on User Behaviors at Social Network Sites // Cornell University, Tech. Rep. 2012.

5. *Friedrichsen M., Mühl-Benninghaus W.* Handbook of Social Media Managment Value Chain and Business Models in changing media marketing. Springer-Verlag Berlin Heidelberg, 2013. 880 p.

6. *Junco R.* Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // Computers in Human Behavior. 2012. V. 28, No 1. P. 187–198.

7. *Junco R.* The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement Received // Magazine Computers & Education. 2012. V. 58, No 1. P. 162–171.

8. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Magazine PNAS. 2013. V. 110, No 15. P. 5802–5805.

9. *Мацута В.В., Киселев П.Б., Фещенко А.Б., Гойко В.Л., Сузанова Е.А., Степаненко А.А.* Методы и инструменты выявления перспективных абитуриентов в социальных сетях // Открытое и дистанционное образование. 2017. № 4. С. 45–52.

10. *Penetration of leading social networks in Russia as of 4th quarter 2017* //

Statistica. URL: https://www.statista.com/statistics/284447/russia-social-network-penetration/.

11. *Мотивы проявления студентами колледжей социальной активности в социальных сетях: регионального аспекта* // Электронный научный архив УрФУ. URL: http://elar.urfu.ru/bitstream/ 10995/59123/1/978-5-91256-403-1_2018_053.pdf.

12. *Социальная сеть Вконтакте.* URL: https://vk.com/.

13. *Политика конфиденциальности VK.com // Социальная сеть Вконтакте.* URL: https://vk.com/privacy.

14. *VK.com python API wrapper* // GitHub. URL: https://github.com/ voronind/vk.

15. *Kaggle.* URL: https://www.kaggle.com/.

16. *What are outliers in the data* // Engineering statistics handbook. URL: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

17. *Histograms and density plots in python* // Towards data science. URL: https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0.

18. *How to normalize and standardize your machine learning data in weka* // Machine learning mastery. URL: https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/.

19. *Generalized Linear Models* // Scikit-learn. URL: https://scikit-learn.org/ stable/supervised_learning.html.

20. *Overfitting vs underfitting: a conceptual explanation* // Towards data science. URL: https://towardsdatascience.com/overfitting-vs-underfitting-a-conceptual-explanation-d94ee20ca7f9.

21. *Что такое кросс-валидация* // Data Science. URL: http://datascientist. one/cross-validation/.

22. *What is the difference between a parameter and a Hyperparameter?* // Machine Learning Mastery. URL: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

MACHINE LEARNING METHODS FOR DETERMINING THE RELATIONSHIP BETWEEN ACADEMIC SUCCESS AND DATA OF SOCIAL NETWORK PROFILE

I. R. Ikhsanov[1], I. S. Shakhova[2]

Higher School of Information Technologies and Intelligent Systems, Kazan (Volga region) Federal University

[1]ilias.ihsanov@gmail.com, [2]is@it.kfu.ru

Abstract

The paper is aimed to propose the machine learning model for determining the relationship between data of social network profile and academic success of students and predicting the success using the data.

Keywords: *machine learning, social networks, psychometrics, academic success, education, abiturient*

REFERENCES

1. *Pritchard M.* Using Emotional and Social Factors to Predict Student Success // Journal of College Student Development. 2003. V. 44, No 1. P. 18–28.

2. *What your Facebook likes say about you*. URL: https://www.cbc.ca/news/technology/facebook-likes-like-a-gift-1.3893298.

3. *Psihometricheskij vstupitel'nyj ekzamen v Izraile //* Oficial'nyj sajt putevoditelya po Izrailyu. URL: https://guide-israel.ru/country/37376-psixometricheskij-vstupitelnyj-ekzamen/.

4. *Shuotian B., Tingshao Z., Li C.* Big-Five Personality Prediction Based on User Behaviors at Social Network Sites // Cornell University, Tech. Rep. 2012.

5. *Friedrichsen M., Mühl-Benninghaus W.* Handbook of Social Media Managment Value Chain and Business Models in changing media marketing. Springer-Verlag Berlin Heidelberg, 2013. 880 p.

6. *Junco R.* Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // Computers in Human Behavior. 2012. V. 28, No 1. P. 187–198.

7. *Junco R.* The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement Received // Magazine

Computers & Education. 2012. V. 58, No 1. P. 162–171.

8. *Kosinski M., Stillwell D., Graepel T.* Private traits and attributes are predictable from digital records of human behavior // Magazine PNAS. 2013. V. 110, No 15. P. 5802–5805.

9. *Macuta V.V., Kiselev P.B., Feshchenko A.B., Gojko V.L., Suzanova E.A., Stepanenko A.A.* Metody i instrumenty vyyavleniya perspektivnyh abiturientov v social'nyh setyah // Otkrytoe i distancionnoe obrazovanie. 2017. No 4. S. 45–52.

10. *Penetration of leading social networks in Russia as of 4th quarter 2017* // Statistica. URL: https://www.statista.com/statistics/284447/russia-social-network-penetration/.

11. *Motivy proyavleniya studentami kolledzhej social'noj aktivnosti v social'nyh setyah: regional'nogo aspekta* // Elektronnyj nauchnyj arhiv UrFU. URL: http://elar.urfu.ru/bitstream/ 10995/59123/1/978-5-91256-403-1_2018_053.pdf.

12. *Vkontakte.* URL: https://vk.com/.

13. *Politika konfidencial'nosti VK.com //* Vkontakte. URL: https://vk.com/privacy.

14. *VK.com python API wrapper* // GitHub. URL: https://github.com/voronind/vk.

15. *Kaggle.* URL: https://www.kaggle.com/.

16. *What are outliers in the data* // Engineering statistics handbook. URL: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

17. *Histograms and density plots in python* // Towards data science. URL: https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0.

18. *How to normalize and standardize your machine learning data in weka* // Machine learning mastery. URL: https://machinelearningmastery.com/normalize-standardize-machine-learning-data-weka/.

19. *Generalized Linear Models* // Scikit-learn. URL: https://scikit-learn.org/stable/supervised_learning.html.

20. *Overfitting vs underfitting: a conceptual explanation* // Towards data science. URL: https://towardsdatascience.com/overfitting-vs-underfitting-a-conceptual-explanation-d94ee20ca7f9.

21. *Chto takoe cross-validatsiya* // Data Science. URL: http://datascientist.one/cross-validation/.

22. *What is the difference between a parameter and a Hyperparameter?* // Machine Learning Mastery. URL: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

СВЕДЕНИЯ ОБ АВТОРАХ

ИХСАНОВ Ильяс Раисович – бакалавр Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета по направлению «Прикладная информатика».

