УДК 004.021 DOI: 10.17587/it.27.195-201

Т. В. Нгуен, аспирант, e-mail: vietqn1987@gmail.com,
А. Г. Кравец, д-р техн. наук, проф., e-mail: agk@gde.ru,
Волгоградский государственный технический университет, г. Волгоград

Оценка и прогнозирование тенденций развития научных исследований на основе библиометрического анализа публикаций*

Предлагается подход к анализу и прогнозированию тематической эволюции исследований путем выявления восходящего тренда ключевых слов. Статистический анализ лексики публикаций позволяет проследить глубину проникновения новых идей и методов, которая может быть задана частотой появления слов, кодирующих целые концепции. В статье представлены разработанные метод анализа исследовательских тенденций и алгоритм ранжирования статей, основанный на структуре сети прямого цитирования. Данные для эксперимента были извлечены из Web of Science Core Collection, собраны 6696 публикаций в области искусственного интеллекта за период 2005—2016 гг. Для оценки предложенного метода было собрано 3211 публикаций с 2017 по 2019 гг. В результате метод оценивался путем проверки присутствия предсказанных ключевых слов в наборе самых частых терминов за период 2017—2019 гг. и обеспечил точность 73,33 %.

Ключевые слова: *прогнозирование тенденций, тематическая эволюция, библиометрический анализ, ранжирование статей, искусственный интеллект, база данных Web of Science*

Введение

В настоящее время число научных публикаций растет быстрыми темпами, и становится невозможным оставаться в курсе всего, что публикуется. Это затрудняет способность накапливать знания и анализировать данные предыдущих исследовательских работ. Один из самых важных исследовательских вопросов в вычислительном анализе научной литературы заключается в том, содержат ли обширные коллекции научного текста важные подсказки о динамике развития науки, которые помогут предсказать рост и падение научных идей, методов и даже областей знаний. Поэтому обзоры литературы все чаще принимают на себя решающую роль в обобщении результатов прошлых исследований, чтобы эффективно использовать существующую базу знаний, продвигать направление исследований и предоставлять основанное на фактических данных понимание практики применения и поддержания профессиональной экспертизы. При этом возможность заблаговременного прогнозирования научных тенденций может потенциально революционизировать методы исследований, например, предоставляя финансирующим учреждениям возможность оптимизировать распределение ресурсов на перспективные направления.

Известно, что ученые используют различные качественные и количественные подходы к обзору литературы, чтобы понять и систематизировать более ранние результаты. Среди них библиометрия обладает потенциалом для введения систематического, прозрачного и воспроизводимого процесса обзора, основанного на статистических измерениях научной деятельности [1]. Во многих областях исследований используются библиометрические методы для изучения влияния научного направления, влияния научной школы или влияния конкретной статьи [2—4].

По этой причине основной целью данной работы является представление общего подхода к анализу и прогнозированию тематической эволюции конкретной области исследования путем выявления ключевых слов восходящего тренда. Этот подход представлен в виде метода анализа тенденций, алгоритма ранжирования публикаций и визуализации оценок терминов во временных рядах для обнаружения тематических направлений исследований.

*Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 19-07-01200 и № 20-37-90092.

В качестве примера исследования предложенный подход применяется для анализа тематической эволюции исследовательской области искусственного интеллекта (ИИ) с 2005 по 2016 гг. по наукометрическим данным платформы Web of Science (WoS).

Статья организована следующим образом: вначале представлен библиометрический анализ, собранный набор данных и некоторые полученные из библиометрического анализа результаты. Далее описаны алгоритмы ранжирования статей и извлечения значимых ключевых слов, а также результаты тематических тенденций исследования с указанием ключевых слов восходящего тренда. Затем метод оценивается путем обнаружения присутствий предсказанных ключевых слов в наборе истинных (наиболее часто встречающихся) терминов. В заключении приведены будущие направления исследования.

Анализ тенденций научных исследований на основе библиометрического анализа

В библиометрии есть две основные процедуры: анализ производительности и научное картирование. Анализ производительности нацелен на оценку групп научных субъектов (стран, университетов, департаментов, исследователей) и влияния их деятельности на основе библиографических данных. Научное картирование или библиометрическое картирование — это пространственное представление того, как дисциплины, области, специальности и отдельные статьи или авторы связаны друг с другом [5, 6].

В данной работе рассмотрены данные из платформы WoS. Вначале по ключевому слову "Artificial intelligence" и категории "Computer science Artificial intelligence" отфильтровывались релевантные статьи, опубликованные за период 2005—2016 гг. и проиндексированные в ядре базы данных WoS (Web of Science Core Collection). На момент написания этой статьи были собраны для эксперимента 6696 публикаций. Кроме того, для оценки предложенного метода было собрано 3211 публикаций с 2017 по 2019 годы [7, 8].

