Палий Александр Викторович – Южный федеральный университет; e-mail: a.v._paliy@mail.ru; 347928, Таганрог, пер. Некрасовский, 44; тел.: 88634371603; кафедра конструирования электронных средств; к.т.н.; доцент.

Чернов Николай Николаевич – e-mail: nik-chernov@yandex.ru; тел.: 88634371795; кафедра электрогидроакустической и медицинской техники; д.т.н.; профессор.

Прибыльский Алексей Васильевич – Государственное бюджетное профессиональное образовательное учреждение Ростовской области «Таганрогский колледж морского приборостроения»; e-mail: pribylsku.al@mail.ru; 347910, г. Таганрог, ул. Ленина 222 В, кв. 7; тел.: 89094328815; преподаватель специальных дисциплин.

Paliy Alexander Viktorovich – Southern Federal University; e-mail: a.v._paliy@mail.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371603; the department of electronic apparatuses design; cand. of eng. sc.

Chernov Nikolay Nikolaevich – e-mail: nik-chernov@yandex.ru; phone: +78634371795; the department of acoustics and medical technology; dr. of eng. sc.; professor.

Prybylski Alexey Vasilevich – The State budget professional educational institution of the Rostov Region "Taganrog College of marine engineering"; e-mail: pribylsku.al@mail.ru; 347910, Taganrog, Lenina street 222, AP. 7; phone: +7909432-88-15; teacher of special subjects.

УДК 004.041 DOI 10.23683/2311-3103-2018-2-163-173

А.В. Семенова, В.М. Курейчик

АНСАМБЛЬ КЛАССИФИКАТОРОВ ДЛЯ АВТОМАТИЧЕСКОГО ПОПОЛНЕНИЯ ОНТОЛОГИЙ*

Искусственный интеллект в настоящее время является одной из перспективных областей научного и практического знания. В искусственном интеллекте онтологии используются для формальной спецификации знаний. В статье предложен подход к автоматизации процесса пополнения онтологии по коллекции текстовых документов, относящихся к одной тематике. Ключевой целью работы является разработка ансамбля классификаторов для задачи автоматического пополнения онтологии предметной области. Основной задачей создания ансамбля является повышение точности прогноза агрегированного классификатора по сравнению с точностью прогнозирования каждого индивидуального базового классификатора. Для достижения поставленной цели предложен новый вариант ансамбля классификаторов, основанный на методе опорных векторов (SVM-классификатор), нейронной сети (LSTM-классификатор) и методах дистрибутивной семантики (Fasttext, word embedding), и отличающийся от известных подходов способом представления решения и возможностью формирования коллективов классификаторов. В процессе оптимизации происходит определение параметров, как отдельных классификаторов, так и всего ансамбля. Разработка ансамбля классификаторов выполнена среде Matlab с применением пакета Text Analytics Toolbox. Ансамбль классификаторов построен на наборе данных для машинного обучения Reuters-21578 (выборка новостных статей). Для обучения моделей дистрибутивной семантики выбрана обученная на Wikipedia 2014 коллекция GloVe векторов для английского языка. Сравнительное тестирование показало преимущества использования предложенного ансамбля классификаторов при работе с многомерными данными, характеризующимися большим количеством признаков. Предложенный ансамбль классификаторов может применяться для определения тематики документа, для извлечения терминов из текстовых документов и построения тезауруса. Отличительными особенностями разработанного ансамбля классификаторов являются: мягкие требования к

* Работа выполнена за счет частичного финансирования ГЗ №2.5537.2017/6.7 в ЮФУ. Грант РФФИ № 18-07-00050.

исходным данным; автоматическое выделение терминов области знания; возможность использования алгоритма для построения онтологий разных областей научного знания без его модификации; высокое качество классификации данных при приемлемых временных затратах.

Классификация; ансамбль; онтологии; термины; база знаний; предметная область; признаки; корпус текстов; нейронная сеть.

