УДК 004.9
DOI: 10.17277/voprosy.2018.02.pp.203-216

МЕТОДИКА КЛАССИФИКАЦИИ И ОБРАБОТКИ ДОКУМЕНТОВ В СИСТЕМЕ УПРАВЛЕНИЯ ЭЛЕКТРОННЫМ ДОКУМЕНТООБОРОТОМ НАУЧНО-ОБРАЗОВАТЕЛЬНОГО УЧРЕЖДЕНИЯ

М. Н. Краснянский, А. Д. Обухов

ФГБОУ ВО «Тамбовский государственный технический университет», г. Тамбов, Россия

Рецензент д-р пед. наук, профессор Н. В. Молоткова

Ключевые слова: классификация; машинное обучение; обработка документов; система электронного документооборота.

Аннотация: Рассмотрена методика классификации документов в системах управления электронным документооборотом на основе фасетного подхода. Приведены основные признаки классификации и типы документов в данной предметной области. Предложена методика предварительной обработки документов на основе поэтапной фильтрации текста с использованием приоритетных блоков, стоп-листов и лемматизации. Приведенная методика позволяет повысить эффективность и точность работы методов машинного обучения за счет снижения размерности исходной задачи обработки текста. Полученные результаты будут использоваться в дальнейших исследованиях применения методов машинного обучения для обработки, классификации и маршрутизации документов.

Введение

Постоянно растущие объемы информации и высокие требования к надежности ее хранения и скорости передачи привели к широкому распространению систем управления электронным документооборотом (**СУЭД**), позволяющих автоматизировать процессы движения документов в организации, перейти от бумажного документооборота к электронному, оптимизировать структуру информационных потоков организации в целом. Разработка масштабных, многофункциональных СУЭД является длительным и трудоемким процессом, требующим комплексного подхода к проектированию, формализации структуры и параметров информационной

Краснянский Михаил Николаевич – доктор технических наук, профессор, ректор ТамбГТУ; Обухов Артем Дмитриевич – кандидат технических наук, ассистент кафедры «Компьютерно-интегрированные системы в машиностроении», e-mail: obuhov.art@gmail.com, ТамбГТУ, г. Тамбов, Россия.

системы. Кроме того, необходимо учитывать факторы предметной области, что приводит к адаптации СУЭД под конкретные задачи. С другой стороны, требуется разработка универсальных проектных решений с использованием передовых информационных технологий. Решить данные задачи без использования методов системного анализа и математического моделирования невозможно [1].

Трендом последних лет является повсеместное использование методов искусственного интеллекта и машинного обучения для решения задач обработки больших объемов информации, их кластеризации и классификации, что позволяет обнаруживать новые закономерности в уже исследованных процессах и формировать на их основе новые гипотезы.

Вопросы обработки документов с использованием искусственного интеллекта рассматриваются с научно-практической точки зрения уже много лет [2 – 6]. Отметим, что все авторы признают важность этапа подготовки и предварительной обработки документов, так как данные процессы позволяют значительно повысить эффективность и точность работы алгоритмов машинного обучения без существенной переработки существующих методов обработки документов на основе машинного обучения.

В рамках данной статьи рассмотрим две важные задачи подготовки документов при использовании методов машинного обучения: классификацию и обработку документов. Первая задача направлена прежде всего на получение списка классов (категорий) документов, которые мы будем использовать при классификации на основе машинного обучения. Вторая задача позволит нам повысить точность и эффективность работы алгоритмов машинного обучения за счет повышения качества исходного текста. В данных задачах ключевое значение будет играть специфика рассматриваемой предметной области научно-образовательного учреждения.

Применение машинного обучения для обработки информации в научно-образовательном учреждении

Перед тем как перейти непосредственно к решению поставленных задач классификации и обработки информации, проведен анализ предметной области научно-образовательного учреждения, определена ее специфика, а также те задачи, в которых возможно и оправдано применение методов машинного обучения.

Разработка СУЭД для научно-образовательного учреждения связана с рядом трудностей: отдельные модули системы могут быть направлены на решение абсолютно разных задач административной, образовательной и научно-инновационной деятельности, обладать различной функциональностью.

Для решения данной проблемы предложена универсальная структура модулей СУЭД, учитывающая возможные функциональные требования отдельных структурных подразделений научно-образовательного учреждения и позволяющая реализовать общую структуру СУЭД на основе модульного подхода к проектированию данных систем и компоновки модулей из отдельных элементов в зависимости от поставленных перед ними задач.

