УДК 004.912:19.254
DOI: 10.24160/1993-6982-2022-1-111-119

Построение терминологических профилей научных сотрудников на основе публикаций в цифровой библиотеке eLIBRARY.RU

П.А. Козлов, А.С. Мохов, В.О. Толчеев

Проведено построение терминологических профилей специалистов на основе публикаций из цифровой библиотеки eLIBRARY. Усреднение индивидуальных терминологических профилей позволяет составить обобщенный профиль («портрет») малого научного коллектива (кафедры, лаборатории, отдела). Сопоставление индивидуальных профилей с помощью мер близости (например, косинусной меры) дает возможность группировать схожие профили и выявлять группы сотрудников, проводящих исследования в одной предметной области. Это помогает определять специализацию научного коллектива средствами Text Mining без использования субъективных экспертных оценок. Результаты, полученные с помощью профильного подхода, подтверждены путем построения графов соавторства и графа терминов в программе Gephi.
Составление терминологических профилей использовано также при разработке персонализированных систем поддержки научной деятельности. Данная система предназначена для оказания помощи пользователю (специалисту-предметнику) в выборе релевантных научных конференций и поиске полезных (по возможности пертинентных) публикаций. Для описания текстовых документов в работе взята векторная модель, веса терминов установлены с помощью расчета частоты встречаемости термина (или формулы tfc-взвешивания). На стадии предварительной обработки отсечены стоп-слова, редкочастотные слова и проведена лемматизация.
Разработанный профильный подход апробирован на примере малого научного коллектива, специализирующегося в области компьютерных наук (Computer Science). Построение терминологических профилей и их анализ позволил выделить направления, по которым специализируется члены коллектива, и разработать персонализированную систему поддержки научной деятельности, отслеживающую в автоматизированном режиме публикации в eLIBRARY по одному из актуальных направлений («Интеллектуальный анализ данных»).

Ключевые слова: интеллектуальный анализ текстовых данных, терминологический профиль, персонализированная система поддержки научной деятельности, взвешивание терминов, косинусная мера, граф Gephi.

Для цитирования: Козлов П.А., Мохов А.С., Толчеев В.О. Построение терминологических профилей научных сотрудников на основе публикаций в цифровой библиотеке eLIBRARY.RU // Вестник МЭИ. 2022. № 1. С. 111—119. DOI: 10.24160/1993-6982-2022-1-111-119.

Drawing Up Researcher Terminological Profiles Based on Publications in the Digital Library eLIBRARY.Ru

P.A. Kozlov, A.S. Mokhov, V.O. Tolcheev

The terminological profiles of specialists are drawn up based on publications from the digital library eLIBRARY. By averaging the individual terminological profiles, it is possible to draw up a generalized profile (“portrait”) of a small research team (a department, laboratory, or sector). Comparison of individual profiles with the use of proximity measures (for example, a cosine measure) makes it possible to group similar profiles and identify groups of employees who conduct research in the same subject area. This helps determine the research team specialization by means of Text Mining tools without using subjective expert assessments. The results obtained using the profile approach are confirmed by constructing graphs of co-authorship and a graph of terms in the Gephi computer program.
The compilation of terminological profiles was also used in the development of personalized scientific activity support systems. This system is intended for helping the user (a specialist in a subject area) in choosing relevant scientific conferences and searching for useful (pertinent as far as possible) publications. For describing text documents, a vector model is used, and the weights of terms are determined by calculating the term occurrence frequency (or the tfc-weighting formula). At the preprocessing stage, the stop-words and rarely encountered words are removed, and lemmatization is carried out. The developed profile approach has been approbated on the example of a small research team specializing in computer science. The terminological profiles were constructed and analyzed, based on which the areas in which the team members specialize have been identified, and a personalized scientific activity support system has been developed, that tracks, in an automated mode, publications in the eLIBRARY in one of the relevant areas (Data Mining).

Key words: text data mining, terminological profile, personalized scientific activity support system, term weighting, cosine measure, Gephi graph.

For citation: Kozlov P.A., Mokhov A.S., Tolcheev V.O. Drawing Up Researcher Terminological Profiles Based on Publications in the Digital Library eLIBRARY.Ru. Bulletin of MPEI. 2022;1:111—119. (in Russian). DOI: 10.24160/1993-6982-2022-1-111-119.

Введение

Создание и развитие больших специализированных хранилищ текстовых данных существенно расширили возможности автоматизированного построения терминологических профилей (ТП) специалистов на основе их «цифровых» следов: публикаций, докладов на конференциях, диссертационных работ, отчетов по НИР. Под терминологическим профилем понимается вектор, содержащий термины, встречающиеся в публикациях специалиста и отражающие его профессиональные интересы.