A.V. Semenova, V.M. Kureichik

ENSEMBLE OF CLASSIFIERS FOR ONTOLOGY ENRICHMENT

Artificial intelligence is currently one of the promising areas of scientific and practical knowledge. In artificial intelligence, ontologies are used for the formal specification of knowledge. The article proposes an approach to automating the ontology replenishment process in text corpus related to the same domain. The key purpose of the paper is development of ensemble of classifiers for the task of domain ontology population. The main task of creating of the ensemble is to increase precision of the forecast of the aggregated classifier in comparison with the precision of individual baseline classifier. To achieve this goal, a new version of the ensemble of classifiers based on the method of support vector machine (SVM-classifier), neural network (LSTM-classifier) and methods of distributional semantics (Fasttext, word embedding) is developed. The principal difference of the ensemble from the known approaches is the method of representing the solution and the possibility of forming groups of classifiers. In the process of optimization, parameters are determined, both for individual classifiers and for the entire ensemble. The development of the ensemble of classifiers was performed in Matlab using the Text Analytics Toolbox. The ensemble of classifiers is built on a set of data for machine learning Reuters-21578 (news articles). To learn the models of distributional semantics, the GloVe vector collection for the English language trained in Wikipedia 2014 was selected. Comparative testing showed the advantages of using the proposed ensemble of classifiers when working with multidimensional data, characterized by a large number of features. The proposed ensemble of classifiers can be used to define the topic of a document, to extract terms from text documents and construct a thesaurus. Distinctive features of the developed ensemble of classifiers are: soft requirements to the initial data; automatic selection of the terms of the field of knowledge; the possibility of using an algorithm to construct ontologies of different areas of scientific knowledge without modifying it; high quality of data classification at an acceptable time.

Classification; ensemble; ontology; terms; knowledge base; domain; features; text corpus; neural network.

Введение. Искусственный интеллект в настоящее время является одной из перспективных областей научного и практического знания. В искусственном интеллекте онтологии используются для формальной спецификации знаний. Разработка и внедрение интеллектуальных экспертных систем является важным направлением повышения надёжности и эффективности технической эксплуатации промышленных объектов. Одним из современных подходов совершенствования экспертных систем (ЭС) является использование онтологий. Автоматическое построение онтологической модели данных и ее последующее пополнение на основе анализа корпуса научных текстов для определенной предметной области позволит в автоматическом режиме извлекать знания о терминах и отношениях между ними из научных текстов, что повысит эффективность построения онтологий [1].

Стоит отметить, что качество формируемых онтологий предметных областей во многом определяется полнотой учета в онтологической модели наиболее значимых концептов для корпуса анализируемых текстов с учетом их тематической специфики [2]. В связи с этим целесообразно решить задачу формирования множества концептов будущей онтологии. При поиске слов и словосочетаний, которые могут применяться в качестве концептов, сформированное множество концептов претендентов не всегда является оптимальным. Обобщив результаты, полученные в работе [3], были выявлены следующие проблемы, возникающие на этапе объединения онтологических моделей: не всегда удается правильно найти связи между концептами; не всегда удается выделить концепты, имеющие связь с наибольшим количеством

других концептов; найденные связи между концептами будущей онтологии не всегда актуальны для конкретной предметной области, имеет место увеличение пространства признаков. При этом не только повышается используемый объем памяти и увеличивается время на создание онтологии и обработку запросов к ней, но и избыточным становится объем онтологии, что снижает оперативность дальнейшего ее применения. Целесообразно рассмотреть возможность устранения перечисленных трудностей путем снижения размерности пространства признаков за счет классификации текстовой информации.

В последнее время значительно возрос интерес к вопросу увеличения точности моделей, основанных на алгоритмах машинного обучения, посредством объединения возможностей нескольких классификаторов и создания ансамблей классификаторов, что в итоге позволит повысить качество решения прикладных задач.

Учитывая сказанное, задача построения ансамбля классификаторов, который может являться ансамблем из нескольких известных алгоритмов, с различными весами и параметрами, и работающих со слабоструктурированными данными, является актуальной.

Анализ и современное состояние исследований по проблеме. В настоящее время вопросы построения эффективных средств автоматической классификации текстов достаточно активно рассматриваются как в отечественных, так и в зарубежных работах [4].