Универсальную структуру модуля представим в виде множества компонентов типового модуля и для каждого из них опишем методы машинного обучения, которые возможно использовать для повышения эффективности функционирования. Описание элементарных компонентов и возможность применения в них методов машинного обучения приведена в табл. 1 [1].

Отметим, что каждый из компонентов предложенной структурной модели выбирается в конкретной ситуации в зависимости от требований предметной области и технического задания. При этом применение конкретных методов машинного обучения зависит от решаемых модулем задач и необходимости в обработке информации с применением данных технологий [7 – 9].

Таблица 1

Применение методов машинного обучения в компонентах СУЭД

Наименование и описание компонента модуля СУЭД	Применение машинного обучения в рамках компонента
1	2
1. Компоненты обработки документов. Реализуют основные функции по созданию, редактированию, перемещению и управлению документами через обращение к интерфейсным формам СУЭД либо к контекстному меню файлового менеджера. Также включает резервирование и восстановление данных	Пакетное редактирование документов с автоматизацией внесения изменений. Переформатирование документов. Автоматическое резервирование ключевых документов необходимых категорий
2. Компоненты чтения документов. Осуществляют доступ к документам для их просмотра, распознавания, классификации, проверки, подписи и выполнения ряда административных функций	Классификация документов. Распознавание текста и изображений. Проверка правильности заполнения форм
3. Компоненты поиска информации. Включают необходимые инструменты и средства реализации информационного поиска в СУЭД	Поиск документов. Обработка поисковых запросов, коррекция ошибок, лемматизация
4. Компоненты формирования документации. К ним относятся различные средства, инструменты и методы формирования электронных версий документов	Автоматическое формирование документов по заданным шаблонам
5. Административные компоненты. Используются для осуществления контроля работоспособности системы и управления отдельными ее компонентами с целью устранения неполадок. Также включают необходимые инструменты модернизации и модификации СУЭД	Структурный синтез компонентов СУЭД на основе данных о наборе решаемых задач и требуемых функций

Продолжение табл. 1

1	2
6. Компоненты сетевого взаимодействия. Включают встроенные и сторонние инструменты по организации электронных архивов и управлению документами на сервере, передачу данных на различных уровнях взаимодействия пользователей и СУЭД	Маршрутизация документов
7. Компоненты работы с базами данных. Используются для осуществления основных операций с данными: получения, передачи, обновления, анализа и поиска. Включают инструменты по организации работы с базами данных: библиотеки программного кода, графические оболочки, панели управления и т.д.	Обработка и генерация запросов к базам данных
8. Компоненты разграничения доступа к информации. Включают программные и интерфейсные средства по ограничению доступа пользователей к данным, файлам, функциям, разделам информационной системы и отдельным элементам форм	Классификация пользователей и определение их роли в системе на основе атрибутов учетной записи
9. Компоненты формирования статистики. Реализация формирования статистических документов встроенными средствами СУЭД либо сторонними расширениями	Анализ больших объемов данных, поиск закономерностей и зависимостей
10. Вычислительные компоненты. Процедуры, функции, написанные на языке программирования, сторонние модули и библиотеки, позволяющие производить математические расчеты и отображать их результаты	Интеллектуальный выбор алгоритмов и методов в зависимости от контекста задачи
11. Справочные компоненты. Содержат различные учебно-методические, справочные материалы, призванные ускорить процесс освоения СУЭД пользователями, разрешить возникающие при работе затруднения, используя инструменты поиска, тестирования, визуальные средства закрепления полученных знаний	Автоматизация поиска справочной информации, контекстные подсказки на основе методов искусственного интеллекта
12. Компоненты тестирования. Используемые на этапе разработки и внедрения новых функций компоненты, направленные на упрощение процесса отладки и тестирования СУЭД	Генерация тестовых пакетов для проверки работоспособности информационной системы

Таким образом, для разработки модуля СУЭД следует определиться с набором компонентов, при необходимости применить методы машинного обучения для решения конкретных задач в рамках компонентов. Есте-

ственно, что в процессе структурного синтеза не все компоненты будут использованы в рамках одного модуля, так как это сделает его функциональность избыточной, а разработку – нерационально сложной и долгой.

Анализ подходов к применению методов машинного обучения для обработки документов позволил выделить несколько эффективных и популярных решений на основе нейронных сетей и библиотек Python: TensorFlow, Pandas, Scikit-Learn, PDFMiner, Stemming Snowball и др. [10, 11].