Ilias Raisovich IKHSANOV, Bachelor of Science in Applied Informatics from the Higher School of Information Technologies and Intelligent Systems, Kazan (Volga region) Federal University.

email: ilias.ihsanov@gmail.com

ШАХОВА Ирина Сергеевна – ассистент кафедры программной инженерии Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета. Сфера научных интересов – мобильные приложения, цифровые образовательные системы, индивидуализация образования, мобильное обучение.

Irina Sergeevna SHAKHOVA – teacher of the Higher School of Information Technologies and Intelligent Systems, Kazan Federal University. Research interests include mobile applications, digital educational systems, individualization in education, mobile learning.

email: is@it.kfu.ru

Материал поступил в редакцию 20 июня 2019 года

УДК 002.01

О ЛИЦЕНЗИОННОМ ДОГОВОРЕ НА ИЗДАНИЕ СЛУЖЕБНОГО ПРОИЗВЕДЕНИЯ

Т. А. Полилова

Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»

polilova@keldysh.ru

Аннотация

В соответствии с Гражданским кодексом РФ научное произведение является результатом интеллектуальной деятельности, которому предоставляется государственная охрана. Автору научного произведения (в частности, статьи) принадлежат право авторства, право на имя и иные неимущественные права. Если статья создается в рамках выполнения автором своих служебных обязанностей, исключительное право на служебное произведение принадлежит работодателю.

С согласия работодателя автор заключает с издательством лицензионный договор на опубликование статьи в научном издании на условиях, предложенных издателем. Заключение лицензионного договора не влечет за собой переход исключительного права к издателю. Даже если работодатель поручил автору заключить авторский договор с издателем на условиях исключительной лицензии, работодатель сохраняет за собой право использовать произведение, в том числе, право опубликовать произведение на своем сайте.

За автором (правообладателем) навсегда сохраняется право создавать производные произведения. Нередко навязываемые издателем условия лицензионного договора, ограничивающие право автора создавать произведения на основе ранее опубликованной статьи, ничтожны (т. е. не имеют юридической силы).

Опубликование автором производных произведений, содержащих фрагменты текста из предыдущих статей автора, не должно огульно считаться

нарушением издательской этики. Термины «самоплагиат», «автоплагиат» являются некорректными.

В Гражданском кодексе РФ закреплен механизм простых (неисключительных) лицензий, позволяющий нескольким издателям опубликовать статью без ее переработки. Опубликование статьи в нескольких изданиях — это один из законных способов реализации права автора (правообладателя) на широкое обнародование произведения.

***Ключевые слова**: научная публикация, служебное произведение, исключительное право, лицензионный договор, авторский договор, исключительная лицензия, простая лицензия, производное произведение, повторное использование текста, избыточная публикация*

ВВЕДЕНИЕ

Написанная автором научная статья является результатом интеллектуальной деятельности, которому предоставляется государственная охрана. Автору статьи принадлежат право авторства, право на имя и иные неимущественные права. Статья создается с целью ее дальнейшего опубликования в научном издании. Взаимоотношения автора и издателя регулируются положениями части IV Гражданского кодекса Российской Федерации (далее — ГК РФ) [1].

Для большинства ученых представление об авторском праве сводится к простым и привычным для всего научного сообщества правилам, относящимся скорее к области морально-этических норм. Во времена расцвета советской науки свои авторские права ученый мог реализовать путем опубликования научной статьи и в виде институтского препринта, и в научных журналах, не опасаясь какого-либо конфликта интересов. Сейчас при заключении авторского договора с издателем автору следует внимательно ознакомиться с предлагаемыми условиями, в частности, выяснить, не нарушает ли договор права автора и работодателя.

Издатель может включить в договор пункт о передаче ему исключительной лицензии на публикацию статьи. Большинство авторов незнакомы с законодательством в области авторского права в необходимом объеме и часто не вполне понимают смысл и последствия подписания договора

на предложенных условиях. Автор может задать издателю вопрос, насколько целесообразны и оправданы накладываемые издательством ограничения прав автора. Однако ответ, развернутый и обоснованный с позиций ГК РФ, он вряд ли получит.

Многие авторы не знают, нужно ли иметь согласие работодателя на опубликование статьи? Сохраняет ли автор право опубликовать свою ранее вышедшую статью, возможно, с некоторыми изменениями, в другом издании? Каков при этом минимально допустимый объем переработанного текста для новой статьи?

Ответы на эти вопросы требуют от автора определенной юридической грамотности в области авторского права. Научные сотрудники вынуждены погружаться в непривычный для них мир юридических понятий, испытывать на практике действие юридических норм.

СЛУЖЕБНОЕ ПРОИЗВЕДЕНИЕ

Каждый сотрудник заключает с организацией трудовой договор, где описаны права и обязанности работника в соответствии с его квалификацией, права и обязанности работодателя. В договоре должны также отражаться вопросы правового статуса интеллектуальной продукции, создаваемой в результате выполнения служебных обязанностей.

В ГК РФ вводится понятие «служебное произведение». Согласно статье 1295 ГК РФ исключительное право на служебное произведение принадлежит работодателю, если трудовым или гражданско-правовым договором между работодателем и автором не предусмотрено иное. При этом автору результата интеллектуальной деятельности принадлежат право на имя и иные личные неимущественные права. Статья 1228 ГК РФ гласит:

«Право авторства, право на имя и иные личные неимущественные права автора неотчуждаемы и непередаваемы. Отказ от этих прав ничтожен».

Что подразумевает «исключительное право»? Статья 1229 ГК РФ определяет, что правообладатель может использовать служебное произведение по своему усмотрению любым способом, не противоречащим закону.

Правообладатель может обнародовать служебное произведение, а также потребовать, чтобы при использовании служебного произведения указывались

имя или наименование правообладателя. Правообладатель может указать, какие действия над произведением допустимы, а какие действия запрещены.

Где и в какой форме работодатель может обнародовать служебное произведение? В работе [2] высказывается мнение, что наиболее подходящее место для обнародования (опубликования) служебных произведений, созданных по результатам проводимых в организации исследований, — сайт этой научной организации. Одна из возможных форм первичного представления результатов исследования — препринт. Согласно ГОСТ 7.60-2003, препринт — научное издание, содержащее материалы предварительного характера, опубликованные до выхода в свет издания, в котором они могут быть помещены.

Полезное качество препринта — лояльное отношение к опубликованным таким образом материалам со стороны журналов: на основе препринтов авторы могут в дальнейшем, в полном соответствии с определением препринта в ГОСТе, опубликовать статью в журнале. Типичный авторский договор, в котором имеется пункт о возможном принятии к опубликованию ранее опубликованных в виде препринта материалов, можно, например, видеть на сайте издателя журнала «Физическая мезомеханика» [3].