Методы визуализации используются для представления научной карты и результатов различных аналитических обзоров. Например, исследовательские сети могут быть представлены для проведения анализа научных карт в целях выявления и визуализации тем исследований и их эволюции в течение периодов [9]. На рис. 1 (см. третью сторону обложки) показана эволюция в последовательные периоды времени тематических областей ИИ на базе информации об авторских ключевых словах, полученной из R-пакета Biblioshiny [10].

Альтернативно временной анализ предназначен для обнаружения концептуальной, интеллектуальной или социальной эволюции исследовательской области путем выявления закономерностей, тенденций и сезонности. Среди них обнаружение "взрыва слов" (word burst detection) представляет собой временной анализ, который нацелен на выявление терминов, имеющих высокую интенсивность использования в течение конечных периодов времени [11, 12].

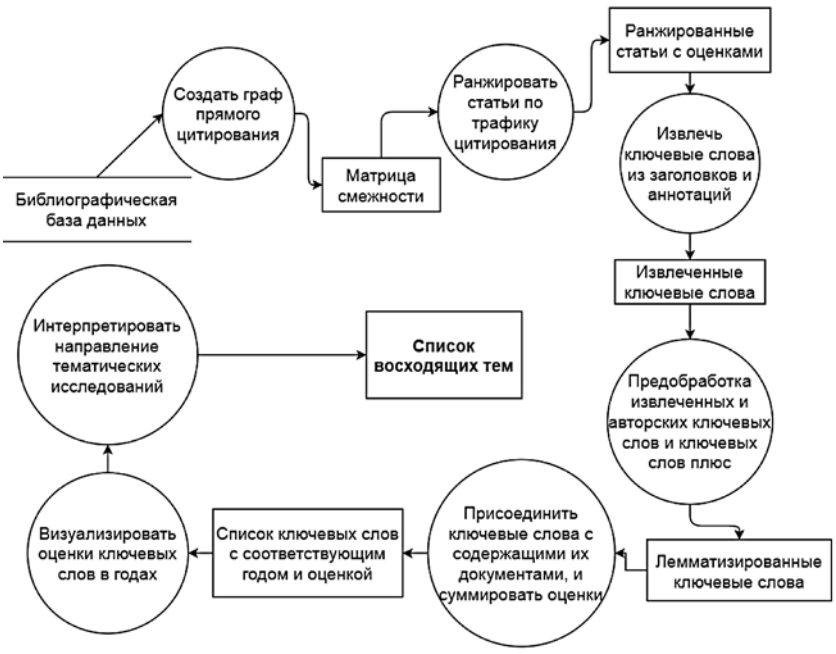

Рис. 2. Диаграмма потока данных метода анализа тенденций научных исследований

Прогнозирование тенденций научных исследований на основе алгоритма ранжирования статей

Анализ вышеупомянутых методов показал, что их недостатком является отсутствие возможности агрегации данных из разделов публикации "Заголовок", "Аннотация", "Авторские ключевые слова" и "Ключевые слова плюс" (Keywords plus). В WoS "Ключевые слова плюс" являются индексными терминами, автоматически генерируемыми из названий статей, и включают фразы из одного или нескольких слов.

Кроме того, существует потребность в анализе и обобщении всех

аспектов самых престижных и влиятельных публикаций (с учетом импакт-фактора). Темы тренда исследования и динамика слов просто идентифицируются по частоте терминов, не учитывая импакт-фактор статей, содержащих эти ключевые слова. При этом все упомянутые выше источники не показывают однозначно тематическую эволюцию через определенный промежуток времени.

Таким образом, предлагается новый метод анализа тенденций научных исследований путем отслеживания оценок ключевых слов в последовательные периоды времени. На рис. 2 продемонстрирована диаграмма потоков данных для предлагаемого метода.

Далее подробно описываются наиболее важные части предлагаемого метода.

Алгоритм ранжирования статей

Цитатный анализ или анализ библиографических ссылок в научных публикациях позволяет определять связи между публикациями, выявлять структуру областей знания и даже прогнозировать их развитие. Авторы работы [13] недавно предположили, что прямое цитирование более точно в представлении фронта исследования, чем библиографическое сочетание и совместное цитирование. Таким образом, предлагаемый алгоритм ранжирования основан на структуре сети прямого цитирования, сформированной из научных работ. Важность и влияние исследовательской работы хорошо отражается в ее цитировании в других публикациях. Исследовательские работы цитируют другие статьи, из которых принимают аргументы и доказательства. Влияние публикации прямо пропорционально важности, качеству и числу исследовательских работ, в которых она цитируется. Поэтому такое допущение используется в нашем алгоритме для ранжирования научных работ путем присвоения каждой из них авторитетной (импакт) оценки. Далее описываются процессы вычисления таких оценок и алгоритм ранжирования статей.