Таким образом, проведенный анализ позволяет говорить о высокой актуальности и важности применения машинного обучения при реализации СУЭД. Однако, учитывая определенную специфику научно-образовательного учреждения, необходимо проанализировать и адаптировать решение задач классификации документов и их предварительной обработки.

Подходы к классификации документов научно-образовательного учреждения

Классификация документов – одна из задач информационного поиска, заключающаяся в определении отношения документа к одной из нескольких категорий на основании содержания документа или каких-то иных признаков или атрибутов. Долгое время классификация документов осуществлялась вручную либо на основе заданного набора правил, однако распространение методов машинного обучения в последние годы позволило решать данную задачу намного эффективнее [12, 13].

Отметим, что необходимо разделять задачи классификации и кластеризации документов. В первом случае разбиваем множество всех документов по заранее заданным категориям в соответствии с известными критериями, во втором, – подобная группировка заранее не определена и проводится на основе поступивших на вход алгоритма документов.

Перед тем как перейти непосредственно к анализу подходов к применению машинного обучения для классификации документов, рассмотрим теоретическую базу данного процесса, то есть сформулируем систему классификации документов, те признаки и критерии, по которым их можно сгруппировать.

Итак, под системой классификации будем понимать совокупность методов и правил классификации и ее результат, то есть систему классификационных группировок (классов). Система классификации может быть представлена в виде перечня признаков классификации (фасетов) и их значений и правил образования группировок путем комбинирования признаков-фасетов. Основными методами классификации являются иерархический и фасетный [14, 15].

Для последовательного разделения множества объектов на группировки используется иерархический метод классификации, позволяющий получить иерархическую древовидную структуру в виде ветвящегося графа, в котором объекты разделяются по уровням деления в соответствии со своими признаками.

Другим подходом является параллельное деление множества объектов с использованием фасетного метода классификации, при котором выделяется множество фасет-признаков, совокупность которых позволяет

образовывать отдельные системы классификации, определяемые соответствующими фасетными формулами. Для машинного обучения более применимым является фасетный метод, так как каждому типу документов можно сопоставить перечень признаков.

Методика классификации документов на основе фасетного метода заключается в следующем: каждому объекту документооборота $u_i \in U$ соответствует некоторое подмножество признаков классификации f_j, на основе которых осуществляется их группировка в массивы Cat_j:

$$\begin{aligned} & u_i \to \{f_j^i\}, \\ & Cat_j : f_j \to \{u_i \mid f_j \in \{f_j^i\}\}, \\ & f_j^i \in F, u_i \in U, \end{aligned} \qquad (1)$$

где f_j^i – признак f_j классификации, характеризующий документ u_i; u_i – элемент множества документов, каждому из которых ставится в соответствие множество признаков $\{f_j^i\}$; Cat_j – функция формирования классификации по признаку f_j, которая ставит в соответствие каждому признаку f_j набор документов $\{u_i\}$, обладающих данным признаком.

Используя представленную методику, можно разбить множество всех документов по набору признаков, что значительно упростит подготовку информации для последующей ее обработки в методах классификации.

Перечислим основные и дополнительные признаки документов, сформулированные на основе работы [16] (рис. 1):

1) наименование документа, определяющее его разновидность (тип): приказы, распоряжения, планы, отчеты, акты, протоколы, договоры, инструкции, справки, докладные, объяснительные записки, служебные письма и т.д.;

2) структурная принадлежность документа (определяется его отношением к отдельным структурным подразделениям);

3) способ хранения документов: рукописный, машинописный, печатный, электронный, графический. Для электронных версий документов также возможна группировка по формату документов;

4) степень сложности документации, зависящая от объема и структуры документа: простая (документ направлен на решение одного конкретного вопроса или задачи), сложная (включает множество задач различного уровня вложенности);

5) секретность: открытые (несекретные) и закрытые (с ограниченным доступом) документы. Вторая категория также может подразделяться по различной степени секретности (конфиденциальные, секретные и совершенно секретные, для служебного пользования);

6) юридическая сила документов, разделяющая их на подлинные и подложные. Первые, в свою очередь, делятся на действительные (имеющие юридическую силу) и недействительные (срок действия которых уже закончился);

7) срок исполнения: срочный и несрочный;

Рис. 1. Структурная схема признаков классификации документов

8) срок хранения: временный и постоянный;

9) происхождение документа: внутреннее (создан внутри организации) и внешнее (поступил извне);

10) обязательность исполнения: информационные (служат для изложения различных фактов и сведений по деятельности организации) и директивные (обладают юридической силой и обязательны для исполнения);

11) степень унификации: индивидуальные (обладающие своим уникальным стилем и форматированием), типовые (стандартные по стилю и оформлению), шаблоны, анкеты, таблицы;

12) характер происхождения: первичный (содержит исходные данные) и вторичный (получен на основе анализа, обработки или обобщения исходной информации).