При заинтересованном отношении администрации научного учреждения к первичной публикации служебных произведений препринты могут приобрести черты полноценного научного издания. Примером такого издания является научное издание «Препринты ИПМ им. М.В. Келдыша» (http://library.keldysh.ru/preprints/).

В ИПМ им. М.В. Келдыша РАН препринты выпускаются уже не один десяток лет. Многие годы для сотрудников Института препринты являлись одним из основных способов опубликования результатов научной работы. Препринты поступали в библиотеку Института, передавались в Российскую книжную палату и попадали тем самым в ведущие библиотеки страны. С 2002 года на сайте Института стали размещаться на регулярной основе в открытом доступе полные тексты препринтов.

Сотрудники Института довольно быстро оценили преимущества онлайнового размещения препринтов на институтском сайте. Во-первых, автор

теперь может оперативно исправить ошибку, обнаруженную в препринте. Важно, что такая возможность у автора сохраняется в течение всей последующей творческой жизни. Во-вторых, автор может уточнить формулировки, среагировать на критические замечания и пожелания со стороны своих коллег по улучшению онлайнового текста. Автор получает возможность постоянно актуализировать свой текст, внося в него последние сведения о своих достижениях, достижениях коллег.

Преимущества в размещении препринта (статьи) на сайте Института очевидны в случае внесения в онлайновый вариант статьи мультимедийных иллюстраций. Видеоматериалы или анимация, внедренные в текст статьи, позволяют наглядно продемонстрировать читателю результаты моделирования процессов. Подобные средства научной визуализации только начинают осваиваться авторами, и не всегда автор самостоятельно справляется с возникающими техническими трудностями. Однако практика показала, что автору не составляет особого труда установить контакт с IT-специалистами Института, которые всегда готовы оказать содействие в реализации представления статьи на сайте [4].

Перечисленные возможности институтских препринтов оказались широко востребованными. Жаль, что они практически недоступны авторам, публикующим статьи в современных журналах.

Публикация служебных произведений на сайте организации является убедительным способом продемонстрировать творческий потенциал коллектива, показать качество проводимых научных исследований. Опубликованные на сайте результаты интеллектуальной деятельности являются своего рода неформальным отчетом организации.

ЛИЦЕНЗИОННЫЙ ДОГОВОР

Статье, опубликованной на сайте организации, полезно получить дополнительную оценку экспертного сообщества. На сегодняшний день наиболее эффективным, убедительным способом получения положительной экспертной оценки является публикация статьи в авторитетном научном журнале. Традиционно высоким авторитетом и высокими рейтинговыми показателями обладают академические журналы, привлекающие корпус

квалифицированных рецензентов из числа ведущих специалистов в области исследования. Статья, опубликованная в академическом журнале, приобретает высокий статус, как прошедшая строгое и ответственное рецензирование.

Рассмотрим следующую типичную ситуацию. Автор как работник научной организации написал статью в рамках выполнения своих должностных обязанностей. Автору, безусловно, принадлежат право авторства, право на имя. Исключительные права на произведение, как указано в статье 1295 ГК РФ, принадлежит работодателю. Автор планирует опубликовать свою работу в научном журнале. Какие шаги он обязан предпринять?

Для начала следует обратиться к трудовому договору с работодателем. Действия автора не должны нарушать права работодателя. Если в трудовом договоре указано, что использование служебного произведения возможно при получении согласия работодателя, то согласие от работодателя желательно получить в письменной форме. Работодатель может также потребовать от автора указать в служебном произведении аффилиацию автора как сотрудника работодателя, источник финансирования («работа выполнена в рамках бюджетного финансирования по теме …») и др. Работодатель вправе обязать автора получить экспертное заключение о возможности опубликования материалов в открытой печати.

Заручившись согласием работодателя, автор заключает с издателем журнала лицензионный договор.

Статья 1235 ГК РФ определяет, что по лицензионному договору обладатель исключительного права на результат интеллектуальной деятельности (лицензиар) обязуется предоставить другой стороне (лицензиату) право использования такого результата в пределах, предусмотренных в договоре. Право, прямо не указанное в договоре, не считается предоставленным лицензиату. В предмете лицензионного договора с издателем должны быть явно указаны результат интеллектуальной деятельности, а также способы использования такого результата.

При заключении лицензионного договора автору следует обратить внимание на формулировки положений о возможности заключения лицензиатом сублицензионных договоров. По сублицензионному договору

сублицензиату может быть предоставлено право использовать результат интеллектуальной деятельности только в пределах тех прав и тех способов использования, которые предусмотрены лицензионным договором для лицензиата (статья 1238 ГК РФ).

Лицензионный договор заключается только в письменной форме. Несоблюдение требования письменной формы влечет недействительность лицензионного договора. Территория действия договора — территории Российской Федерации, если в договоре территория не указана явным образом. Лицензионный договор не может заключаться бессрочно. Срок, на который заключается лицензионный договор, не может превышать срока действия исключительного права на результат интеллектуальной деятельности. Согласно ГК РФ, исключительное право на произведение действует в течение всей жизни автора и далее в течение семидесяти лет после смерти автора. В случае, когда в лицензионном договоре срок его действия не указан, договор считается заключенным на пять лет.

Статья 1236 ГК РФ определяет два вида лицензионного договора, предусматривающие два различных способа опубликования научного произведения:

1) на условиях простой (неисключительной) лицензии, когда лицензиату (издателю) предоставляется право использовать результат интеллектуальной деятельности с сохранением за лицензиаром (автором) права выдачи лицензий другим лицам;

2) на условиях исключительной лицензии, когда лицензиату предоставляется право использовать результат интеллектуальной деятельности без сохранения за лицензиаром права выдачи лицензий другим лицам.

Пункт 1.1 статьи 1236 ГК РФ уточняет: если лицензионным договором не предусмотрено иное, лицензия предполагается простой (неисключительной).

Опубликование произведения на условиях простой (неисключительной) лицензии не лишает автора возможности опубликовать работу в других изданиях. Автор, реализующий свое право на обнародование результата интеллектуальной деятельности, через опубликование произведения в нескольких изданиях расширяет свою читательскую аудиторию.

В то же время появление в интернете множества официально разрешенных копий произведения в некоторых случаях может вызывать определенные неудобства и даже ущемлять интересы правообладателей. Например, поисковые сервисы по запросу на поиск статьи выдадут адреса нескольких сайтов, где размещены копии статьи, при этом сайт правообладателя статьи никаким образом не будет отмечен в качестве основного.