В большинстве обзоров, посвященных методам классификации текстов, документы представляются как векторы, в виде мешка слов (bag of words, bow). Так, Andrews и Fox в работе 2007 года [5] описывают способы представления набора документов в виде векторной модели, в том числе различные способы предобработки текстов, а также алгоритмы их кластеризации, такие как модификации k-means (или метод k-средних), EM-алгоритм и спектральная кластеризация. Так как одним из главных недостатков представления документов в виде мешка слов является высокая размерность и разреженность получаемых векторов, авторы также представляют методы понижения размерности векторного пространства.

Представление документов в данном пространстве TF-IDF заключается в формировании вектора частот слов, умноженного на вектор обратной подокументной частоты слов (как правило, lg(N/df), где N – общее количество документов в коллекции, а df – количество документов, в котором встречалось данное слово). Параметр IDF (обратная частота встречаемости в корпусе) указывает на общую встречаемость слова во всем корпусе. Следовательно, появление в документе часто встречаемых слов нормализуется. Высокие веса будут у термов, которые встречаются в документе часто, но редко во всей коллекции. Авторы работы [6] продемонстрировали другой вариант взвешивания значимости слов – метод BM25, который позволил получить более высокие результаты кластеризации текстов. В предложенном методе ограничивается значимость частоты n-граммы, а также она не только нормализуется по его размеру, но и ограничивается сверху, что позволяет избежать присваивания слову слишком большого веса [7].

Рассматривая разделительные (partitional) алгоритмы кластеризации документов, в частности k-means, более детально, Huang представляет описание и сравнение мер близости между bow-векторами [8]. В статье описаны шесть различных мер близости, между которыми проведено экспериментальное сравнение на алгоритме k-means; лучшие результаты по метрикам чистоты (purity) и энтропии показала кластеризация, использующая в качестве меры близости коэффициент Жаккара (Jaccard coeficient) и коэффициент корреляции Пирсона (Pearson correlation coefficient). Sathiyakumari и др. [9] также рассматривают кластеризацию документов только применительно к их представлению в виде мешка слов. Они выделяют четыре группы методов кластеризации таких представлений: разделительная кластеризация,

иерархическая кластеризация, k-средних и EM-алгоритм, хотя во многих других работах k-means включается в группу разделительных алгоритмов. Как видно в вышеупомянутых работах, классификация документов обычно сводится к классификации их векторных представлений в виде мешка слов.

Кластеризации векторов в общем случае, безотносительно к текстовым документам, также посвящено множество работ. Более широкий спектр возможных векторных представлений документа разбирается в одной из глав книги Mining Text Data [10]. В частности, в ней описываются методы, использующие в качестве признаков документов часто встречающиеся наборы слов, а также методы тематического моделирования. Кроме того, обозреваются подходы к онлайн-кластеризации текстов, использованию графовых методов кластеризации (в случае если между текстами существуют связи) и имеющейся заранее информации для кластеризации на основе алгоритмов частичного обучения (semi-supervised).

В некоторых обзорах авторы отдельно выделяют методы семантической классификации. Marchionini и др. [11] считают определяющим отличием семантической кластеризации от традиционной, основанной на мешке слов, использование семантических отношений между словами документов. Авторы относят к методам семантической кластеризации несколько групп алгоритмов: алгоритмы, основанные на онтологиях, таких как WordNet; алгоритмы, использующие в качестве признаков документа наборы связанных по смыслу слов; а также алгоритмы, основанные на графах концептов или именованных сущностей с семантическими отношениями между ними.

Таким образом, к настоящему времени произведено множество обзоров и экспериментальных сравнений методов классификации, но в большей их части не рассматриваются современные методы, например векторные представления слов, полученные с помощью нейронных сетей (word embedding), а также не учитывается специфика научных статей, в частности, тот факт, что во многих практических приложениях необходимо классифицировать статьи, принадлежащие одной предметной области, по более узким направлениям, причем полные тексты статей не всегда доступны. Одним из подходов к решению задачи классификации является усиление простых классификаторов путём комбинирования примитивных слабых классификаторов в один сильный. Под силой классификаторов подразумевается эффективность (качество) решения задачи классификации [12].

Описание используемых в работе моделей классификации. Решение задачи классификации состоит из множества последовательных этапов. Общая схема процесса мультиклассовой классификации представлена на рис. 1.

Рис. 1. Этапы процесса мультиклассовой классификации

Современные методы машинного обучения ориентированы на признаковое описание объектов. Поэтому текстовые документы переводят в векторные представления.