Данные признаки могут быть расширены за счет дополнительных, специфических особенностей предметной области организации, в которой осуществляется классификация документов. Так как в рамках статьи рассматривается документооборот научно-образовательного учреждения, то приведем несколько примеров специфических признаков из данной предметной области, которые расширяют множество вышеперечисленных основных признаков [17]:

13) наименование деятельности организации, в рамках которой существует документ: административная, научно-инновационная, образовательная;

14) состояние документа, определяющее степень его готовности: начальное, временное, текущее, финальное;

15) архитектура документа: простая (один документ), составная (включает в себя несколько взаимосвязанных документов);

16) актуальность документа: актуальный (соответствует текущим стандартам Министерства образования и науки РФ), устаревший;

17) категория исполнителей документа: административные работники, профессорско-преподавательский состав, научные сотрудники, обучающиеся.

Перечисленные основные и специфические признаки позволяют осуществить классификацию документов в научно-образовательной организации, причем как аналитически, так и с использованием современных методов машинного обучения. Таким образом, переходим от первой задачи (построения системы классификации) ко второй – подготовительной обработке документов.

Методика обработки документов научно-образовательного учреждения

Машинное обучение является эффективным инструментом при реализации алгоритмов классификации, маршрутизации, обработки и поиска документов, однако определяющее значение в этих процессах имеет качество исходных данных. Именно поэтому проведение подготовки исходных документов, их предварительная обработка позволяет значительно повысить точность результатов, получаемых в ходе применения машинного обучения.

Учитывая специфику предметной области и вышеперечисленные признаки классификации документов, сформулирована методика предварительной обработки документов научно-образовательного учреждения, направленная на повышение точности работы методов машинного обучения. Представим ее в виде алгоритма (рис. 2) и опишем основные этапы [18].

На вход алгоритма поступает исходный документ. Первым этапом является определение его формата, то есть расширение файла (txt, doc, pdf и т.д.). На основе этой информации осуществляется выбор библиотеки программного кода, с помощью которой можно корректно прочитать и проанализировать документ.

Используя выбранную библиотеку, осуществляется извлечение данных из документа в виде неформатированного текста. Так как размеры документов в научно-образовательном учреждении сильно варьируются от небольших (например, служебные записки) до огромных (дипломные

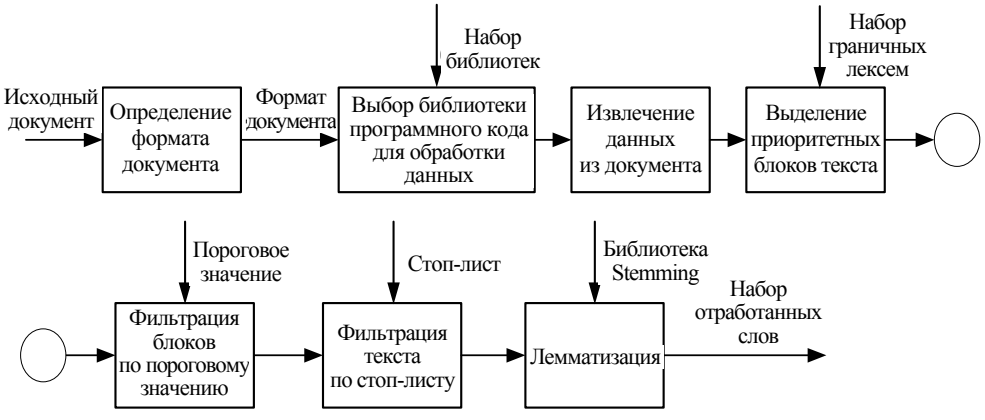


Рис. 2. Алгоритм представительной обработки документов

работы, пояснительные записки и научно-технические отчеты), то необходимо провести процедуру выделения приоритетных блоков. Исходный текст документа разбивается на несколько частей в соответствии с набором граничных лексем (списка определенных слов, позволяющих выделить границы блоков), после чего каждая часть ранжируется в зависимости от положения в документе. Отметим, что в научно-образовательном учреждении наибольший приоритет имеют начальные блоки, так как они содержат информацию с титульных листов, наименование документов и списки отправителей, исполнителей и получателей. Также приоритет отдается последним блокам, так как они могут содержать подписи и список ответственных за исполнение документа.