Построение профилей («портретов») необходимо при разработке персонализированных систем поддержки научной деятельности, автоматизирующих поиск полезных публикаций, конференций и т. д. [1 — 3]. В этом случае на основе пользовательского ТП «предсказывается» информационная потребность и выбираются наиболее ценные сведения из предметной области. Создание ТП полезно при анализе деятельности научных коллективов (прежде всего на уровне кафедры, лаборатории, отдела) [4, 5]. Путем усреднения индивидуальных ТП можно составить обобщенный профиль коллектива и, введя меру близости между ТП, выявить тематики исследований, по которым проводятся наиболее интенсивные работы. Это позволяет установить специализацию научного коллектива средствами Text Mining без использования субъективных экспертных оценок [6, 7]. В образовательном учреждении индивидуальные ТП могут служить для оценки степени соответствия ТП преподавателя и терминологического состава разделов учебной программы читаемых куров [8 — 10].

Опишем формальную постановку задачи.

Пусть имеется множество русскоязычных публикаций сотрудников малого научного коллектива — $\{\boldsymbol{X}\}$. Требуется:

- построить индивидуальные профили сотрудников $\boldsymbol{X}_j$, состоящие из терминов, наиболее часто встречающихся в публикациях ($j = 1, \ldots, n$, n — число членов коллектива и, соответственно, профилей);
- сформировать терминологический портрет научного коллектива $\bar{\boldsymbol{X}}$ путем усреднения индивидуальных профилей;
- изучить терминологический портрет научного коллектива на однородность (соответствие всех публикаций одной предметной области), в случае неоднородности — выявить узкоспециализированные тематики, по которым работают несколько специалистов;
- на основе построенных профилей разработать персонализированную систему поддержки научной деятельности, которая в автоматизированном режиме будет отслеживать в eLIBRARY публикации по наиболее активно развивающимся направлениям специализации научных сотрудников.

Формирование выборок и предварительная обработка данных

В настоящее время публикационная активность ученых и преподавателей в российских (и переводных) изданиях отражается в открытой электронной библиотеке eLIBRARY (https://www.elibrary.ru). На основе сведений из нее рассчитываются различные наукометрические показатели, отслеживаются перспективные тенденции в предметных областях, проводится анализ выполнения программ развития вузов, строятся профили студентов и преподавателей [4, 5, 11]. В настоящей работе для построения терминологических профилей специалистов использованы библиографические описания научных работ, проиндексированные в eLIBRARY (описания включают название, аннотацию, ключевые слова, сведения об авторе и вспомогательную информацию).

Для составления выборки в библиотеке последовательно выполняются следующие операции: «Навигатор» → «Авторский указатель» → «Поиск автора» (по фамилии, имени и отчеству) → «Получение списка публикаций автора».

Выборка, характеризующая деятельность малого научного коллектива, состоит из 387 публикаций, проиндексированных в eLIBRARY в период с 1991 по 2019 гг. Для отсечения сотрудников, в настоящее время не участвующих в исследованиях, взяты статьи только тех специалистов, которые за 2009 — 2019 гг. имели в библиотеке не менее пяти работ. Размер малого научного коллектива — 13 человек.

Для описания документов использована модель «мешок слов», представляющая тексты в виде отдельных терминов (без учета контекста их появления и связи с другими словами) [12, 13]. Отметим, что применение словосочетаний (в качестве единого дескриптора) в подобных задачах малоэффективно, поскольку при этом снижается совстречаемость понятий в документах и ослабляется терминологическое сходство между тематически близкими текстами и профилями. Работа с более сложными нейросетевыми моделями представления текста также нецелесообразна из-за небольшого размера исходной выборки [14, 15].

Предварительная обработка документальных массивов включает следующие этапы [12, 16]:

- удаление знаков препинания и понижение регистра;
- исключение стоп-слов, не несущих полезной (лексической) информации (предлогов, союзов, междометий, частиц и т. д.);
- удаление редко встречающих слов (отсекались слова, встречавшиеся в выборке один или два раза);
- приведение слов к начальной форме (нормализация).

Для нормализации в Text Mining используются два подхода [12].

Стемминг — поиск начальной формы с помощью обрезания (заранее заданных в программе) окончаний и суффиксов. Для русского языка, в котором (в отличие, например, от английского) широко распространены сложные формы словообразования (падежи, склонения, времена и т.д.), стемминг чаще всего малоэффективен и плохо интерпретируется.

Лемматизатор — определение начальной формы на основе применения специальных словарей, учитывающих всевозможные варианты словообразования. Качество лемматизации существенно зависит от полноты словаря и специфики терминологии предметной области.

В нашем случае признаковое пространство, в котором описываются документы, формируется из лемматизированных терминов, оставшихся после удаления стоп- и редко встречающихся слов.

Построение терминологических профилей и анализ результатов

Построим терминологический профиль $\boldsymbol{X}_j$ в виде вектора:

$$\boldsymbol{X}_j = \begin{bmatrix} x_j^{(1)} \\ \dots \\ x_j^{(M)} \end{bmatrix}, \qquad (1)$$

где $j = 1, \dots, n$; n — размер малого научного коллектива ($n = 13$); M — размер словаря (все информативные термины, содержащиеся в индивидуальных профилях).

Веса в формуле (1) определяются частотой встречаемости слова:

$$x_j^{(i)} = f_{ij},$$

где f_{ij} — частота встречаемости i-го термина в документах j-го специалиста.