Выход из этой ситуации лежит в использовании технологии задания метаданных статьи, специфицирующих текст статьи на сайте. Так, например, в наборе элементов метаданных Дублинского ядра [5] присутствуют конструкции Rights и RightsHolder. Указание с их помощью правообладателя статьи может выглядеть следующим образом:

<meta name="DC.rights.rightsHolder" content="Исключительные права на произведение принадлежат ИПМ им. М.В. Келдыша РАН http://keldysh.ru/" />

Поисковый сервис, обнаружив подобную конструкцию, вообще говоря, мог бы оценить принадлежность адреса (URL) очередного найденного результата поиска адресному пространству правообладателя, и у сайта правообладателя появился бы убедительный шанс попасть в первую строчку списка результатов.

ПРЕДОСТАВЛЕНИЕ ИСКЛЮЧИТЕЛЬНОЙ ЛИЦЕНЗИИ

Среди издателей научных журналов нередко бытует философия рыночного подхода, когда издатель в интересах поддержания своего бизнеса требует от автора заключения авторского договора на условиях исключительной лицензии.

Одним из примеров предоставления издателю исключительной лицензии является типовой авторский договор издательства «Наука» [6], до недавнего времени издававшего академические журналы.

В авторском договоре читаем:

«Автор предоставляет Лицензиату исключительную лицензию на использование Статьи следующими способами».

По данному договору издатель получает от автора исключительную лицензию на воспроизведение в любой материальной форме, распространение, доведение статьи до всеобщего сведения.

С какой целью здесь требуют исключительной, а не простой лицензии? По-видимому, издатель журнала надеется, что исключительная лицензия позволит ему уверенно чувствовать себя на рынке научных публикаций, не встречая конкуренции со стороны других издательств. Но насколько исключительная лицензия защищает монополию издателя на опубликование статьи?

Заключение лицензионного договора не влечет за собой переход исключительного права к лицензиату. Даже если работодатель поручил автору заключить авторский договор с издателем на условиях исключительной лицензии, работодатель сохраняет за собой исключительные права (пункт 1 статьи 1233 ГК РФ):

«Заключение лицензионного договора не влечет за собой переход исключительного права к лицензиату».

Здесь и далее речь идет об отношениях издателя и правообладателей в Российской Федерации. Законодательные нормы других стран, возможно, допускают иные отношения.

По причине терминологической близости двух понятий «исключительное право» и «исключительная лицензия» часто возникает путаница в понимании этих юридических терминов. Нужно помнить, что «исключительное право» — понятие существенно более широкое. Понятие «исключительная лицензия» существует только в контексте лицензионного договора и означает лишь одно: переданная лицензиаром исключительная лицензия на использование произведения лишает лицензиара права выдавать лицензии другим лицам.

Что именно включает в себя исключительное право на произведение? Согласно пункту 1 статьи 1270 ГК РФ правообладатель может использовать произведение в любой форме и любым не противоречащим закону способом:

«1. Автору произведения или иному правообладателю принадлежит исключительное право использовать произведение в соответствии со статьей 1229 настоящего Кодекса в любой форме и любым не противоречащим закону способом (исключительное право на произведение), в том числе способами, указанными в пункте 2 настоящей статьи».

В пункте 2 статьи 1270 перечислены следующие способы использования произведения:

- воспроизведение произведения, т. е. изготовление одного и более экземпляра произведения или его части в любой материальной форме;
- распространение произведения;
- импорт (ввоз на таможенную территорию РФ) оригинала или экземпляров произведения в целях распространения;
- перевод или другая переработка произведения;

и др.

Передав издателю исключительную лицензию на использование статьи, правообладатель (работодатель) сохраняет за собой исключительные права на произведение. Такие виды использования произведения, как «воспроизведение» и «распространение» произведения, подразумевают возможность, например, опубликования произведения на сайте правообладателя (работодателя). Таким образом, издатель, получивший исключительную лицензию на использование произведения, монополистом не становится.

Рассмотрим другую ситуацию. Работодатель передает исключительные права на произведение автору. Но и в этом случае работодатель не лишается возможности использовать служебное произведение в соответствии с пунктом 3 статьи 1295 ГК РФ:

«В случае, если в соответствии с пунктом 2 настоящей статьи исключительное право на служебное произведение принадлежит автору, работодатель имеет право использования соответствующего служебного произведения на условиях простой (неисключительной) лицензии …».

Из этого положения следует, что, если автор, получив от работодателя исключительное право на служебное произведение, заключает с издателем лицензионный договор о передаче произведения на условиях исключительной лицензии, то это не гарантирует, что произведение не будет повторно опубликовано. Такое опубликование вправе выполнить как автор, так и работодатель, реализующий свое право использования служебного произведения на условиях простой (неисключительной) лицензии.

Вернемся к тексту договора [6]. В разделе 3 «Гарантии сторон» читаем:

«3.1. Автор гарантирует, что:

- он является законным правообладателем Статьи;

- на момент вступления в силу настоящего Договора Автору ничего не известно о правах третьих лиц, которые могли быть нарушены предоставлением исключительной лицензии на использование Статьи по Договору».

К сожалению, приведенные в договоре [6] формулировки гарантий могут вводить автора в заблуждение. Автор не обязан, вообще говоря, разбираться в юридических тонкостях. Ему, например, известно, что он обладает правом авторства. Но он может не знать, что из авторства, строго говоря, не следует возникновение исключительного права на произведение.

Договор требует от автора гарантии того, что ему ничего не известно о правах третьих лиц. Не погруженный в юридические вопросы автор без всякого злого умысла может искренне подтвердить свою неосведомленность о правах третьих лиц в силу своей некомпетентности в вопросах авторского права. Хотя в договоре издатель на самом деле намерен возложить на автора ответственность в случае нарушения интересов третьих лиц.

Как уже отмечалось, исключительное право на статью как служебное произведение принадлежит работодателю. Если работодатель не заключил с автором договора об отчуждении исключительного права на созданное служебное произведение и не поручил в письменном виде автору заключать договор с издательством на условиях исключительной лицензии, то автор не может гарантировать, что права третьих лиц не будут нарушены.

Добросовестный издатель, ориентируясь на неподготовленных в юридических вопросах авторов, должен был бы, во-первых, разъяснить понятия, используемые в договоре. Во-вторых, издатель должен затребовать от автора подтверждающие документы: или договор-отчуждение исключительных прав в пользу автора, или договор-поручение от работодателя о возможности заключения авторского договора с издателем на условиях исключительной лицензии, или доказательства написания статьи вне рамок трудовой деятельности. Без подтверждающих документов договор издателя и автора может быть оспорен в суде, поскольку он, возможно, нарушает права работодателя.