Дальнейшим шагом алгоритма является фильтрация блоков по заданному пороговому значению, которое может быть определено экспериментально на основе анализа объема документов или содержания первого блока. Все блоки, значение приоритета которых ниже порогового значения, исключаются из дальнейшего анализа.

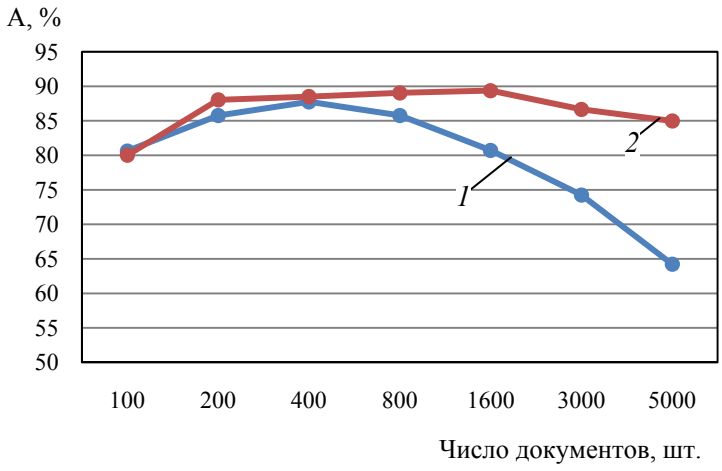
Следующий этап – фильтрация слов текста по стоп-листу. Стоп-лист представляет собой набор коротких слов, предлогов, частиц, знаков препинания, символов, чисел и т.д., не несущих смысловой нагрузки для последующего анализа. Их удаление из исходного текста позволяет повысить его смысловую ценность, что особенно важно для задачи классификации, а также эффективность работы обучающих алгоритмов, так как число слов напрямую влияет на размерность задачи обучения и, следовательно, время ее решения.

Заключительным этапом алгоритма является лемматизация, то есть процесс приведения слов к леммам, их нормальным словесным формам. Для реализации процесса лемматизации можно использовать библиотеку программного кода Python Stemming Snowball, которая позволяет привести все русские и английские слова к нормальной форме, удалив окончания и необязательные суффиксы. Таким образом, значительно снижается размерность задачи, исключая множество различных вариаций одного и того же слова.

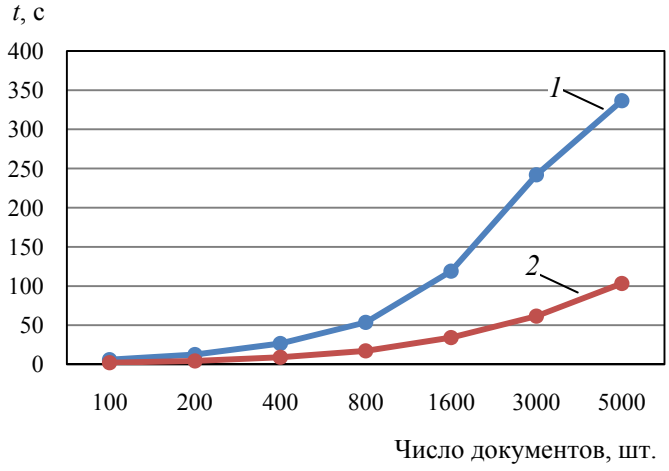
Полученный после выполнения лемматизации набор слов уже может использоваться для проведения машинного обучения и решения конкретных задач: классификации, маршрутизации, поиска и т.д.

Представленная методика позволяет значительно ускорить выполнение процедуры машинного обучения и повысить ее точность для осуществления классификации документов научно-образовательного учреждения. Эффективность работы классификаторов до применения разработанной методики и после представлены на рис. 3. В качестве выбранной технологии машинного обучения использовался наивный байесовский классификатор (naive Bayes Classifier) [19].