Для проведения сравнения размер всех ТП должен быть одинаковым, в исследуемом научном коллективе он составляет $M = 1500$. Такой словарь получается после удаления стоп- и редкочастотных слов ($f_{ij} = 1$ и $f_{ij} = 2$), а также проведения лемматизации.

Следует отметить, что профили получаются разреженными — слова, не встречающиеся в публикациях сотрудника, имеют в его профиле нулевой вес.

В качестве меры близости между двумя профилями $\boldsymbol{X}_j$ и $\boldsymbol{X}_k$ использована косинусная мера (Cosine similarity), рассчитываемая следующим образом [16, 17]:

$$\cos\alpha = \frac{(X_j, X_k)}{|X_j||X_k|} = \frac{\sum_{i=1}^{M} x_j^{(i)} x_k^{(i)}}{\sqrt[2]{\sum_{i=1}^{M}\left(x_j^{(i)}\right)^2}\sqrt[2]{\sum_{i=1}^{M}\left(x_k^{(i)}\right)^2}}. \qquad (2)$$

Здесь $x_j^{(i)}$, $x_k^{(i)}$ — частоты встречаемости i-го слова в профилях j-го и k-го преподавателей ($j, k = 1, \dots, n$; $i = 1, \dots, M$).

Для построения терминологического портрета малого научного коллектива $\bar{\boldsymbol{X}}$ усредним веса каждого термина по профилям сотрудников:

$$f_i^{\text{сред}} = \frac{1}{n}\sum_{j=1}^{n} f_{ij}.$$

Терминологический портрет научного коллектива выполнен с помощью программы Word Cloud и показан на рис. 1. Размер термина зависит от частоты его встречаемости в публикациях сотрудников кафедры.

Анализ полученного портрета позволяет предположить его неоднородность, т. е. наличие различных тематик (направлений) исследований. Для подтверждения этого предположения определим близость между профилями сотрудников с помощью косинусной меры (2) и проиллюстрируем полученные результаты дендрограммой (рис. 2). Для обозначения членов малого научного коллектива на рис. 2, 3 взяты их инициалы.

В ходе исследований использованы различные способы объединения кластеров, давшие достаточно близкие результаты. На рисунке 2 показан наилучший

Рис. 1. Терминологический портрет исследуемого малого научного коллектива

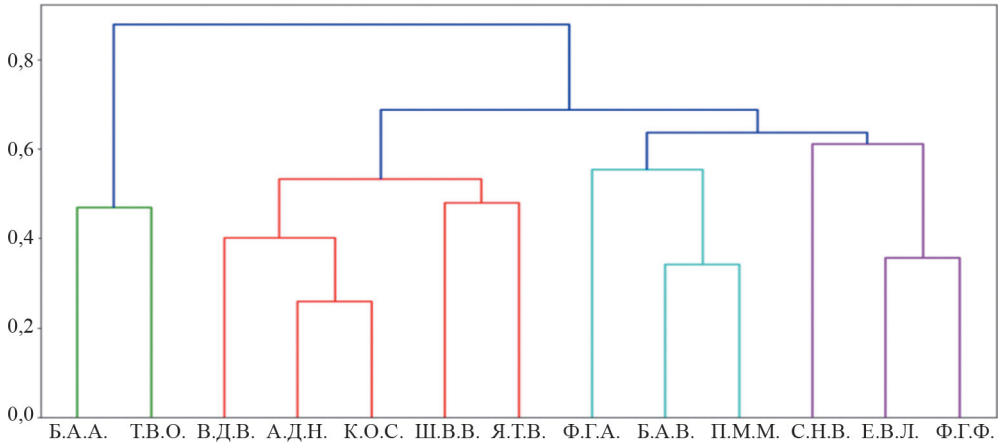

Рис. 2. Дендрограмма для косинусной меры и метода полной связи

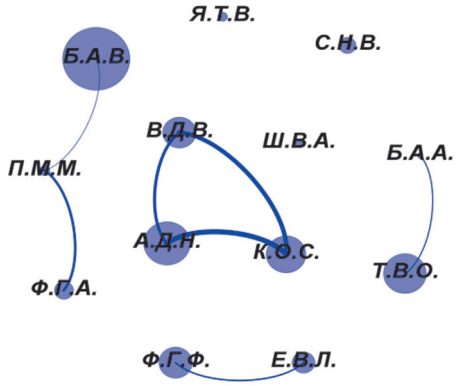

Рис. 3. Граф соавторства сотрудников кафедры УИТ

вариант разбиения с помощью метода полной связи (правила «дальнего соседа»).

Изучение индивидуальных профилей и полученных кластеров позволяет сделать вывод, что в малом научном коллективе есть четыре группы, состоящие, как минимум, из двух сотрудников. Анализ ТП, вошедших в один кластер, позволил назвать их следующим образом: «Нечеткие системы управления», «Нейросети», «Информационно-аналитические системы», «Анализ текстовых данных».

Возникает закономерный вопрос — можно ли сделать вывод о специализации малого научного коллектива на основе только одного исследования (с применением ТП)? Для подтверждения устойчивости и достоверности результатов стоит провести исследования с помощью других способов анализа данных. Для этого был выбран графовый метод, обладающий хорошей интерпретируемостью и реализованный в программе Gephi [18].