Шаг 1. Исходные данные. Пусть G — ориентированный граф n статей и m цитирований. Тогда $V(G)$ — множество статей, а $E(G)$ множество цитирований.

Далее, пусть для $p_i, p_j \in V(G)$ статья p_i цитирует p_j, если $e = \{p_i, p_j\} \in E(G)$.

Если использовать нотацию в виде графа для описания статьи p, то множество ссылок в списке литературы статьи p можно представить как набор $C_G^+(p) = \{r \in V(G) | (p,r) \in E(G)\}$, мощность которого определяется выходной степенью вершины p; $O(p)$ — число вершин, выходящих из p.

Число цитирований, полученных статьей p, равно $I(p)$ — это число входящих в p вершин, которое является мощностью множества $C_G^-(p) = \{r \in V(G) | (r,p) \in E(G)\}$. Отсюда следует, что $|C_G^-(p)| = I(p)$ и $|C_G^+(p)| = O(p)$.

Шаг 2. Инициализация. Пусть оценки статей обозначаются вектором $\mathbf{r} = [r_1, ..., r_i, ..., r_n]$, где r_i — импакт-оценка статьи i. Более того, нормализованным ограничением является уравнение $\sum_{i=1}^{n} r_i = 1$.

Инициализированные оценки на итерации $t = 0$: $r_i^{(t=0)} = \frac{1}{n}$, $i = \overline{1,n}$.

Шаг 3. Цикл вычисления. На каждой итерации (t) оценка статьи r_i рассчитывается последовательно по двум представленным ниже формулам:

$$r_i^{(t)} = \sum_{j \in C_G^-(i)} \frac{r_j^{(t-1)}}{O(j)};$$

$$r_i^{(t)} := r_i^{(t)} + \frac{1 - \sum_{k=1}^{n} r_k^{(t)}}{n}.$$

Шаг 4. Проверка остановки цикла. Цикл останавливается, когда вектор оценок $\mathbf{r}$ сходится с заданной точностью ε:

$$\mathbf{r}^{(t)} - \mathbf{r}^{(t-1)} < \varepsilon.$$

Если это условие не удовлетворено, то процесс вычисления возвращается к шагу 3.

Шаг 5. Вычисление итоговой оценки статей. Необходимо принять фактор времени при ранжировании исследовательских работ, чтобы уменьшить смещение по сравнению с недавними работами, которые получают меньше времени для изучения, следовательно, меньше цитируются исследователями по сравнению с более давними работами. Исходя из идеи m-индекса [14], который определяется как h/y, где h — индекс Хирша, y — число лет с момента публикации первой статьи ученого, для получения конечной оценки статьи делим каждую оценку на соответствующий возраст статей (множитель 100 был принят для того, чтобы последующая визуализация была более наглядна):

$$s_i = 100 \frac{r_i}{y_i}, \quad i = \overline{1,n},$$

где s_i — конечная оценка статьи; r_i — оценка статьи после выхода из цикла; y_i — возраст статьи.

Прогнозирование тенденции научных исследований в области искусственного интеллекта

При применении предложенного алгоритма для ранжирования статей к собранным данным области "Искусственный интеллект" за период 2005—2016 получим оценки каждой статьи и отсортируем их по оценкам в порядке убывания, которые представлены на рис. 3.

Далее создаем корпус текстов, агрегируя заголовки, аннотации, авторские ключевые слова, ключевые слова плюс от топ-20 лучших статей из приведенной выше таблицы рейтинга. После этого используется функция "Создание карты на основе текстовых данных" в библиометрическом программном обеспечении VOSviewer [15], чтобы создать карту терминов совместного использования (term co-occurrence map) на основе полученного корпуса текста, а затем извлечь все термины (ключевые слова) из этой карты. В программе VOSviewer этап идентификации терминов состоит из следующих пяти этапов:

- удаление заявлений об авторских правах в аннотациях;
- обнаружение и разбиение на предложения;
- маркировка части речи (Part-of-speech tagging): с использованием алгоритма маркировки части речи, предоставляемого библиотекой Apache OpenNLP, каждому слову присваивается часть речи, такая как глагол, существительное, прилагательное, предлог и т. д.;
- идентификация словосочетания: VOSviewer-программа определяет словосочетание как последовательность из одного или нескольких последовательных слов в предложении, последнее слово которой представляет собой существительное, а каждое из остальных слов является существительным или прилагательным. Чтобы определить фразы с существительными, VOSviewer рассматривает только самые длинные фразы из существительных в предложении;
- объединение словосочетаний: объединение словосочетаний осуществляется путем удаления большинства не алфавитно-цифровых символов, удаления акцентов из символов, преобразования прописных символов в строчные и преобразования словосочетаний множественного числа в единственное число.