Из полученного графика очевидно, что в среднем точность классификации повысилась с 80 до 87 %, а время работы классификатора снизилось в 3,2 раза за счет выделения приоритетных блоков документов и сокращения общего объема анализируемого текста.



а)



б)

Рис. 3. Влияние предварительной обработки данных на точность *A* (*а*) и время *t* (*б*) работы классификатора документов до *1* и после *2* применения методики

Далее проведено исследование эффективности представленной методики при классификации документов с помощью библиотеки Tensor Flow (нейронная сеть) [20]. Несмотря на то что такого же значительного сокращения времени обучения не произошло (показатель улучшился в диапазоне 1,33 – 1,5 раза), точность классификации возросла в среднем с 73 до 89 %.

Полученные результаты, прежде всего, обусловлены характером классифицируемых документов: они могут значительно отличаться как по структуре, так и размеру, причем данные признаки варьируются в очень широких пределах, например, на вход подавались документы объемом в несколько сотен страниц. Естественно, для классификации подобного набора документов необходимы дополнительные операции по предварительной обработке.

Таким образом, представленная методика в рамках классификации документов предметной области научно-образовательного учреждения доказывает свою применимость и эффективность.

Заключение

В рамках данной статьи рассмотрены несколько важных вопросов в области применения машинного обучения для обработки информации в СУЭД научно-образовательных учреждений.

Проведен анализ структуры научно-образовательного учреждения, рассмотрены универсальная структура типового модуля СУЭД и возможность применения методов машинного обучения в компонентах модулей. Определив, таким образом, ряд задач по применению машинного обучения в СУЭД, рассмотрены конкретные подходы к классификации документов и их предварительной обработке.

Классификация документов научно-образовательного учреждения проводилась на основе фасетного подхода. Приведены основные признаки классификации и типы документов в данной предметной области, что позволяет сформулировать набор категорий, по которым в дальнейшем будет осуществляться классификация документов.

Рассмотрена методика предварительной обработки документов для их последующего анализа при использовании методов машинного обучения. Используя такие подходы, как ранжирование приоритетных блоков текста, фильтрация по стоп-листу, лемматизация, достигнут конкретный результат в виде повышения качества классификации и ускорения процесса обучения, что подтверждает применимость изложенной методики для предварительной обработки информации.

Полученные результаты будут использоваться в дальнейших исследованиях применения методов машинного обучения для обработки, классификации и маршрутизации документов.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ в рамках гранта Президента РФ МК-1666.2018.9.

Список литературы

1. Проектирование информационных систем управления документооборотом научно-образовательных учреждений : монография / М. Н. Краснянский [и др.]. – Тамбов : Изд-во ФГБОУ ВПО «ТГТУ», 2015. – 216 с.

2. Mathematical Language Processing: Automatic Grading and Feedback for Open Response Mathematical Questions / A. S. Lan [et al.] // Proceedings of the Second (2015) ACM Conference on Learning Scale. – ACM, 2015. – P. 167 – 176.

3. Piernik, M. Clustering XML Documents by Patterns / M. Piernik, D. Brzezinski, T. Morzy // Knowledge and Information Systems. – 2016. – Vol. 46, No. 1. – P. 185 – 212.

4. A Novel Contextual Topic Model for Multi-Document Summarization / G. Yang [et al.] // Expert Systems with Applications. – 2015. – Vol. 42, No. 3. – P. 1340 – 1352.

5. Canhasi, E. Multi-Document Summarization Via Archetypal Analysis of the Content-Graph Joint Model / E. Canhasi, I. Kononenko // Knowledge and Information Systems. – 2014. – Vol. 41, No. 3. – P. 821 – 842.

6. Yang, W. A Discriminative Topic Model Using Document Network Structure / W. Yang, J. Boyd-Graber, P. Resnik // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). – 2016. – Vol. 1. – P. 686 – 696.

7. Sebastiani, F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys (CSUR). – 2002. – Vol. 34, No. 1. – P. 1 – 47.

8. Cohen, W. W. Context-Sensitive Learning Methods for Text Categorization / W. W. Cohen, Y. Singer // ACM Transactions on Information Systems (TOIS). – 1999. – Vol. 17, No. 2. – P. 141 – 173.

9. Radovanović, M. Document Representations for Classification of Short Web-Page Descriptions / M. Radovanović, M. Ivanović // International Conference on Data Warehousing and Knowledge Discovery. – Springer, Berlin, Heidelberg, 2006. – P. 544 – 553.