На рисунке 3 дан граф, построенный на основе матрицы «авторство–соавторство». Размер вершин на полученном графе определяли числом публикаций автора, наличие и толщина ребер зависела от количества статей, написанных в соавторстве.

На графе соавторства отчетливо выделяются четыре самостоятельные научные группы, что практически полностью совпадает с результатами, полученными с помощью профильного подхода (см. рис. 2). Выявленные группы немногочисленны и компактны, между ними отсутствуют связи, которые приводили бы к образованию одного большого кластера и свидетельствовали об однородности тематик и их высокой взаимосвязанности. Граф соавторства позволяет установить дополнительную полезную информацию — часть специалистов, объединенных на дендрограмме в общий кластер (при достаточно больших значениях меры близости), не имеют общих публикаций и ведут «независимые» исследования. Например, среди сотрудников С.Н.В, Ф.Г.Ф, Е.В.Л, объединенных в один кластер на основе близости их профилей, у С.Н.В. нет общих публикаций с двумя другими членами кластера. Граф соавторства, а также знание некоторых социальных связей внутри коллектива, позволяет сделать вывод, что наиболее часто совместные публикации имеют научные руководители и аспиранты, а также исполнители общей НИР.

Для подтверждения неоднородности исследований, ведущихся внутри малого научного коллектива, с помощью программы Gephi построена карта терминов, визуализирующая матрицу «термин–термин», и проведена раскраска, объединяющая общим цветом термины, наиболее часто встречающиеся совместно (рис. 4). Для упрощения графа и упрощения его интерпретации использовано только 10% наиболее высокочастотных терминов [19].

Изучение имеющейся выборки с помощью программы Gephi подтверждает результаты, полученные с помощью ТП, позволяет сделать вывод о неоднородности исследований и свидетельствует о наличии в малом научном коллективе нескольких самостоятельных научных групп, специализирующихся в различных научных направлениях. При этом большая часть исследований, судя по публикациям, смещена в область интеллектуальных технологий или, более конкретно, интеллектуального анализа данных (ИАД, Data and Text Mining). Публикации, сделанные в рамках ИАД, присутствуют в кластерах «Нейросети», «Информационно-аналитические системы», «Анализ текстовых данных».

Разработка персонализированной системы поддержки научной деятельности

С учетом высокого интереса сотрудников исследуемого коллектива к предметной области «Интеллектуальный анализ данных» было принято решение о разработке персонализированной системы поддержки научной деятельности, способной в автоматизированном режиме отслеживать публикации по ИАД, проводя ежемесячный мониторинг цифровой библиотеки eLIBRARY, и предоставлять специалистам для изучения наиболее ценные публикации. Данная система должна обеспечивать:

- подключение к API цифровой библиотеки.
- отправку поискового запроса с соответствующими параметрами.
- выгрузку списка ссылок на найденные статьи.
- выгрузку найденных статей по их ссылкам.

Персонализированная система поддержки научной деятельности предусматривает выявление статей по ИАД из общего документального потока. Фильтрация в нашем случае рассматривалась как задача бинарной классификации [12, 16].

Для формирования класса «ИАД» взят запрос в eLIBRARY: «Интеллектуальный анализ данных». Вид запроса и дополнительные условия поиска приведены на рис. 5. По заданному поисковому запросу по тематике ИАД было найдено 1600 статей, изданных в период с 1991 по 2019 гг.

Для формирования класса «не ИАД» использован запрос в eLIBRARY: «Управление в технических си-