Описанная выше стадия идентификации термина дает набор терминов (ключевых слов), которые были идентифицированы в текстовых данных, доступных для VOSviewer. На следующем этапе, начиная с набора идентифицированных терминов, выполняется выбор терминов путем исключения терминов на основе порога частоты появления или с низким показателем релевантности, а также путем ручного исключения, а именно: оставляем только значимые термины, связанные с методом, алгоритмом, областью или разделом исследования, и исключаем вручную все общие или неинформативные термины, такие как "direction", "first step", "action" и т. д.

Суммируя оценки статей, рассчитанные по разработанному алгоритму, сформируем итоговые оценки для ключевых слов, а затем сгруппируем полученные результаты по годам. Таким образом, можно продемонстрировать множество ключевых слов по годам во временных рядах и наблюдать, какие ключевые слова сохраняют восходящий тренд. Для ясности сохраняем только довольно частые и недавние ключевые слова (присутствуют в статьях за последние минимум четыре года) и разбиваем их на четыре подгруппы, одна из которых продемонстрирована на рис. 4 (см. четвертую сторону обложки).

Затем отбираем ключевые слова, число присутствия которых в исследуемом периоде более 9 (это "ant colony", "artificial bee colony", "honey bee", "particle swarm optimization", "swarm intelligence", "tree search"). Далее для проверки и прогнозирования тенденции изменения оценок ключевых слов необходимо приспособить модель полиноминальной регрессии [16]. Результаты такого приспособления модели показаны на рис. 5 (см. четвертую сторону обложки), из которого видно, что следующие ключевые слова сохраняют восходящие тенденции: "ant colony", "artificial bee colony", "particle swarm optimization", "swarm intelligence", "tree search".

Аналогичным способом в итоге собраны 30 явно восходящих по ключевым словам исследователь-

	Title	Year	Score	Abstract	Author_keywords	Keyword_plus
2	On the performance of artificial bee colony (ab	2008	0.195	Artificial bee colony (ABC) algorit	swarm intelligence;	Differential Evo
3	A comprehensive survey: artificial bee colony	2014	0.054	Swarm intelligence (SI) is briefly d	Swarm intelligence;	Quantum Evolu
4	An artificial bee colony algorithm for the leaf-c	2009	0.043	Given an undirected, connected, w	Artificial bee colony	Abc Algorithm;
5	A survey of monte carlo tree search methods	2012	0.042	Monte Carlo tree search (MCTS) is	Artificial intelligenc	Game
6	A modified artificial bee colony (abc) algorithm	2011	0.028	Artificial Bee Colony (ABC) algorit	Swarm intelligence;	Evolutionary Al
7	The 2014 general video game playing competit	2016	0.027	This paper presents the framewor	Competitions; evolu	
8	The best-so-far selection in artificial bee colon	2011	0.026	The Artificial Bee Colony (ABC) al	Artificial Bee Colony	Image Registrat
9	A granular intrusion detection system using ro	2016	0.024	Security in computer networks is a	Intrusion detection	Fuzzy; Decision
10	Affect control processes: intelligent affective in	2016	0.024	This paper describes a novel meth	Affect; Emotion; Soc	Emotion; Identi
11	Classification of dna microarrays using artificia	2016	0.024	DNA microarray is an efficient nev	DNA microarrays; A	Particle Swarm
12	Computational interpretation of comic scenes	2016	0.024	Understanding intellectual produc	Computational mod	
13	Dynamic modeling based on a temporal-causal	2016	0.024	This paper presents a dynamic mo	Modeling; Dynamic;	Recurrent Neur
14	Future progress in artificial intelligence: a surv	2016	0.024	There is, in some quarters, concer	Artificial intelligenc	Ai
15	Optimization of electricity markets participati	2016	0.024	The electricity markets environm	Artificial intelligenc	
16	When thinking never comes to a halt: using for	2016	0.024	The recognition that human minds	Cognitive systems; C	Approximability
17	Universal intelligence: a definition of machine i	2007	0.022	A fundamental problem in artificia	AIXI; complexity the	Science
18	A universal measure of intelligence for artificia	2005	0.022			
19	Artificial and natural intelligence integration	2015	0.021	The large amount of data generate	Data mining; visuali	
20	The coming of age of artificial intelligence in m	2009	0.019	This paper is based on a panel disc	Artificial intelligenc	Decision-Suppo
21	The primary language of ancient battles	2011	0.018	Linguistic Geometry (LG) is a type	Linguistic Geometry	

Рис. 3. Топ-20 лучших (влиятельных) статей, полученных по предлагаемому алгоритму