10. Stamatatos, E. Automatic Text Categorization in Terms of Genre and Author / E. Stamatatos, N. Fakotakis, G. Kokkinakis // Computational linguistics. – 2000. – Vol. 26, No. 4. – P. 471 – 495.

11. Агеев, М. С. Автоматическая рубрикация текстов: методы и проблемы / М. С. Агеев, Б. В. Добров, Н. В. Лукашевич // Ученые записки Казанского университета. Сер.: Физико-математические науки. – 2008. – Т. 150, № 4. – С. 25 – 40.

12. Найденова, К. А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований / К. А. Найденова, О. А. Невзорова // Ученые записки Казанского университета. Сер.: Физико-математические науки. – 2008. – Т. 150, №. 4. – С. 5 – 24.

13. Мбайкоджи Э. Метод автоматической классификации коротких текстовых сообщений / Э. Мбайкоджи, А. А. Драль, И. В. Соченков // Информационные технологии и вычислительные системы. – 2012. – № 3. – С. 93 – 102.

14. Гулин, В. В. Сравнительный анализ методов классификации текстовых документов / В. В. Гулин // Вестник МЭИ. – 2011. – №. 6. – С. 100 – 108.

15. Tuia, D. Multiclass Feature Learning for Hyperspectral Image Classification: Sparse and Hierarchical Solutions / D. Tuia, R. Flamary, N. Courty // ISPRS Journal of Photogrammetry and Remote Sensing. – 2015. – Vol. 105, July 2015. – P. 272 – 285. doi: 10.1016/j.isprsjprs.2015.01.006

16. Samanta, S. Space-Time Facet Model for Human Activity Classification / S. Samanta, B. Chanda // IEEE Transactions on Multimedia. – 2014. – Vol. 16, No. 6. – P. 1525 – 1535.

17. Элова, Г. В. Основы документооборота в таможенных органах : учеб. пособие / Г. В. Элова . – СПб. : ИЦ Интермедия, 2014. – 206 с.

18. Обухов, А. Д. Алгоритм структурно-параметрического синтеза системы электронного документооборота научно-образовательного учреждения / А. Д. Обухов // Вопросы современной науки и практики. Ун-т им. В.И. Вернадского. – 2016. – №. 1(59). – С. 199 – 209.

19. Peng, F. Combining Naive Bayes and n-Gram Language Models for Text Classification / F. Peng, D.Schuurmans // European Conference on Information Retrieval. – Springer, Berlin, Heidelberg, 2003. – P. 335 – 350.

20. TensorFlow: A System for Large-Scale Machine Learning / M. Abadi [et al.] // OSDI. – 2016. – Vol. 16. – P. 265 – 283.

References

1. Krasnyanskij M.N., Karpushkin S.V., Ostrouh A.V., Obuhov A.D., Kasatonov I.S., Bukreev D.V., Karpov S.V., Dedov D.L. *Proektirovanie informacionnyh sistem upravleniya dokumentooborotom nauchno-obrazovatel'nyh uchrezhdenij*

[Designing information management systems for document management of scientific and educational institutions], Tambov: Izdatel'stvo FGBOU VPO «TGTU», 2015, 216 p. (In Russ.)

2. Lan A.S., Vats D., Waters A.E., Baraniuk R.G. [Mathematical language processing: Automatic grading and feedback for open response mathematical questions], *Proceedings of the Second ACM Conference on Learning Scale*, ACM, 2015, pp. 167-176.

3. Piernik M., Brzezinski D., Morzy T. [Clustering XML documents by patterns], *Knowledge and Information Systems*, 2016, vol. 46, no. 1, pp. 185-212.

4. Yang G., Wen D., Chen N.S., Sutinen E. [A novel contextual topic model for multi-document summarization], *Expert Systems with Applications*, 2015, vol. 42, no. 3, pp. 1340-1352.

5. Canhasi E., Kononenko I. [Multi-document summarization via archetypal analysis of the content-graph joint model], *Knowledge and information systems*, 2014, vol. 41, no. 3, pp. 821-842.

6. Yang W., Boyd-Graber J., Resnik P. [A discriminative topic model using document network structure], *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers), 2016, vol. 1, pp. 686-696.

7. Sebastiani F. [Machine learning in automated text categorization], *ACM computing surveys (CSUR)*, 2002, vol. 34, no. 1, pp. 1-47.