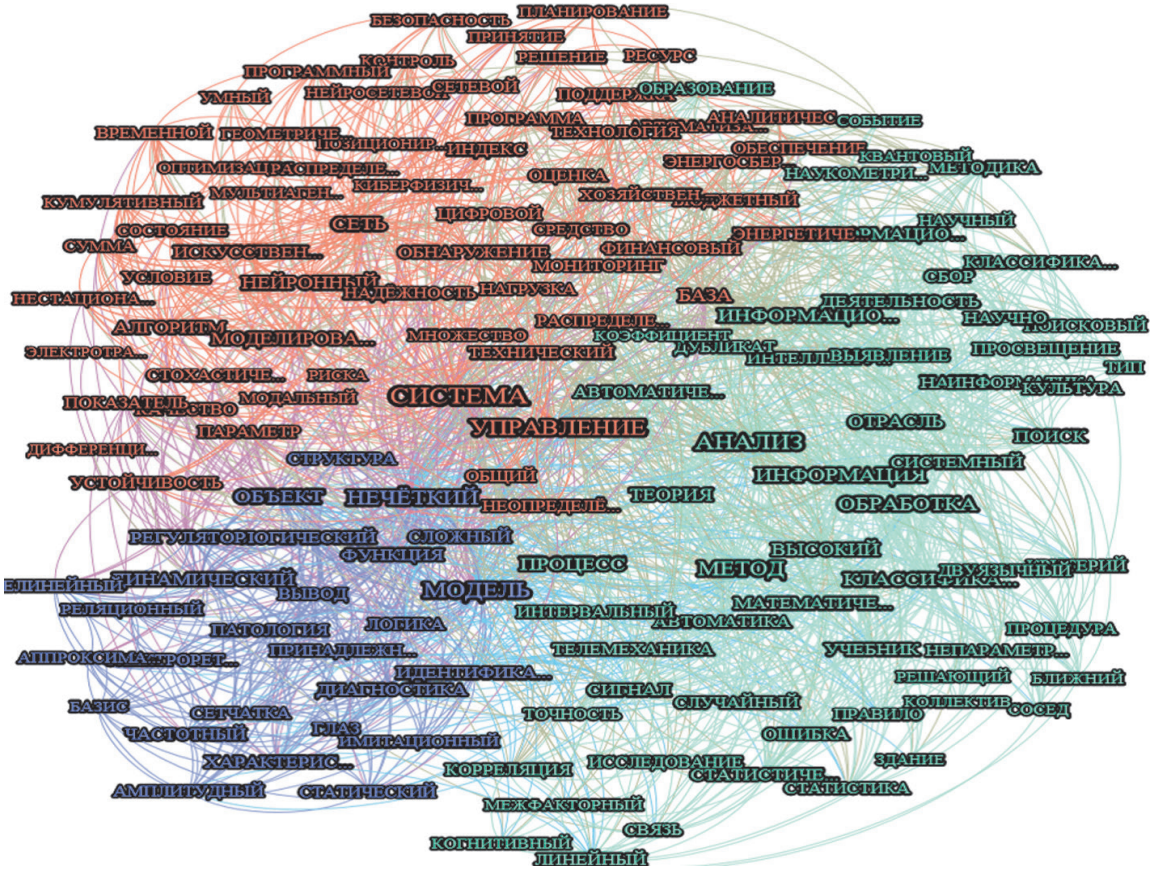

Рис. 4. Граф терминов, построенный с помощью программы Gephi

стемах», поскольку эта тема является одним из направлений исследований изучаемого научного коллектива. Окончательный размер полученной выборки составил $N = 3100$ статей (1600 статей принадлежат классу «ИАД» и 1500 статей — классу «не ИАД»). Указанные данные использованы для обучения (параметры настраивали с помощью пятикратной перекрестной проверки). Сформирована выборка из публикаций за 2020 г., состоящая из 400 статей, где 200 статей относятся к тематике «ИАД» и 200 статей — к тематике «не ИАД». Эти данные необходимы для тестирования.

Для описания сформированных выборок взята матрица «документ–термин»:

$$X = \begin{bmatrix} x_1^{(1)} \ldots x_1^{(i)} \ldots x_1^{(M)} \\ \ldots\ldots\ldots\ldots \\ x_j^{(1)} \ldots x_j^{(i)} \ldots x_j^{(M)} \\ \ldots\ldots\ldots\ldots \\ x_N^{(1)} \ldots x_N^{(i)} \ldots x_N^{(M)} \end{bmatrix},$$

где, по аналогии с (1), j обозначает документы выборки ($j = 1, \ldots, N$), а i — термины ($i = 1, \ldots, M$).

Взвешивание терминов проводили с помощью tfc–взвешивания, при котором веса терминов принимали значения от 0 до 1:

$$x_j^{(i)} = \frac{f_{ij} \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[f_{ij}\log\left(\frac{N}{N_i}\right)\right]^2}},$$

где f_{ij} — частота слова i в документе j; N_i — число документов, в которых присутствует i-й термин; N — размер выборки.

Суммирование в знаменателе дроби проведено по всем терминам j-го документа, имеющего i-й термин.

Для распознавания релевантных и нерелевантных документов обучаются бинарные классификаторы [12, 17, 20]. В качестве мер качества классификации использованы Accuracy (верность, правильность), Precision (точность), Recall (полнота), рассчитанные по матрице ошибок (табл. 1).

Приведем формулы для мер качества:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN};$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = TPR = \frac{TP}{TP + FN}.$$