ских тенденций, и это также прогнозируемые тенденции после 2016 г.: "artificial neural network", "computational intelligence", "diagnosis", "disease", "emotion", "genetic algorithm", "machine learning", "ant colony", "cognitive science", "evolutionary algorithm", "evolutionary computation", "fuzzy set", "game theory", "granular computing", "human brain", "intrusion detection system", "machine intelligence", "markov decision process", "medicine", "multilayer perceptron", "particle swarm optimization", "pattern recognition", "radial basis function", "support vector machine", "swarm intelligence", "cognitive psychology", "health care", "rough set theory", "tree search", "fuzzy cognitive map", "pattern classification", "recurrent neural network", "artificial bee colony", "cognitive system", "differential evolution", "empirical evidence", "honey bee", "human reasoning", "network intrusion detection", "sociology", "tumor", "computer go", "electricity market", "gene-expression data".

Среди них такие ключевые слова, как "artificial neural network", "computational intelligence", "disease", "emotion", "genetic algorithm", "machine learning" показывают высокий темп эволюции, и можно логично предположить, что эти тематические исследования все еще сильно будут проявляться в последующие годы.

Экспериментальная оценка предложенного метода

Для проверки качества метода результаты оцениваются на основе статистического анализа лексики 3211 собранных статей из библиографической базы данных платформы WoS при поиске с запросом "Artificial Intelligence" и фильтром категории "Computer science Artificial Intelligence" за период 2017—2019 гг.

В первую очередь создаем корпус текста, комбинируя заголовки, аннотации, авторские ключевые слова, ключевые слова плюс всех статей из изложенного набора данных. Затем метод "Создание карты на основе текстовых данных" снова используется в библиометрической программе VOSviewer для создания карты терминов совместного использования на основе полученного корпуса текста, тогда извлекаются все ключевые слова из этой карты совместного использования с порогом присутствия 50. Аналогично оставляем только значимые термины, касающиеся методов, алгоритмов, области или разделов исследования, и исключаем вручную общие и неинформативные термины, такие как "argument", "distance", "thing", и т. д.

В результате получается набор топ-50 используемых ключевых слов в порядке убывания частоты с соответствующими показателями релевантности, которые представлены на рис. 6, *а, б*. Следует отметить, что термины с высоким показателем релевантности представляют конкретные темы в тексте, а термины с низкой релевантностью носят общий характер и не являются репрезентативными для какой-либо конкретной темы. После исключе-