8. Cohen W.W., Singer Y. [Context-sensitive learning methods for text categorization], *ACM Transactions on Information Systems (TOIS)*, 1999, vol. 17, no. 2, pp. 141-173.

9. Radovanović M., Ivanović M. [Document representations for classification of short web-page descriptions], *International Conference on Data Warehousing and Knowledge Discovery*, Springer, Berlin, Heidelberg, 2006, pp. 544-553.

10. Stamatatos E., Fakotakis N., Kokkinakis G. [Automatic text categorization in terms of genre and author], *Computational linguistics*, 2000, vol. 26, no. 4, pp. 471-495.

11. Ageev M.S., Dobrov B.V., Lukashevich N.V. [Automatic text classification: methods and problems], *Uchenye zapiski Kazanskogo universiteta. Ser.: Fiziko-matematicheskie nauki* [Scientific notes of Kazan University. Ser.: Physical and mathematical sciences], 2008, vol. 150, no. 4, pp. 25-40. (In Russ.)

12. Najdenova K.A., Nevzorova O.A. [Machine learning in the tasks of natural language processing: a review of the current state of research], *Uchenye zapiski Kazanskogo universiteta. Ser.: Fiziko-matematicheskie nauki* [Scientific notes of Kazan University. Ser.: Physical and mathematical sciences], 2008, vol. 150, no. 4, pp. 5-24. (In Russ.)

13. Mbajkodzhi E.H., Dral' A.A., Sochenkov I.V. [Method of automatic classification of short text messages], *Informacionnye tekhnologii i vychislitel'nye sistemy* [Information Technologies and Numerical Systems], 2012, no. 3, pp. 93-102. (In Russ.)

14. Gulin V.V. [Comparative analysis of classification methods for text documents], *Vestnik MEHI* [Bulletin of MPEI], 2011, no. 6, pp. 100-108. (In Russ.)

15. Tuia D., Flamary R., Courty N. [Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions], *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015, vol. 105, pp. 272-285. doi: 10.1016/j.isprsjprs.2015.01.006

16. Samanta S., Chanda B. [Space-time facet model for human activity classification], *IEEE Transactions on Multimedia*, 2014, vol. 16, no. 6, pp. 1525-1535.

17. Elova G.V. *Osnovy dokumentooborota v tamozhennyh organah* [Basics of document circulation in customs bodies], St. Petersburg: IC Intermediya, 2014, 206 p. (In Russ.)

18. Obuhov A.D. [Algorithm of structural-parametric synthesis of the electronic document management system of the scientific-educational institution], *Voprosy sovremennoj nauki i praktiki. Universitet im. V.I. Vernadskogo* [Problems of Contempo-

rary Science and Practice. Vernadsky University], 2016, no. 1(59), pp. 199-209. (In Russ., abstract in Eng.)

19. Peng F., Schuurmans D. Combining naive Bayes and n-gram language models for text classification, *European Conference on Information Retrieval*, Springer, Berlin, Heidelberg, 2003, pp. 335-350.

20. Abadi M., Barham P., Chen J., Chen Zh. [et al.]. [TensorFlow: A System for Large-Scale Machine Learning], *OSDI*, 2016, vol. 16, pp. 265-283.

The Method of Classification and Management of Documents in Electronic Document Management System of Academic Institutions

M. N. Krasnyansky, A. D. Obukhov

Tambov State Technical University, Tambov, Russia

Keywords: electronic document management system; document management; classification; machine learning.

Abstract. This article discusses the approach to the classification of documents in electronic document management systems based on the facet approach. The analysis of the structure of academic institutions, considered the universal structure of the model module and the possibility of using machine learning methods in the components of modules. Having defined, thus, a number of tasks on application of machine learning in EDMS, further concrete approaches to classification of documents and their preliminary processing are considered.

Classification of documents of scientific and educational institution was carried out on the basis of the facet approach. The main features of the classification and types of documents in this subject area are given, which allows formulating a set of categories for which the classification of documents will be carried out in the future.

The method of preliminary processing of documents for their subsequent analysis by using machine learning methods is proposed. The application of such approaches as ranking of priority blocks of text, filtering by stoplist, lemmatization, achieved a concrete result in the form of improving the quality of classification and acceleration of the learning process confirms the applicability of the above methodology for preprocessing of information.

The main features of classification and types of documents in this subject area are considered. The results obtained will be used in further studies of the application of machine learning methods for processing, classification and routing of documents.