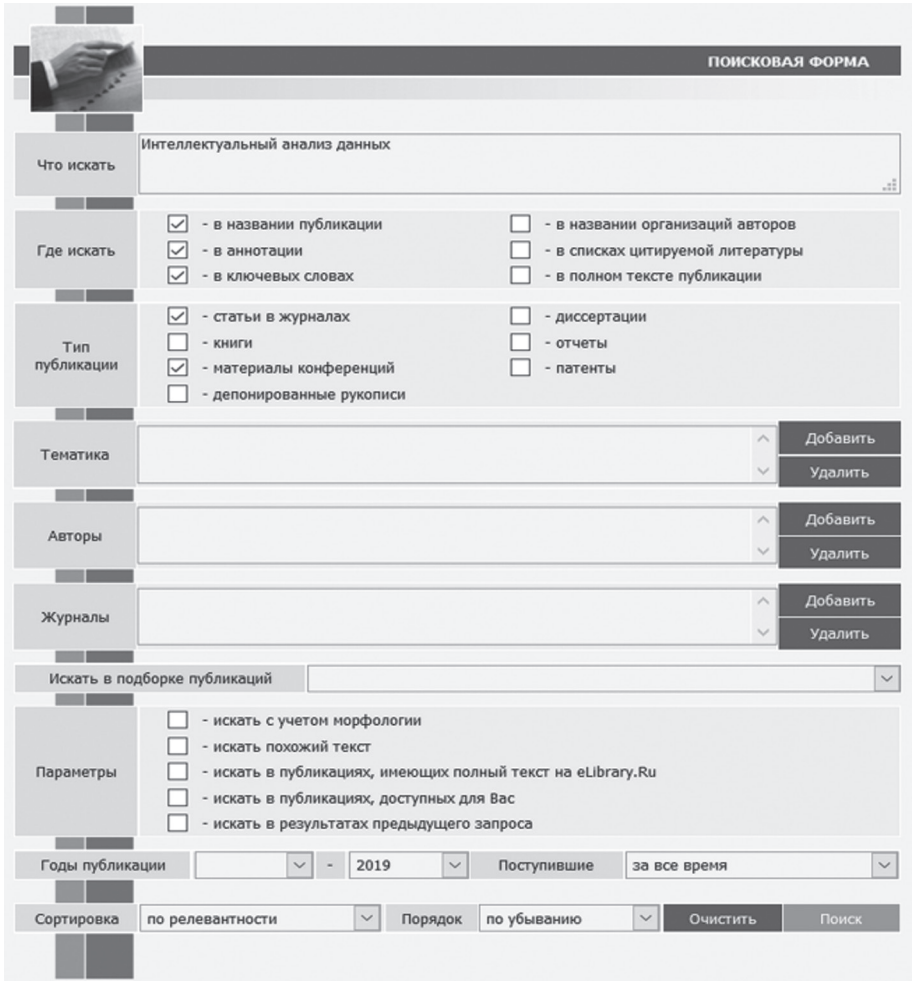

Рис. 5. Вид поискового запроса в eLIBRARY

Таблица 1

Матрица ошибок классификации

Оценка классификатора	Истинная оценка	
	Положительная	Отрицательная
Положительная	TP	FP
Отрицательная	FN	TN

Размер словаря определяли путем отсечения терминов по частоте встречаемости при проведении пятикратной кросс-валидации на обучающей выборке и расчете Accuracy. В ходе проведенных экспериментов размер словаря составил $M = 773$ терминов.

В качестве бинарных классификаторов применяли стандартные алгоритмы, содержащиеся в библиотеке Scikit-learn: метод к-ближайших соседей, логистическую регрессию, «случайный лес», метод опорных векторов. Все выбранные методы реализуют различные решающие правила, что позволяет подобрать алгоритм, наилучшим образом учитывающий специфику исходных данных. Параметры классификаторов настраиваются с помощью пятикратной кросс-валидации.

Результаты, полученные на тестовой выборке, приведены в табл. 2, где под положительным классом понимается «ИАД», под негативным — «не ИАД».

Несомненный лидер на тестовой выборке — «случайный лес». Высокое качество классификации обеспечили также логистическая регрессия и метод опорных векторов. Однако содержательный анализ публикаций в классе «ИАД» показывает, что высокое качество классификации не гарантирует пользователю (специалисту-предметнику) получение пертинентных публикаций, соответствующих информационной потребности. В дальнейшем для повышения качества рекомендаций и ранжирования публикаций «ИАД» планируется использовать терминологические портреты специалистов и учитывать их особые предпочтения («Нейросети», «Информационно-аналитические системы», «Анализ текстовых данных»).

Выводы

Решена задача построения терминологических профилей специалистов и научного портрета малого научного коллектива на основе публикаций из цифровой библиотеки eLIBRARY. С помощью кластеризации ТП

Таблица 2

Значения Precision и Recall для тестовой выборки

Классификатор	Класс	Precision	Recall
«Случайный лес»	Негативный	0,98	0,98
	Положительный	0,99	0,98
Логистическая регрессия	Негативный	0,95	0,97
	Положительный	0,97	0,96
Метод опорных векторов	Негативный	0,95	0,97
	Положительный	0,97	0,96
Метод к-ближайших соседей	Негативный	0,84	0,63
	Положительный	0,73	0,89

выявлены группы сотрудников, выполняющих исследования в одной предметной области, и определены основные направления специализации научного коллектива. Для подтверждения результатов профильного подхода построены граф соавторства и карта терминов с помощью программы Gephi. Показана целесообразность применения ТП при разработке персонализированных систем поддержки научной деятельности.

Литература

1. **Aggarwal C.C.** Content-based Recommender Systems. N.-Y.: Springer, 2016. Pp. 139—166.

2. **Андреев А.М., Березкин Д.В., Козлов И.А.** Подход к автоматизированному мониторингу тем на основе обнаружения событий в потоке текстовых документов // Информационно-измерительные и управляющие системы. 2017. № 3. С. 49—60.

3. **Barakhnin V.B., Kozhemyakina O.Yu., Mukhamediev R.I., Borzilova Yu.S., Yakunin K.O.** The Design of Structure of the Software System for Processing Text Document Corpus // Business Informatics. 2019. No. 4. Pp. 60—72.