В **модели bag-of-words** каждый документ представляется в виде неупорядоченного набора термов [12]:

$$d_i = \{w_{1i}, w_{2i}, \dots, w_{mi}\},$$

где w_{ji} – j-ый терм (слово) в i-м документе и m – общее количество различных термов во всех документах коллекции (размер словаря).

Представление TF-IDF (Term Frequency – Inverse Document Frequency) – представление документов в данном пространстве заключается в формировании вектора частот слов, умноженного на вектор обратной подокументной частоты слов (как правило, lg(N/df), где N – общее количество документов в коллекции, а df – количество документов, в котором встречалось данное слово). Параметр IDF (обратная частота встречаемости в корпусе) указывает на общую встречаемость слова во всем корпусе. Значение признаков для n-граммы t_i в документе d_j в этом методе рассчитывается по следующей формуле [12]

$$idf(t_i) \cdot \frac{tf(t_i, d_j) \cdot (k_1 + 1)}{k_1 \cdot \left(1 + b + b \cdot \frac{|d_j|}{|d_{avg}|}\right) + tf(t_i, d_j)}$$

где $|d_j|$ – длина данного документа; $|d_{avg}|$ – средняя длина документов в наборе; k_1 и b – свободные параметры.

В методах «Word Embeddings» вектор документа представляет собой усредненную сумму векторных представлений каждого слова. Наиболее популярные методы данного класса – word2vec [13], GloVe [14] и fastText [15]. Первые два метода раскладывают матрицу терм-терм на две матрицы основную и контекстную. Word2Vec использует для этого минимизацию логарифма правдоподобия, взятого с отрицательным знаком (минимизация кросс-энтропии) (Word2Vec можно представить также трёхслойной нейронной сетью, состоящей из линейного скрытого слоя и softmax выходного слоя или в случае модели скипграмм – сигмоидной функцией активации), а GloVe – минимизацию взвешенной суммы квадратов ошибок (взвешенная линейная регрессия).

FastText. В отличие от первых двух методов FastText использует дополнительную информацию о морфологии слова при построении вектора слова, представляя его (слово) виде суммы буквенных н-грамм. Преимущество данного варианта по сравнению с предыдущим в том, что вектора не зависят от количества документов и имеют произвольную размерность. Недостаток – отсутствие интерпретируемости координат полученных векторов. Интерпретируемость данного варианта возможна только относительно меры сходства между векторами (чаще всего косинусная мера сходства).

Появление **word2vec** – семейства методов построения векторных представлений слов в пространстве низкой размерности (word embeddings) – позволило свести задачу оценки семантической близости слов к вычислению косинуса угла между векторами этих слов. Векторы, построенные на крупном неразмеченном корпусе текстов, дают лучшие результаты, чем классические методы на основе расстояний между словами в семантических сетях.

Тематическое моделирование. Тематическое моделирование – способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему [16]. Переход из пространства терминов в

пространство найденных тематик помогает разрешать синонимию и полисемию терминов, а также эффективнее решать такие задачи, как тематический поиск, классификация, суммаризация и аннотация коллекций документов и новостных потоков. Преимущество подхода заключается в том, что полученные векторы получаются разреженными, что, очевидно, позволит определить какие темы доминируют в документе и, как следствие, к какому классу относится документ.

Метод машин опорных векторов (SVM) основан на построении в векторном пространстве объектов-документов разделяющей гиперплоскости. По какую сторону классифицируемый объект расположен относительно этой гиперплоскости, к тому классу он и принадлежит. В основе линейных методов (построения гиперплоскости) лежит оптимизация функционала специального вида (метод минимизации эмпирического риска) [17].

Проанализировав современные методы мультиклассовой классификации для задачи индексации документов, методов построения и обучения классификатора с помощью машинного обучения, выполним оценку эффективности методов мультиклассовой классификации документов, проанализировав результаты соревнований мультиклассовой классификации документов.