ПРОИЗВОДНОЕ ПРОИЗВЕДЕНИЕ

Производное произведение появляется в результате переработки исходного произведения. Переработка является одним из способов реализации исключительного права автора (правообладателя) на произведение (пункт 9 статьи 1270 ГК РФ). Гражданский кодекс не регламентирует конкретные допустимые способы и объемы переработки научного произведения. Целесообразность создания производного произведения полностью относится к компетенции и ответственности автора (правообладателя) произведения. Производным произведением является, в частности, перевод произведения на другой язык.

У автора нет никаких ограничений в создании производных произведений на основе исходного произведения. Никакие заключенные лицензионные договора и даже заключенный договор об отчуждении исключительного права на произведение не лишают автора права создавать производные произведения. Пункт 4 статьи 1233 ГК РФ гласит:

Условия договора об отчуждении исключительного права или лицензионного договора, ограничивающие право гражданина создавать результаты интеллектуальной деятельности определенного рода или в определенной области интеллектуальной деятельности либо отчуждать исключительное право на такие результаты другим лицам, ничтожны.

Таким образом, в отношении будущих произведений Гражданский кодекс ограждает автора от договоров, ущемляющих интересы автора. Вновь повторим: в Российской Федерации издатель в договоре не может потребовать от автора не создавать в будущем производных произведений.

Производное произведение может быть создано автором специально для функционирования в интернете. Такое произведение реализуется не в виде текстового файла, а на специальной технологической платформе. По сравнению с привычным текстом произведение может обладать новыми качествами: гибкой визуализацией текста в технике адаптивного дизайна, мультимедийными иллюстрациями, онлайновыми вычислениями и т. д.

Автор может создавать свое произведение в технологии живых публикаций. Автор живой публикации берет на себя обязанность не только

постоянно совершенствовать свое произведение, но и следить за событиями в исследуемой области и систематически отражать все новое в своем онлайновом тексте [8]. В какой-то момент автор может опубликовать в журнале статью — зафиксированный временной срез живой публикации — и далее продолжать развивать свое произведение. Разумеется, все изменения, вносимые автором в живую публикацию, должны протоколироваться. Накопив новые, достаточно интересные факты и результаты, автор может вновь опубликовать очередной временной срез своей непрерывно развивающейся работы. Таким образом, первоначальное произведение может порождать семейство производных произведений, предназначенных для публикации в журналах.

Пример такой живой публикации — статья М.М. Горбунова-Посадова «Интернет-активность как обязанность ученого» (http://keldysh.ru/gorbunov/duty. htm). На сайте живой публикации указываются журнальные статьи, опубликованные по материалам живой публикации — производные произведения (рис. 1). За десять лет существования этой живой публикации автор опубликовал в журналах две статьи:

- Горбунов-Посадов М.М. Интернет-активность как обязанность ученого // Информационные технологии и вычислительные системы. 2007, № 3. С. 88–93.
- Горбунов-Посадов М.М. Жизненный путь научной публикации // Информационные технологии и вычислительные системы. 2014, № 4. С. 79–88.

Приведем еще один пример визуализации семейства производных произведений, представленный на сайте сериального издания «Препринты ИПМ им. М.В. Келдыша» [9]. Авторам препринтов предоставляется возможность разместить ссылки на журнальные публикации, выполненные на основе выпущенного препринта (рис. 2).

Препринт «Летные испытания алгоритмов управления ориентацией микроспутника “Чибис-М”», авторы Иванов Д.С., Ивлев Н.А., Карпенко С.О., Овчинников М.Ю., Ролдугин Д.С., Ткачев С.С., был опубликован в 2012 г. На основе этого препринта в 2014 г. были подготовлены и опубликованы три производных произведения:

- M. Ovchinnikov, D. Ivanov, N. Ivlev, S. Karpenko, D. Roldugin, S. Tkachev, Development, integrated investigation, laboratory and in-flight testing of Chibis-M microsatellite ADCS // Acta Astronautica. 2014. V. 93. P. 23—33.
- Д.С. Иванов, Н.А. Ивлев, С.О. Карпенко, М.Ю. Овчинников, Д.С. Ролдугин, С.С. Ткачев. Результаты летних испытаний системы ориентации микроспутника Чибис-М // Космические исследования. 2014. Т. 52, № 3. С. 218—228.
- S. Ivanov, N.A. Ivlev, S.O. Karpenko, M.Yu. Ovchinnikov, D.S. Roldugin, S.S. Tkachev. The results of flight tests of an attitude control system for the Chibis-M microsatellite // Cosmic Research. 2014. V. 52, No 3. P. 205—215.