4. **Васенин В.А., Афонин С.А., Голомазов Д.Д.** К созданию системы управления научной информацией на основе семантических технологий // Знания — Онтологии — Теории: Материалы Всеросс. конф. с международным участием. Новосибирск, 2011. С. 78—87.

5. **Валько Д.В.** Рекомендательная система на основе интеллектуального анализа наукометрического профиля исследователя // Программные продукты и системы. 2018. № 2. С. 275—283.

6. **Shvets A., Devyatkin D., Sochenkov I., Tikhomirov I., Popov K., Yarygin K.** Detection of Current Research Directions Based on Full-text Clustering // Proc. Sci. and Information Conf. London, 2015. Pp. 483—488.

7. **Голицына О.Л., Куприянов В.М., Максимов Н.В.** Информационные и технологические решения в задачах управления знаниями // Научно-техническая информация. 2015. Сер. 1. № 8. С. 1—12.

8. **Slater S., Joksimovic S., Kovanovic V., Baker R.S., Gasevic D.** Tools for Educational Data Mining: a Review // J. Educational and Behavioral Statistics. 2017. V. 42(1). Pp. 85—106.

9. **Мохов А.С., Сафин Ш.И., Толчеев В.О.** Анализ соответствия между научной и учебной деятель-

References

1. **Aggarwal C.C.** Content-based Recommender Systems. N.-Y.: Springer, 2016:139—166.

2. **Andreev A.M., Berezkin D.V., Kozlov I.A.** Podkhod k avtomatizirovannomu Monitoringu Tem na Osnove Obnaruzheniya Sobytiy v Potoke Tekstovykh Dokumentov. Informatsionno-izmeritel'nye i Upravlyayushchie Sistemy. 2017;3:49—60. (in Russian).

3. **Barakhnin V.B., Kozhemyakina O.Yu., Mukhamediev R.I., Borzilova Yu.S., Yakunin K.O.** The Design of Structure of the Software System for Processing Text Document Corpus. Business Informatics. 2019;4:60—72.

4. **Vasenin V.A., Afonin S.A., Golomazov D.D.** K Sozdaniyu Sistemy Upravleniya Nauchnoy Informatsiey na Osnove Semanticheskikh Tekhnologiy. Znaniya — Ontologii — Teorii: Materialy Vseross. Konf. s Mezhdunarodnym Uchastiem. Novosibirsk, 2011:78—87. (in Russian).

5. **Val'ko D.V.** Rekomendatel'naya Sistema na Osnove Intellektual'nogo Analiza Naukometricheskogo Profilya Issledovatelya. Programmnye Produkty i Sistemy. 2018; 2:275—283. (in Russian).

6. **Shvets A., Devyatkin D., Sochenkov I., Tikhomirov I., Popov K., Yarygin K.** Detection of Current Research Directions Based on Full-text Clustering. Proc. Sci. and Information Conf. London, 2015:483—488.

7. **Golitsyna O.L., Kupriyanov V.M., Maksimov N.V.** Informatsionnye i Tekhnologicheskie Resheniya v Zadachakh Upravleniya Znaniyami. Nauchno-tekhnicheskaya Informatsiya. 2015;1;8:1—12. (in Russian).

8. **Slater S., Joksimovic S., Kovanovic V., Baker R.S., Gasevic D.** Tools for Educational Data Mining: a Review. J. Educational and Behavioral Statistics. 2017;42(1): 85—106.

9. **Mokhov A.S., Safin Sh.I., Tolcheev V.O.** Analiz Sootvetstviya Mezhdu Nauchnoy i Uchebnoy Deyatel'-

ностью кафедры с использованием информационных технологий // Дистанционные образовательные технологии: Сб. статей IV Всерос. науч.-практ. конф. 2019. С. 232—236.

10. **Маслихов С.Р., Мохов А.С., Толчеев В.О.** Применение технологий интеллектуального анализа для оценки соответствия научного профиля кафедры и тематик лекционных курсов // «ИНФОТЕХ — 2019»: Сб. статей Всерос. науч.-техн. конф. 2019. С. 129—133.

11. **Бершадский А.М., Бурукина И.П., Акимов А.А.** Информационная система мониторинга деятельности кафедры // Информатизация образования и науки. 2011. № 3(11). С. 12—23.

12. **Маннинг К., Рагхаван П., Шютце Х.** Введение в информационный поиск. М.: Вильямс, 2014.

13. **Chen K., Zhang Z., Long J., Zhang H.** Turning from TF-IDF to TF-IGM for Term Weighting in Text Classification // Expert Syst. Appl. 2016. V. 66. Pp. 245—260.

14. **Joulin A., Grave E., Bojanowski P., Mikolov T.** Bag of Tricks for Efficient Text Classification // Proc. 15 Conf. European Chapter Association for Computational Linguistics, 2017. V. 2. Pp. 427—431.

15. **Rani N., Sharma A., Pathak S.** Text Classification Using Machine Learning Techniques: Comparative study // Intern. J. Future Revolution in Computer Sci. & Communication Eng. 2018. Iss. 3. Pp. 551—555.