Методы, извлечения признаков из онтологий. Все методы выбирают глубину разбиения онтологии на категории. Выбор глубины разбиения выполняется на основе того, чтобы в каждой категории было много данных, все категории были приблизительно сбалансированы на всех используемых при классификации языках. В работе для извлечения признаков из онтологий предлагается использовать вероятностные тематические модели. Вероятностные тематические модели поддерживают подключение онтологий [18]. Имеется возможность получить иерархическую информацию из онтологий и обучиться на коллекциях данных из различных источников (например, Википедии). Но у вероятностных тематических моделей есть недостаток: для эффективного их использования необходима предварительная обработка слов с целью повысить совстречаемость различных комбинаций слово-документ, прежде чем делать стохастическое матричное разложение. Этот недостаток легко преодолевается, если использовать векторные представления слов («Word Embedding») и по полученным векторам проводить контекстную кластеризацию. Контекстная кластеризация работает следующим образом. Вначале проводится кластеризация слов с помощью сферического k-means (работает на косинусном расстоянии), затем, выбирается вектор центроида каждого кластера; слово описывается как вектор близости к каждому кластеру. Затем каждый элемент вектора близости приводится к вероятностному значению с помощью нормированного гауссового ядра.

Способы объединения информации от признаков из онтологий и обучающей коллекции: линейный классификатор.

Реализация и экспериментальная оценка ансамбля классификации. Ансамбль состоит из разных модулей, каждый из которых реализует свой тип классификации и дополнительные признаки. Выбор данного подхода основывается на проведенном выше анализе и исходит из соображений, что не удается однозначно определить модель, превосходящую все остальные известные методы классификации, поэтому необходим ансамбль, который будет использовать преимущество каждой из описанных в документе моделей. В качестве априорной информации для классификации предлагается использовать онтологии (для извлечения дополнительных признаков для лучшего разделения классов).

Использование классификаторов позволяет повысить точность классификации. Построение k классификаторов осуществляется независимо друг от друга на обучающих множествах, полученных из исходного случайной заменой документов (размер обучающего множества остается прежним, просто одни документы отсутствуют, а другие встречаются несколько раз).

Идея подхода заключается в последовательном построении k классификаторов и объединении их результатов классификации. Классификатор строится на исходном обучающем множестве, документы которого учавствуют в обучении с некоторыми весовыми коэффициентами . После обучения классификатор проверяется на исходной обучающейся выборке и происходит пересчет коэффициентов. Коэффициент уменьшается, если документ классифицирован верно, и увеличивается в противном случае.

План разработки классификатора. Вначале разрабатывается каркас ансамбля классификаторов. Затем итеративно заполняется каждая его часть. Наращивание функционала каждой части происходит после полной итерации по всем частям ансамбля, а именно:

1. Разработка инфраструктуры: сбор корпусов, выборок и онтологий для создания классификатора; извлечение признаков из тренировочной выборки; построение верхнего уровня ансамбля; оценка качества работы системы; добавление признаков из онтологий; оценка качества работы системы с добавочными признаками из онтологии.

2. Итерации улучшения работы классификатора: добавить признаки из тренировочной выборки; оценка качества работы системы с новыми признаками; добавить признаки из онтологии; оценка качества работы системы с новыми признаками; оценка качества работы классификатора на основном выборке.

На рис. 2 показан каркас предлагаемого ансамбля классификаторов.

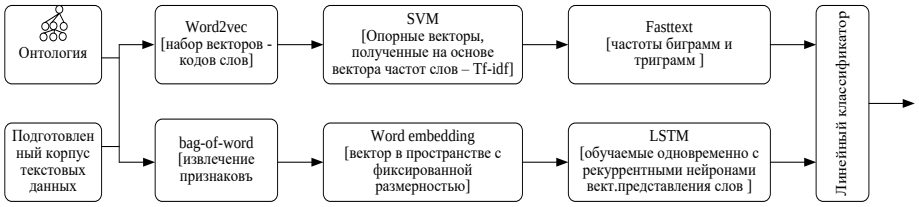

Рис. 2. Каркас ансамбля классификаторов

Реализация предлагаемого классификатора выполнена в среде Matlab. Используемые функции: extractFileText (чтение текстовых данных), erasePunctuation (очистка текста от пунктуационных знаков), normalizeWords (нормализация слов), stopWords (удаление стоп-слов), tfidftokenizedDocument (текенизация документов) bagOfWords (построение «мешка слов»), trainWordEmbedding (обучение векторной модели), построение модели word2vec и др. [19].