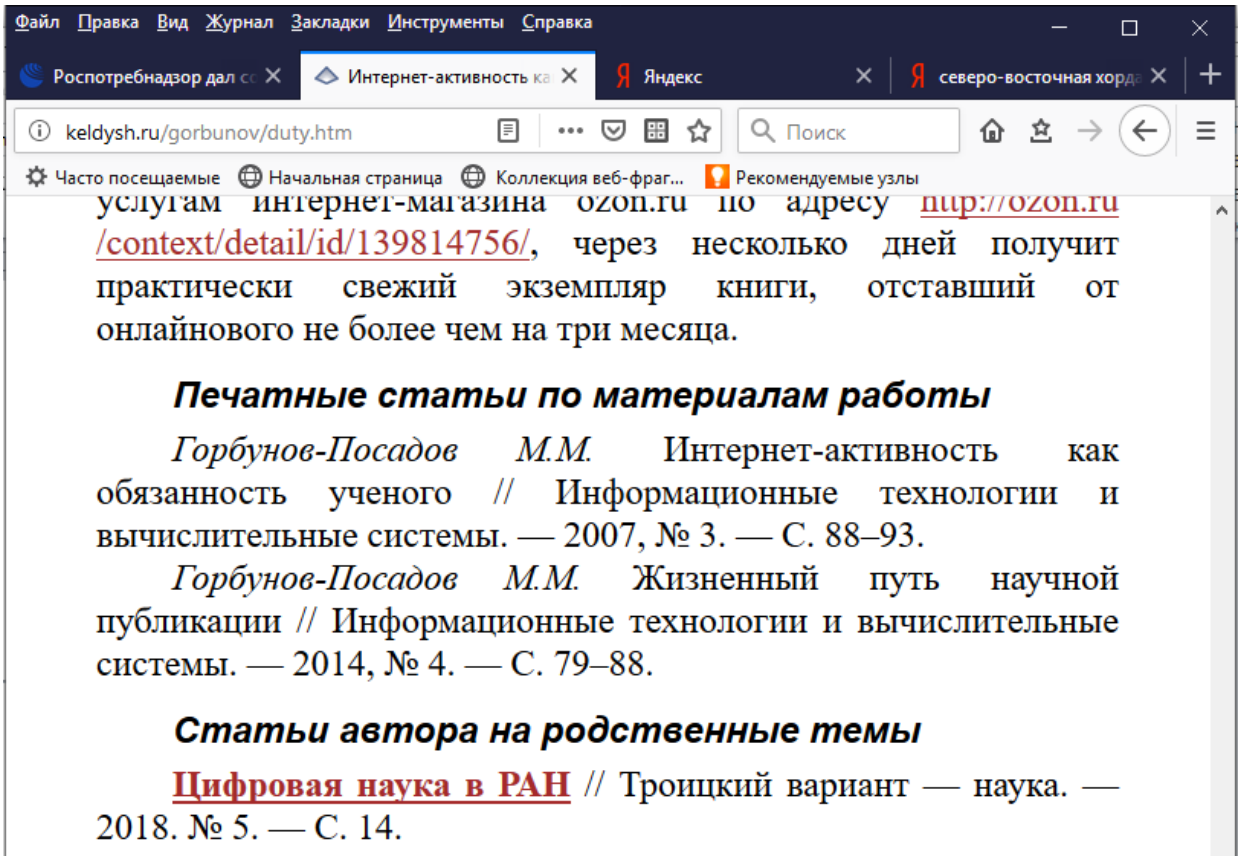


Рис. 1. Журнальные статьи, опубликованные по материалам живой публикации М.М. Горбунова-Посадова «Интернет-активность как обязанность ученого»

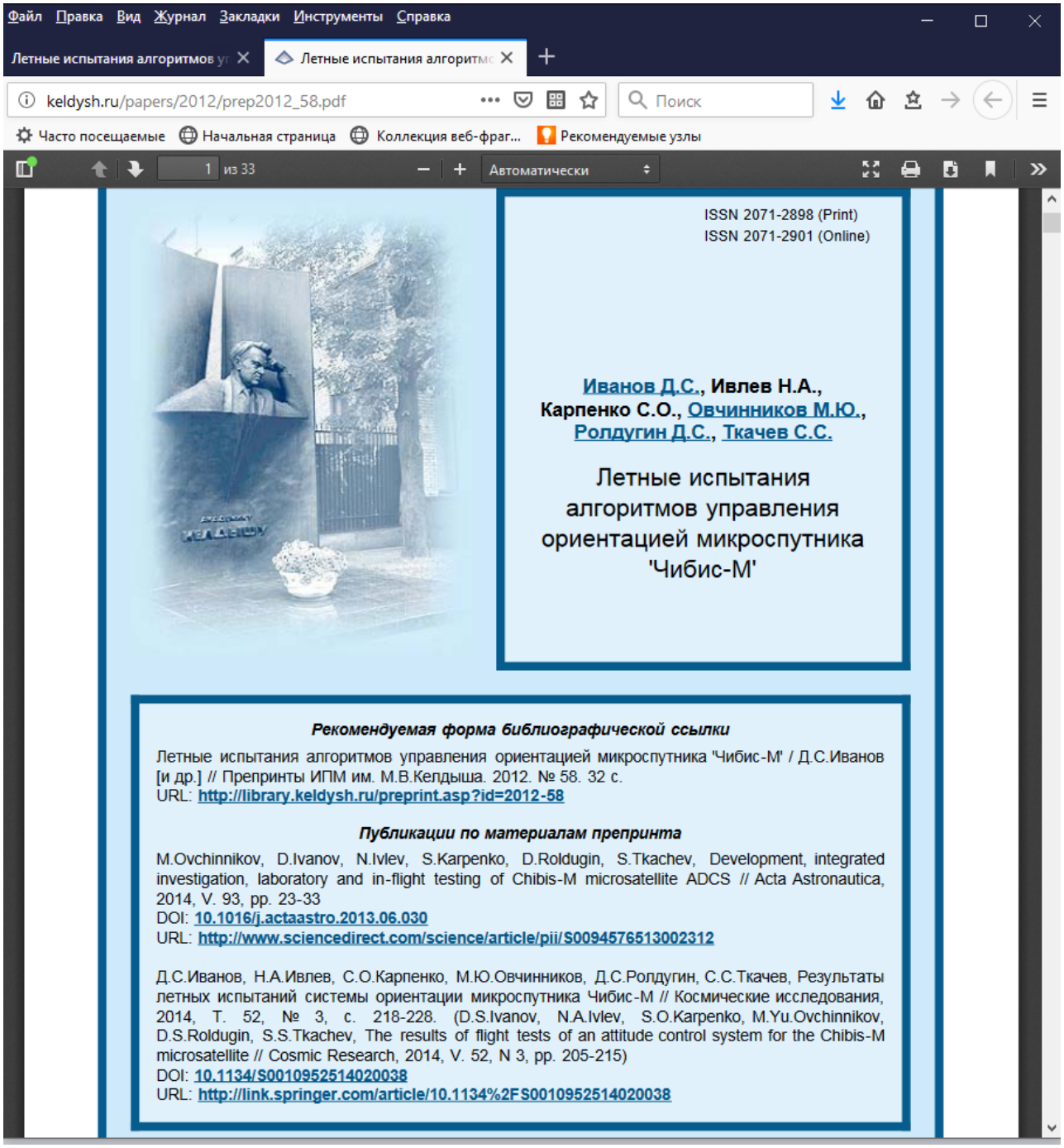


Рис. 2. Семейство производных произведений, созданных по материалам препринта, на сайте «Препринты ИПМ им. М.В. Келдыша»

Ссылки на производные произведения размещены на головной странице исходного препринта http://keldysh.ru/papers/2012/prep2012_58.pdf. Тем самым читатель препринта имеет возможность познакомиться с первичной публикацией и со всеми производными произведениями.

САМОПЛАГИАТ И МНОЖЕСТВЕННАЯ ПУБЛИКАЦИЯ

В предыдущем разделе был описан естественный процесс опубликования результатов научных исследований, в котором появляются производные произведения (статьи), предназначенные для опубликования в нескольких журналах.

В последнее время в научном и издательском сообществе развернулась дискуссия: этично или неэтично публиковать производные произведения.