16. **Aggarwal C.C.** Machine Learning for Text. N.Y.: Springer, 2018.

17. **Специализированный** информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электрон. ресурс] www.machinelearning.ru (дата обращения 22.02.2021).

18. **Айсина Р.М.** Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1584—1618.

19. **Козлов П.А., Мохов А.С., Толчеев В.О.** Кластеризация научных публикаций кафедры (на основе данных из библиотеки elibrary.ru) // Нечеткие системы, мягкие вычисления и интеллектуальные технологии: Сб. трудов VIII Междунар. науч.-практ. конф. 2020. Т. 2. С. 189—199.

20. **Флах П.** Машинное обучение — наука и искусство построения алгоритмов. М.: ДМК-пресс, 2015.

nost'yu Kafedry s Ispol'zovaniem Informatsionnykh Tekhnologiy. Distantsionnye Obrazovatel'nye Tekhnologii: Sb. Statey IV Vseros. Nauch.-prakt. Konf. 2019:232—236. (in Russian).

10. **Maslikhov S.R., Mokhov A.S., Tolcheev V.O.** Primenenie Tekhnologiy Intellektual'nogo Analiza dlya Otsenki Sootvetstviya Nauchnogo Profilya Kafedry i Tematik Lektsionnykh Kursov. «INFOTEKH — 2019»: Sb. Statey Vseros. Nauch.-tekhn. Konf. 2019:129—133. (in Russian).

11. **Bershadskiy A.M., Burukina I.P., Akimov A.A.** Informatsionnaya Sistema Monitoringa Deyatel'nosti Kafedry. Informatizatsiya Obrazovaniya i Nauki. 2011; 3(11):12—23. (in Russian).

12. **Manning K., Ragkhavan P., Shyuttse Kh.** Vvedenie v Informatsionnyy Poisk. M.: Vil'yams, 2014. (in Russian).

13. **Chen K., Zhang Z., Long J., Zhang H.** Turning from TF-IDF to TF-IGM for Term Weighting in Text Classification. Expert Syst. Appl. 2016;66:245—260.

14. **Joulin A., Grave E., Bojanowski P., Mikolov T.** Bag of Tricks for Efficient Text Classification. Proc. 15 Conf. European Chapter Association for Computational Linguistics, 2017;2:427—431.

15. **Rani N., Sharma A., Pathak S.** Text Classification Using Machine Learning Techniques: Comparative study. Intern. J. Future Revolution in Computer Sci. & Communication Eng. 2018;3:551—555.

16. **Aggarwal C.C.** Machine Learning for Text. N.Y.: Springer, 2018.

17. **Spetsializirovannyy** Informatsionno-analiticheskiy Resurs, Posvyashchennyy Mashinnomu Obucheniyu, Raspoznavaniyu Obrazov i Intellektual'nomu Analizu Dannykh [Elektron. Resurs] www.machinelearning.ru (Data Obrashcheniya 22.02.2021). (in Russian).

18. **Aysina R.M.** Obzor Sredstv Vizualizatsii Tematicheskikh Modeley Kollektsiy Tekstovykh Dokumentov. Mashinnoe Obuchenie i Analiz Dannykh. 2015;1;11: 1584—1618. (in Russian).

19. **Kozlov P.A., Mokhov A.S., Tolcheev V.O.** Klasterizatsiya Nauchnykh Publikatsiy Kafedry (na Osnove Dannykh iz Biblioteki elibrary.ru). Nechetkie Sistemy, Myagkie Vychisleniya i Intellektual'nye Tekhnologii: Sb. Trudov VIII Mezhdunar. Nauch.-prakt. Konf. 2020;2: 189—199. (in Russian).

20. **Flakh P.** Mashinnoe Obuchenie — Nauka i Iskusstvo Postroeniya Algoritmov. M.: DMK-press, 2015. (in Russian).

Сведения об авторах:

Козлов Павел Андреевич — студент кафедры управления и интеллектуальных технологий НИУ «МЭИ», e-mail: kozlov.pavel.andreevih@yandex.ru

Мохов Андрей Сергеевич — кандидат технических наук, доцент кафедры управления и интеллектуальных технологий НИУ «МЭИ», e-mail: asmokhov@mail.ru

Толчеев Владимир Олегович — доктор технических наук, профессор кафедры управления и интеллектуальных технологий НИУ «МЭИ», e-mail: tolcheevvo@mail.ru

Information about authors:

Kozlov Pavel A. — Student of Control and Intelligent Technologies Dept., NRU MPEI, e-mail: kozlov.pavel.andreevih@yandex.ru
Mokhov Andrey S. — Ph.D. (Techn.), Assistant Professor of Control and Intelligent Technologies Dept., NRU MPEI, e-mail: asmokhov@mail.ru
Tolcheev Vladimir O. — Dr.Sci. (Techn.), Professor of Control and Intelligent Technologies Dept., NRU MPEI, e-mail: tolcheevvo@mail.ru

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов
Conflict of interests: the authors declare no conflict of interest

Статья поступила в редакцию: *23.03.2021*
The article received to the editor: *23.03.2021*