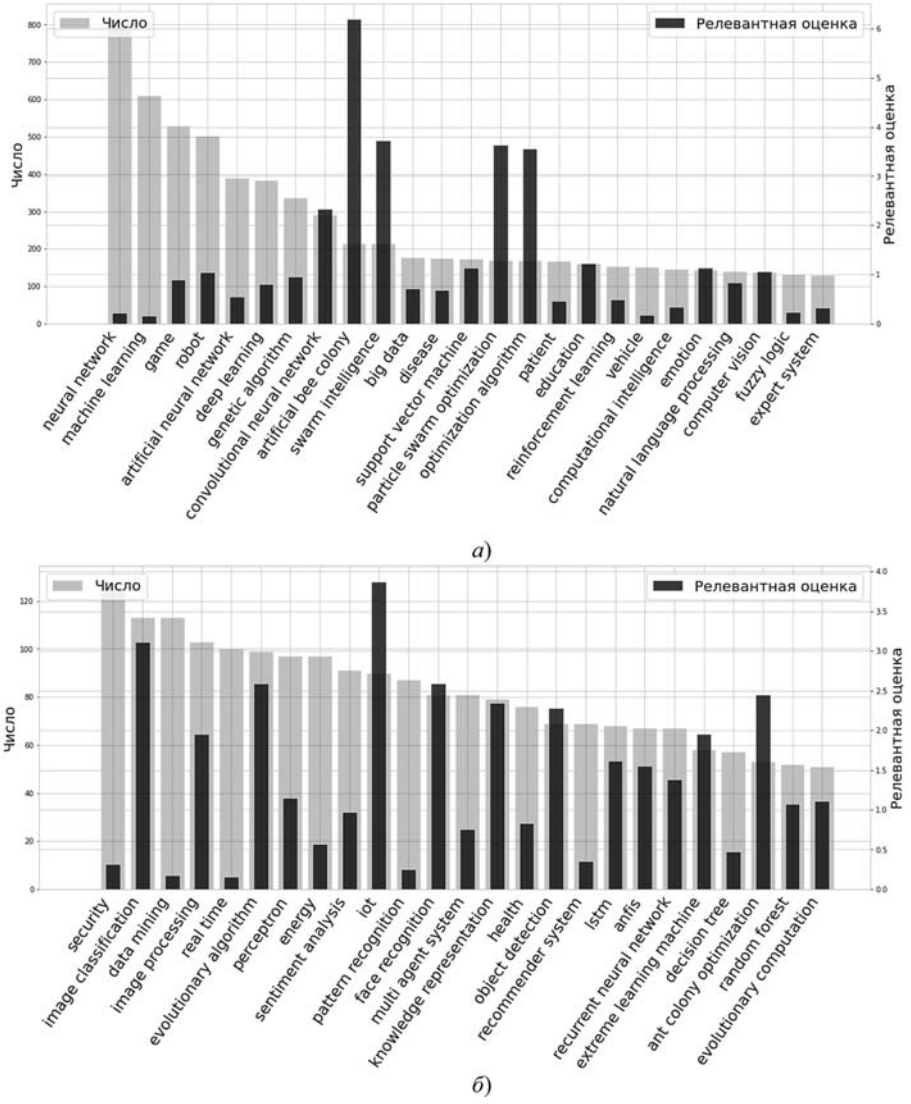

Рис. 6. Топ-50 (*а* — от 1 до 25, *б* — от 26 до 50) самых частых ключевых слов за период 2017—2019 гг.

ния терминов с низким показателем релевантности общие термины отфильтровываются, и фокус смещается на более конкретные и информативные термины [17].

Эти термины считаются принадлежащими фокусным исследованиям в период 2017—2019 гг. Необходимо определить точность нашего прогноза, сравнив число предсказанных ключевых слов с фактическим набором ключевых слов за выбранный период. Кроме того, сравнение ключевых слов выполняется не только по явному совпадению, но и по приблизительному совпадению (например, "game theory" и "game", "fuzzy cognitive map" и "fuzzy logic" считаются сходными ключевыми словами). В итоге обнаружено 22 ключевых слова (из всего 30 прогнозируемых ключевых слов восходящего тренда) в наборе 50 истинных ключевых слов фокусных исследований. Таким образом, точность прогноза составляет около 73,33 % (22/30). Однако полученную точность можно увеличить при расширении промежутка времени для оценки.

Заключение

Из-за неуклонного роста числа научных публикаций трудно вести обзор структуры, содержания и динамичного развития своей области науки и, тем более, связанных научных областей. Электронные ресурсы, в том числе базы данных публикаций, являясь реальным отображением процессов, происходящих в науке, содержат элементы, анализ которых позволяет оценивать и прогнозировать ее развитие.

В данной статье предложен общий подход к анализу и прогнозированию тематической эволюции области исследований ИИ путем определения ключевых слов восходящего тренда. Предложенный подход применен для анализа тематической эволюции исследований в области ИИ за период 2005—2016 гг. из базы данных платформы WoS. Оценка метода реализована путем обнаружения предсказанных ключевых слов в наборе истинных фокусных исследований ИИ за период 2017—2019 гг. Полученная точность прогноза составляет 73,33 %.

В будущем планируется повысить точность прогнозирования за счет улучшения алгоритма ранжирования статей и учета большего числа свойств, таких как факторы влияния авторов и журналов или ранжирование трудов конференций.

Список литературы

1. **Iqbal W., Qadir J., Tyson G., Mian A. N., Saeed-ul H., Crowcroft J.** A bibliometric analysis of publications in computer networking research // Scientometrics. 2019. Vol. 119, N. 2. P. 1121—1155.

2. **Merigo J. M., Pedrycz W., Weber R., de la Sotta C.** Fifty years of Information Sciences: A bibliometric overview // Information Sciences. 2018. Vol. 432. P. 245—268.

3. **Wang Q.** A Bibliometric Model for Identifying Emerging Research Topics // Journal of the Association for Information Science and Technology. 2018. Vol. 69, N. 2. P. 290—304.

4. **Robinson-Garcia N., Sugimoto C. R., Murray D., Yegros-Yegros A., Lariyiere V., Costas R.** The many faces of mobility: Using bibliometric data to measure the movement of scientists // Journal of Informetrics. 2019. Vol. 13, N. 1. P. 50—63.

5. **Cobo M. J., Jurgens B., Herrero-Solana V., Martinez M. A., Herrera-Viedma E.** Industry 4.0: a perspective based on bibliometric analysis // 6th International Conference on Information Technology and Quantitative Management. 2018. Vol. 139. P. 364—371.

6. **Нгуен Т. В., Кравец А. Г.** Метод прогноза исследовательских тенденций на основе алгоритма ранжирования статей // Математические методы в технике и технологиях — ММТТ: сб. тр. XXXIII междунар. науч. конф. ММТТ—33 (г. Казань — г. Калининград — г. Минск (Беларусь) — г. Саратов). В 12 т. Т. 8 / Под общ. ред. А. А. Большакова; КНИТУ, КГТУ, СГТУ им. Гагарина Ю. А., БНТУ. Санкт-Петербург: Изд-во Политехнического ун-та, 2020. С. 123—127.