Рассмотрим определения, данные в итоговом документе конференции «Проблемы качества научной работы и академический плагиат», состоявшейся 26 сентября 2018 г. в РГГУ [10]:

«Под множественной публикацией статьи (тиражированием, самоплагиатом, автоплагиатом) понимается перепечатка автором (авторами) собственных работ, не оправданная какими-либо объективными причинами (внесение изменений в текст статьи, перевод на другой язык, обращение к иной читательской аудитории, включение текста статьи в тематическую подборку или антологию) и без указания источника первоначальной публикации».

Развернутое и весьма взвешенное определение. Согласно этому определению, в частности, статья, опубликованная в журнале после выхода препринта, не является множественной публикацией: здесь определенно имеет место обращение к иной, вообще говоря, более широкой читательской аудитории.

Если автор вносит в первоначальную статью изменения, то измененная статья, следуя логике приведенного выше определения, также не должна иметь статус «множественной» или «избыточной» (читай — неэтичной) публикации. Однако остается открытым вопрос, каким образом формально измерить объем вносимых изменений или объем дублируемого текста, чтобы не нарушить объявленные вполне разумные этические нормы.

Очевидно, что любое определение вряд ли покроет все случаи, когда авторы прибегают к дублированию фрагментов текста. В некоторых этических кодексах используется другой термин — «повторное использование текста» (*text recycling*). Повторное использование текста не является безусловным признаком

нарушения этических норм. В этом случае редакторам предлагается внимательно рассмотреть целесообразность повторного появления текста из опубликованной ранее статьи. В то же время существует, например, такой жанр научного произведения, как диссертация, в которой автор даже обязан привести выдержки из ранее опубликованных статей. Таким способом автор диссертации доказывает, что в соответствии с требованиями ВАК все основные положения его работы были опубликованы.

На наш взгляд, практикуемые сейчас ограничения на опубликование и распространение статьи могут нарушать права автора (правообладателя) на обнародование своего произведения, т. е. доведение его до широкой общественности. Нередко за искаженной трактовкой этических норм скрывается обычная рыночная конкуренция издателей журналов. Связав руки автору в отношении свободного распространения произведения, издатель получает конкурентные преимущества от эксклюзивного опубликования материала. После выхода журнала многие издатели предлагают авторам повторно опубликоваться в тематических сборниках (см., например, предложение издательства «Наука» [11]), получая таким способом дополнительные доходы от повторного распространения и продажи ранее опубликованных статей.

В то же время нельзя полностью отбрасывать этические соображения о неприемлемости множественной публикации одной и той же работы. Но этические нормы вступают в действие только тогда, когда множественная публикация приобретает уродливые формы. Вряд ли можно считать нормальной ситуацию, когда одна и та же статья появляется в десятках или даже сотнях журналах. Здесь действительно налицо нарушение этики, поскольку читатель вынужден раз за разом натыкаться на идентичные тексты. В таком случае обычно легко просматривается попытка недобросовестного автора резко увеличить свои публикационные показатели.

11 апреля 2019 г. состоялся вебинар на тему «Добросовестность в научных исследованиях: основные виды нарушений и конфликты интересов в науке», докладчик — Е.Г. Гребенщикова, доктор философских наук, руководитель Центра научно-информационных исследований по науке, образованию и технологиям ИНИОН РАН. Как следует из презентации вебинара, дублирование

публикаций, т. е. опубликование статьи в нескольких журналах, отнесено к спорным исследовательским практикам.

Что касается автоплагиата — использования частей предыдущих публикаций в новой публикации, то позиция докладчика в этом вопросе достаточно взвешенная. Отмечается, что в некоторых ситуациях цитирование собственных опубликованных материалов следует признать приемлемым (например, дублирование в диссертации частей ранее опубликованных статей). Отношение к автоплагиату, по мнению докладчика, может зависеть от области знания, особенностей национальной исследовательской культуры и т. д.

Отметим, что призывы негативно оценивать авторские работы, в которых присутствуют фрагменты текста из других работ автора, входят в противоречие с нормами статей ГК РФ.

Возможно, близкие по содержанию произведения создаются автором с целью донести до той или иной конкретной аудитории свои результаты или оттенить и расставить новые акценты для улучшения восприятия полученных автором результатов. Те, кто обвиняют автора близких по содержанию статей в графоманстве и недобросовестном увеличении публикационной активности, должны приводить серьезные аргументы и доказательства того, что автор злоупотребляет своим правом на создание производных произведений.

На наш взгляд, термины «самоплагиат», «автоплагиат» являются некорректными. Негативный оттенок, который связан с использованием слова «плагиат», не должен сопутствовать естественной потребности автора создавать варианты своей статьи — производные произведения. Возможность создания производного произведения закреплена в статьях ГК РФ, посвященных авторскому праву.

ЗАКЛЮЧЕНИЕ

В настоящее время многие организации, выполняющие научные исследования, к сожалению, не проводят активной политики в области правовой охраны результатов интеллектуальной деятельности, не реализуют в полном объеме свои исключительные права на создаваемые произведения науки.

Считается обычной практика, когда автор из числа сотрудников организации самостоятельно, без письменного согласия или разрешения

организации, передает в журнал материалы статьи для опубликования и подписывает с издателем авторский договор на условиях, ущемляющих права и самого автора, и работодателя.

Немногие знают, что заключение лицензионного (авторского) договора ни при каких обстоятельствах не влечет за собой переход исключительного права к издателю. Даже если работодатель поручил автору заключить авторский договор с издателем на условиях исключительной лицензии, работодатель всегда сохраняет за собой право опубликовать произведение на своем сайте.

За автором (правообладателем) навсегда сохраняется право создавать производные произведения. Нередко навязываемые издателем условия лицензионного договора, ограничивающие право автора создавать произведения на основе ранее опубликованной статьи, ничтожны (т. е. не имеют юридической силы).

Опубликование автором производных произведений, содержащих фрагменты текста из его предыдущих статей, не должно огульно относиться к нарушениям издательской этики. Термины «самоплагиат», «автоплагиат», появившиеся в угоду вооружившимся программой Антиплагиат администраторам, не только являются юридически некорректными, но и во многих случаях недопустимо сковывают естественную творческую активность автора.

Развернувшаяся в научном сообществе дискуссия на тему, этично ли публиковать статью в нескольких изданиях, лишена смысла. К этической сфере может относиться только обязанность автора сообщить издателю о состоявшейся или же планируемой публикации статьи в другом издании. В Гражданском кодексе РФ закреплен механизм простых (неисключительных) лицензий: любой издатель может получить от правообладателя простую лицензию на опубликование статьи, в том числе без ее переработки. Опубликование статьи в нескольких изданиях — это один из закрепленных в законе естественных способов реализации права автора (правообладателя) на широкое обнародование произведения.