В качестве тестирования была выбрана выборка новостных статей Reuters (Reuters-21578). Reuters-21578 [20] используется во многих исследованиях для проверки качества разработанного алгоритма кластеризации. Приведем характеристики выборок:

- общие характеристики – 91 категория для классификации; обе выборки несбалансированны по количеству входящих в категории элементов;
- тренировочная выборка – 11413 статей для обучения;
- тестовая выборка – 4024 статей для классификации.

Корпуса для обучения. В качестве «Word Embedding» предлагается взять предварительно обученную на Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download) коллекцию GloVe векторов для английского языка.

Онтологии: DBpedia – проект, направленный на извлечение структурированной информации из данных, созданных в рамках проекта Википедия и публикации её в виде доступных под свободной лицензией наборов данных. Baseline. В качест-

ве baseline для сравнения предлагаемого метода предлагается брать методы на предыдущей итерации обновления классификатора. Качество классификации измерялось по метрике precision (точность классификации) с микроусреднением [21].

Результаты экспериментов приведены в таб. 1.

Таблица 1

Результаты экспериментов

Название метода	Признаки	Точность	Скорость обучения
SVM	Tf-idf	73,5 %	Высокая
Fasttext классификатор	Частоты биграмм и триграмм (стандартные настройки классификатора)	71,3 %	Высокая
Word embedding	Обучаемые одновременно со свёрткой векторные представления слов (word embedding). Окрестность свертки векторное представление двух слов	72,3 %	Низкая
LSTM	Обучаемые одновременно с рекуррентными нейронами векторные представления слов (word embedding).	70 %	Крайне низкая

В работе проведен анализ существующих методов классификации для задачи индексации документов, методов построения и обучения классификатора с помощью машинного обучения, проведена оценка эффективности методов мультиклассовой классификации документов:

- SVM-классификатор – настраивать параметры достаточно легко, но необходимо предварительно представить текст в виде модели признаков (tf-idf);
- Fasttext-классификатор – настраивать параметры достаточно легко, нужно варьировать количество итераций в процессе обучения;
- Word embedding - сложнее первых двух методов; нужно строить архитектуру и настраивать гиперпараметры;
- LSTM-классификатор сложнее первых двух методов; нужно строить архитектуру и настраивать гиперпараметры.

В заключении приведено обоснование выбора наиболее приоритетных методов мультиклассовой классификации. Несмотря на то, что все методы проявили себя, с точки зрения точности, приблизительно одинаково, наиболее высокие показатели у SVM, что объясняется небольшим количеством данных. Чем больше данных, тем выше качество нейросетевых методов и ниже у метода опорных векторов. Следовательно:

1. Если данных мало – рекомендуется использовать метод SVM.
2. Если данных на класс много – рекомендуется метод CNN или LSTM.
3. При переходе от малого числа данных к большему стоит попробовать fastText.

Заключение. В статье рассмотрена проблема интеллектуальной обработки научно-технической информации, представленной в открытых источниках, с целью извлечения знаний для построения онтологии предметной области. Предложен алгоритм кластеризации корпуса текстов, реализованный с применением функций Matlab. Предложенный новый вариант ансамбля классификаторов (SVM – метод опорных векторов, LSTM – нейронная сеть, word embedding – дистрибтивная семантика), предназначенный для систем автоматического построения онтоло-