7. **Нгуен Т. В., Кравец А. Г.** Автоматизация сбора информации из открытых интернет источников // Математические методы в технике и технологиях (ММТТ—32): сб. тр. XXXII междунар. науч. конф. В 12 т. Т. 5 / Под общ. ред. А. А. Большакова; Санкт-Петербургский гос. технологический ин-т (техн. ун-т), Санкт-Петербургский ин-т информатики и автоматизации РАН, Санкт-Петербургский политехнический ун-т Петра Великого, Саратовский гос. техн. ун-т им. Гагарина Ю. А. [и др.]. Санкт-Петербург, 2019. С. 131—135.

8. **Нгуен Т. В., Кравец А. Г.** Алгоритм работы веб-краулера для решения задачи сбора данных из открытых интернет источников // Известия Санкт-Петербургского гос. технологического ин-та (технического ун-та). 2019. № 51 (77). С. 115—119.

9. **Moral-Munoz J. A., Arroyo-Morales M., Piper B. F., Cuesta-Vargas A. I., Diaz-Rodriguez L., Cho W. C. S., Herrera-Viedma E., Cobo M. J.** Thematic Trends in Complementary and Alternative Medicine Applied in Cancer-Related Symptoms // Journal of Data and Information Science. 2018. Vol. 3, N. 2. P. 1—19.

10. **Aria M., Cuccurullo C.** bibliometrix: An R-tool for comprehensive science mapping analysis // Journal of Informetrics. 2017. Vol. 11, N. 4. P. 959—975.

11. **Sohrabi B., Vanani I. R., Jalali S. M. J., Abedin E.** Evaluation of Research Trends in Knowledge Management: A Hybrid Analysis through Burst Detection and Text Clustering // Journal of Information & Knowledge Management. 2019. Vol. 18, N. 4. P. 27.

12. **Нгуен Т. В., Кравец А. Г.** Analyzing Recent Research Trends of Computer Science from Academic Open-access Digital Library // 8th International Conference on System Modeling and Advancement in Research Trends (SMART—2019, IEEE Conference ID: 46866) (22nd—23rd November, 2019): Proceedings / Eds.: A. Kr. Saxena, D. Parygin, D. Ather, V. Yadav; College of Computing Sciences & Information Technology, Teerthanker Mahaveer University (Moradabad, India), IEEE UP Section. New Delhi (India), 2019. P. 31—36.

13. **Klavans R., Boyack K. W.** Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? // Journal of the Association for Information Science and Technology. 2017. Vol. 68, N. 4. P. 984—998.

14. **Roldan-Valadez E., Salazar-Ruiz S. Y., Ibarra-Contreras R., Rios C.** Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal

Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics // Irish Journal of Medical Science. 2019. Vol. 188, N. 3. P. 939—951.

15. **Van Eck N. J., Waltman L.** Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics. 2010. Vol. 84, N. 2. P. 523—538.

16. **Polynomial** Regression. [Электронный ресурс]. URL: https://towardsdatascience.com/polynomial-regression-bbe8b9d97491/ (дата обращения: 20.11.2020).

17. **Van Eck N. J., Waltman L.** Citation-based clustering of publications using CitNetExplorer and VOSviewer // Scientometrics. 2017. Vol. 111, N. 2. P. 1053—1070.

T. V. Nguyen, Postgraduate Student, e-mail: vietqn1987@gmail.com,
A. G. Kravets, Dr. Tech. Sciences, Prof., e-mail: agk@gde.ru,
Volgograd State Technical University, Volgograd, Russian Federation

Evaluation and Prediction of Trends in the Development of Scientific Research Based on Bibliometric Analysis of Publications

The article proposes an approach to analyzing and predicting the thematic evolution of research by identifying an upward trend in keywords. Statistical analysis of the vocabulary of publications allows us to trace the depth of penetration of new ideas and methods, which can be set by the frequency of occurrence of words encoding whole concepts. The article presents a developed method for analyzing research trends and an article ranking algorithm based on the structure of a direct citation network. Data for the study was extracted from the Web of Science Core Collection, 6696 publications were collected for the experiment over the period 2005—2016 in the field of artificial intelligence. To evaluate the proposed method, 3211 publications were collected from 2017 to 2019. As a result, the method was evaluated by checking the presence of predicted keywords in the set of the most frequent terms for the period 2017—2019 and provided an accuracy of 73.33 %.

Keywords: *trend forecasting, thematic evolution, bibliometric analysis, citation network, adjacency matrix, VOSviewer, Biblioshiny, article ranking, paper ranking, artificial intelligence, Web of Science database*

Acknowledgments: This research was supported by the Russian Fung of Basic Research (grants No. 19-07-001200, No. 20-37-90092).