Благодарности

Автор выражает глубокую признательность Александру Константиновичу Петренко и Михаилу Юрьевичу Овчинникову за полезные и чрезвычайно точные замечания, сделанные по теме работы.

СПИСОК ЛИТЕРАТУРЫ

1. Гражданский кодекс Российской Федерации. URL: http://www.consultant.ru/document/cons_doc_LAW_5142/

2. *Полилова Т.А.* Инфраструктура научных публикаций // Препринты ИПМ им. М.В.Келдыша. 2009. № 15. 30 с.
URL: http://library.keldysh.ru/preprint.asp?id =2009-15

3. Институт физики прочности и материаловедения СО РАН. Типовой авторский договор с издателем журнала «Физическая мезомеханика».
URL: http:// www.ispms.ru/ru/153/

4. *Горбунов-Посадов М.М.* Цифровая наука в РАН // Троицкий вариант — наука. 2018. № 5. С. 14.
URL: https://trv-science.ru/2018/03/13/cifrovaya-nauka-v-ran/

5. Dublin Core Metadata Initiative. URL: http://dublincore.org/

6. Издательство «Наука»: лицензионный договор о предоставлении права использования статьи в научном журнале.
URL: https://naukapublishers.ru/ avtoram/obraztsy-dogovorov

7. *Полилова Т.А.* Лицензии для научных архивов открытого доступа // Препринты ИПМ им. М.В.Келдыша. 2018. No 233. 20 с.
URL: http://library. keldysh.ru/preprint.asp?id=2018-233

8. *Горбунов-Посадов М.М.* Живая публикация. М.: ИПМ им.М.В.Келдыша, 2011. Редакция от 01.05.2019.
URL: http://keldysh.ru/gorbunov/live.htm

9. Сайт научного сериального издания «Препринты ИПМ им. М.В. Келдыша».
URL: http://library.keldysh.ru/preprints/

10. Итоговый документ конференции «Проблемы качества научной работы и академический плагиат» (Москва, 26 сентября 2018, РГГУ)
URL: http://www.sib-science.info/ru/conferences/itogovyy-dokument-09102018

11. Новые возможности: издание сборников.
URL: https:// naukapublishers.ru/ history/novosti/novosti-izdatelstva/229-novye-vozmozhnosti-izdanie-sbornikov

ABOUT THE LICENSE AGREEMENT FOR THE WORK-FOR-HIRE PUBLICATION

T. A. Polilova

Keldysh Institute of Applied Mathematics Russian Academy of Sciences

polilova@keldysh.ru

Abstract

In accordance with the Civil code of the Russian Federation, a research paper is the result of intellectual activity, which is provided with state protection. The author of the research paper owns the right of authorship, the right to a name and other non-property rights. If the paper is created within the framework of the authors' implementation of their official duties, the exclusive right to the paper belongs to the employer.

With the consent of the employer, the author concludes a license agreement with the publisher for the publication of the paper on the terms proposed by the publisher. Signing of the license agreement does not entail the transfer of the exclusive right to the publisher. Even if the employer has instructed the author to enter into a copyright agreement with the publisher under an exclusive license, the employer reserves the right to use the work, including the right to publish the work on its website.

The author (copyright holder) always retains the right to create derivative works. Often imposed by the publisher terms of the license agreement, limiting the author's right to create works on the basis of previously published articles, have no legal force.

The publication by the author of derivative works containing fragments of the author's previous paper should not be considered as a violation of publishing ethics. The term "self-plagiarism" is incorrect.

The Civil code of the Russian Federation establishes a simple (non-exclusive) license that allows several publishers to publish an article without its processing. The publication of article in several editions — this is one of the legal ways of realization of the rights of the author (copyright holder) on a wide publication of a work.

Keywords*: research paper, work-for-hire, exclusive right, license agreement, copyright agreement, exclusive license, simple license, derivative work, text recycling, redundant publication*

REFERENCES

1. Grazhdanskii kodeks Rossiiskoi Federatsii. URL: http://www.consultant.ru/document/cons_doc_LAW_5142/

2. *Polilova T.A.* Infrastruktura nauchnykh publikatsii // Preprinty IPM im. M.V.Keldysha. 2009. № 15. 30 s. URL: http://library.keldysh.ru/preprint.asp?id=2009-15

3. Institut fiziki prochnosti i materialovedeniia SO RAN. Tipovoi avtorskii dogovor s izdatelem zhurnala «Fizicheskaia mezomekhanika». URL: http://www.ispms.ru/ru/153/

4. *Gorbunov-Posadov M.M.* Tsifrovaia nauka v RAN // Troitskii variant — nauka. 2018. № 5. S. 14. URL: https://trv-science.ru/2018/03/13/cifrovaya-nauka-v-ran/

5. Dublin Core Metadata Initiative. URL: http://dublincore.org/

6. Izdatelstvo «Nauka»: litsenzionnyi dogovor o predostavlenii prava ispolzovaniia stati v nauchnom zhurnale. URL: https://naukapublishers.ru/ avtoram/ obraztsy-dogovorov

7. *Polilova T.A.* Litsenzii dlia nauchnykh arkhivov otkrytogo dostupa // Preprinty IPM im. M.V.Keldysha. 2018. No 233. 20 s. URL: http://library.keldysh.ru/preprint.asp?id=2018-233

8. *Gorbunov-Posadov M.M.* Zhivaia publikatsiia. M.: IPM im.M.V.Keldysha, 2011. Redaktsiia ot 01.05.2019. URL: http://keldysh.ru/gorbunov/live.htm

9. Cait nauchnogo serialnogo izdaniia «Preprinty IPM im. M.V. Keldysha». URL: http://library.keldysh.ru/preprints/

10. Itogovyi dokument konferentsii «Problemy kachestva nauchnoi raboty i akademicheskii plagiat» (Moskva, 26 sentiabria 2018, RGGU) URL: http://www.sib-science.info/ru/conferences/itogovyy-dokument-09102018

11. Novye vozmozhnosti: izdanie sbornikov. URL: https:// naukapublishers.ru/ history/novosti/novosti-izdatelstva/229-novye-vozmozhnosti-izdanie-sbornikov

СВЕДЕНИЯ ОБ АВТОРЕ

ПОЛИЛОВА Татьяна Алексеевна – старший научный сотрудник Института прикладной математики им. М.В. Келдыша РАН, доктор физико-математических наук, лауреат Премии Президента РФ в области образования.

Tatyana Alekseevna POLILOVA – senior researcher of the Keldysh Institute of Applied Mathematics Russian Academy of Sciences.

email: polilova@keldysh.ru

Материал поступил в редакцию 26 июня 2019 года