гий по коллекции тестовых документов, тематически относящихся к одной предметной области, показал высокие результаты по точности классификации. Предлагаемый подход обладает простотой реализации и вычислительной эффективностью. В дальнейшем планируется строить высокоуровневое признаковое описание текста. Для этого необходимо провести синтаксический анализ предложений текста, на основе которого выполнить семантическую разметку ролей и в дальнейшем извлекать факты вида троек – объект, предикат, субъект.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Найханова Л.В.* Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования: монография. – Улан-Удэ: Изд-во БНЦ СО РАН, 2008. – 244 с.
2. *Бубарева О.А.* Математическая модель процесса интеграции информационных систем на основе онтологий // Современные проблемы науки и образования. – 2012. – № 2. – URL: www.science-education.ru/102-6030.
3. *Semenova A.V., Kureichik V.M.* Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis // Proceedings of the 6th Computer Science Online Conference 2017 (CSOC2017). – Vol .1. – P. 302-311.
4. *Пархоменко П.А., Григорьев А.А, Астраханцев Н.А.* Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. – 2017. – Т. 29. – Вып. 2. – С. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6.
5. *Andrews Nicholas O, Fox Edward A.* Recent developments in document clustering: Tech. Rep.: Technical report, Computer Science, Virginia Tech, 2007.
6. *Aggarwal Charu C, Zhai Cheng Xiang.* Mining text data. Springer Science & Business Media, 2012.
7. *Whissell John S., Clarke Charles L.A.* Improving document clustering using Okapi BM25 feature weighting // Information retrieval. – 2011. – Т. 14, No. 5. – P. 513-523.
8. *Huang Anna.* Similarity measures for text document clustering // Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008. – P. 49-56.
9. *Sathiyakumari K., Manimekalai G., Preamsudha V.* A survey on various approaches in document clustering // International Journal of Computer Technology and Applications. – 2011. – Vol. 2 (5). – P. 1534-1539.
10. *Aggarwal Charu C, Zhai Cheng Xiang.* Mining text data. Springer Science & Business Media, 2012.
11. *Marchionini Gary.* Exploratory search: from finding to understanding // Communications of the ACM. – 2006. – Vol. 49, No. 4. – P. 41-46.
12. *Вьюгин В.В.* Математические основы машинного обучения и прогнозирования. Электронное издание. – М.: МЦНМО, 2014. – 304 с.
13. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. – 2013. – https://arxiv.org/abs/1301.3781.
14. *Pennington J., Socher R., Manning Ch. D.* GloVe: Global Vectors for Word Representation. – http://www.aclweb.org/anthology/D14-1162.
15. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching Word Vectors with Subword Information. – https://arxiv.org/abs/1607.04606.
16. *Choi F., Wiemer-Hasting P., Moore J.* Latent semantic Analysis for Text Segmentation // Proceedings of NAACL'01, Pittsburgh, PA, 2001. – P. 109-117.
17. *Gama J.* Knowledge Discovery from Data Streams. Singapore, CRC Press Pubh, 2010. DOI: 10.1201/EBK1439826119.
18. *Tomin N., Zhukov A., Sidorov D., Kurbatsky V., Panasetsky D., Spiryaev V.* Random Forest Based Model for Preventing Large-Scale Emergencies in Power Systems // International Journal of Artificial Intelligence. – 2015. – Vol. 13, no. 1. – P. 221-228.
19. КиберЛенинка. – https://cyberleninka.ru/article/n/modifikatsiya-algoritma-sluchaynogo-lesa-dlya-klassifikatsii-nestatsionarnyh-potokovyh-dannyh.

20. *Дьяконов В., Круглов В.* Математические пакеты расширения MATLAB. Специальный справочник. – СПб.: Питер, 2001. – 480 с.
21. Reuters-21578 Text Categorization Test Collection. – http://www.daviddlewis.com/resources/testcollections/reuters21578.
22. *Агеев М., Кураленок И., Некрестьянов И.* Официальные метрики РОМИП 2006. – Режим доступа: http://romip.ru/romip2006/appendix_a_metrics.pdf.