DOI: 10.17587/it.27.195-201

References

1. **Iqbal W., Qadir J., Tyson G., Mian A. N., Saeed-ul H., Crowcroft J.** A bibliometric analysis of publications in computer networking research, *Scientometrics,* 2019, vol. 119, no. 2, pp. 1121—1155.

2. **Merigo J. M., Pedrycz W., Weber R., de la Sotta C.** Fifty years of Information Sciences: A bibliometric overview, *Information Sciences*, 2018, vol. 432, pp. 245—268.

3. **Wang Q.** A Bibliometric Model for Identifying Emerging Research Topics, *Journal of the Association for Information Science and Technology*, 2018, vol. 69, no 2, pp. 290—304.

4. **Robinson-Garcia N., Sugimoto C. R., Murray D., Yegros-Yegros A., Lariyiere V., Costas R.** The many faces of mobility: Using bibliometric data to measure the movement of scientists, *Journal of Informetrics*, 2019, vol. 13, no. 1, pp. 50—63.

5. **Cobo M. J., Jurgens B., Herrero-Solana V., Martinez M. A., Herrera-Viedma E.** Industry 4.0: a perspective based on bibliometric analysis, *6th International Conference on Information Technology and Quantitative Management*, 2018, vol. 139, pp. 364—371.

6. **Nguyen T. V., Kravets A. G.** Research trend forecasting method based on article ranking algorithm, *Matematicheskie metody v tehnike i tehnologijah — MMTT: sb. tr. XXXIII mezhdunar. nauch. konf. MMTT—33* (g. Kazan — g. Kaliningrad — g. Minsk (Belarus) — g. Saratov), vol. 8, Sankt-Peterburg, Publishing house of Polytech. University, 2020, pp. 123—127 (in Russian).

7. **Nguyen T. V., Kravets A. G.** Automation of collection of information from open Internet sources, *Matematicheskie metody v tehnike i tehnologijah (MMTT—32): sb. tr. XXXII mezhdunar. nauch. konf.*, vol. 5, Sankt-Peterburg, 2019, pp. 131—135 (in Russian).

8. **Nguyen T. V., Kravets A. G.** Algorithm of a web crawler to solve the problem of collecting data from open Internet sources, *Izvestija Sankt-Peterburgskogo gos. tehnologicheskogo in-ta (tehnicheskogo un-ta)*, 2019, no. 51 (77), pp. 115—119 (in Russian).

9. **Moral-Munoz J. A., Arroyo-Morales M., Piper B. F., Cuesta-Vargas A. I., Diaz-Rodriguez L., Cho W. C. S., Herrera-Viedma E., Cobo M. J.** Thematic Trends in Complementary and Alternative Medicine Applied in Cancer-Related Symptoms, *Journal of Data and Information Science*, 2018, vol. 3, no. 2, pp. 1—19.

10. **Aria M., Cuccurullo C.** bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, 2017, vol. 11, no. 4, pp. 959—975.

11. **Sohrabi B., Vanani I. R., Jalali S. M. J., Abedin E.** Evaluation of Research Trends in Knowledge Management: A Hybrid Analysis through Burst Detection and Text Clustering, *Journal of Information & Knowledge Management*, 2019, vol. 18, no. 4, pp. 27.

12. **Nguyen T. V., Kravets A. G.** Analyzing Recent Research Trends of Computer Science from Academic Open-access Digital Library, *8th International Conference on System Modeling and Advancement in Research Trends (SMART—2019, IEEE Conference ID: 46866) (22nd—23rd November, 2019): Proceedings,* College of Computing Sciences & Information Technology, Teerthanker Mahaveer University (Moradabad, India), IEEE UP Section, New Delhi (India), 2019, pp. 31—36.

13. **Klavans R., Boyack K. W.** Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?, *Journal of the Association for Information Science and Technology*, 2017, vol. 68, no. 4, pp. 984—998.

14. **Roldan-Valadez E., Salazar-Ruiz S. Y., Ibarra-Contreras R., Rios C.** Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics, *Irish Journal of Medical Science*, 2019, vol. 188, no. 3, pp. 939—951.

15. **Van Eck N. J., Waltman L.** Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, 2010, vol. 84, no. 2, pp. 523—538.

16. **Polynomial** Regression, available at: https://towardsdatascience.com/polynomial-regression-bbe8b9d97491/ (date of access: 20.11.2020).

17. **Van Eck N. J., Waltman L.** Citation-based clustering of publications using CitNetExplorer and VOSviewer, *Scientometrics,* 2017, vol. 111, no. 2, pp. 1053—1070.