REFERENCES

1. *Nayhanova L.V.* Tekhnologiya sozdaniya metodov avtomaticheskogo postroeniya ontologiy s primeneniem geneticheskogo i avtomatnogo programmirovaniya: monografiya [Technology the development of methods for the automatic construction of ontologies with the use of genetic and automata-based programming: monograph]. Ulan-Ude: Izd-vo BNTs SO RAN, 2008, 244 p.
2. *Bubareva O.A.* Matematicheskaya model protsessa integratsii informatsionnyikh sistem na osnove ontologiy [Mathematical model of the process of integration of information systems based on ontologies], *Sovremennye problemy nauki i obrazovaniya* [Modern problems of science and education], 2012, No. 2. Available at: www.science-education.ru/102-6030.
3. *Semenova A.V., Kureichik V.M.* Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis, *Proceedings of the 6th Computer Science Online Conference 2017 (CSOC2017)*, Vol .1, pp. 302-311.
4. Parkhomenko P.A., Grigorev A.A, Astrakhantsev N.A. Obzor i eksperimental'noe sravnenie metodov klasterizatsii tekstov [Review and experimental comparison of text clustering methods], *Trudyi ISP RAN* [Proceedings of ISP RAS], 2017, Vol. 29, Issue. 2, pp. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6.
5. *Andrews Nicholas O, Fox Edward A.* Recent developments in document clustering: Tech. Rep.: Technical report, Computer Science, Virginia Tech, 2007.
6. *Aggarwal Charu C, Zhai Cheng Xiang.* Mining text data. Springer Science & Business Media, 2012.
7. *Whissell John S., Clarke Charles L.A.* Improving document clustering using Okapi BM25 feature weighting, *Information retrieval*, 2011, Vol. 14, No. 5, pp. 513-523.
8. *Huang Anna.* Similarity measures for text document clustering, *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. 2008*, pp. 49-56.
9. *Sathiyakumari K., Manimekalai G., Preamsudha V.* A survey on various approaches in document clustering, *International Journal of Computer Technology and Applications*, 2011, Vol. 2 (5), pp. 1534-1539.
10. *Aggarwal Charu C, Zhai Cheng Xiang.* Mining text data. Springer Science & Business Media, 2012.
11. *Marchionini Gary.* Exploratory search: from finding to understanding, *Communications of the ACM*, 2006, Vol. 49, No. 4, pp. 41-46.
12. *Vyugin V.V.* Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya [Mathematical foundations of machine learning and forecasting]. Moscow: MTsNMO, 2014, 304 p.
13. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space, *In Proceedings of Workshop at ICLR*, 2013. Available at: https://arxiv.org/abs/1301.3781.
14. *Pennington J., Socher R., Manning Ch. D.* GloVe: Global Vectors for Word Representation. Available at: http://www.aclweb.org/anthology/D14-1162.
15. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching Word Vectors with Subword Information. Available at: https://arxiv.org/abs/1607.04606.
16. *Choi F., Wiemer-Hasting P., Moore J.* Latent semantic Analysis for Text Segmentation, *Proceedings of NAACL'0*1, Pittsburgh, PA, 2001, pp. 109-117.
17. *Gama J.* Knowledge Discovery from Data Streams. Singapore, CRC Press Pubh, 2010. DOI: 10.1201/EBK1439826119.
18. *Tomin N., Zhukov A., Sidorov D., Kurbatsky V., Panasetsky D., Spiryaev V.* Random Forest Based Model for Preventing Large-Scale Emergencies in Power Systems, *International Journal of Artificial Intelligence*, 2015, Vol. 13, no. 1, pp. 221-228.
19. KiberLeninka. Available at: https://cyberleninka.ru/article/n/modifikatsiya-algoritma-sluchaynogo-lesa-dlya-klassifikatsii-nestatsionarnyh-potokovyh-dannyh.

20. *Dyakonov V., Kruglov V.* Matematicheskie pakety rasshireniya MATLAB. Spetsialnyy spravochnik [Mathematical expansion packs MATLAB. Special reference]. Saint Petersburg: Piter, 2001, 480 p.
21. Reuters-21578 Text Categorization Test Collection. Available at: http://www.daviddlewis.com/resources/testcollections/reuters21578.
22. *Ageev M., Kuralenok I., Nekrestyanov I.* Ofitsialnyie metriki ROMIP 2006 [Official metrics of ROMIP 2006]. Available at: http://romip.ru/romip2006/appendix_a_metrics.pdf.

Статью рекомендовал к опубликованию д.т.н., профессор Я.Е. Ромм.

Курейчик Виктор Михайлович – Южный федеральный университет; e-mail: vmkureychik@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; д.т.н., профессор.

Семенова Александра Владимировна – e-mail: alexaforum@rambler.ru; тел.: +79153274378; кафедра систем автоматизированного проектирования; аспирант.

Kureychik Viktor Mikhailovich – Southern Federal University; e-mail: vmkureychik@sfedu.ru; 44, Nekrasovskiy street, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design systems; dr. of eng. sc.; professor.

Semenova Aleksandra Vladimirovna – e-mail: alexaforum@rambler.ru; phone: +79153274378; the department of computer aided design systems; postgraduate.