Новый метод прогнозирования технологических трендов на основе анализа научных статей и патентов

Нгуен Тхань Вьет, А. Г. Кравец

***Аннотация*—Для достижения конкурентоспособности в условиях быстро меняющейся науки, важно следить за развитием существующих технологий и открывать новые и перспективные технологии. Фирмам необходимо разработать стратегию развития технологий с помощью прогноза технологических трендов, чтобы получить конкурентное преимущество при использовании ограниченных ресурсов. С другой стороны, в настоящее время количество научных статей, патентов и других разнородных данных растет быстрыми темпами, и становится невозможными оставаться в курсе всего, что публикуется. Однако, несмотря на все усилия, не были найдены методологические и технологические результаты, дающие возможность осуществить создание моделей и методов для целостного восприятия вычислительной системой разнородной информации – научных публикаций и патентов, которая содержится в открытых источниках. При этом большинство существующих исследований предназначено для анализа и раннее выявления новых технологий или мониторинга трендов в одних конкретных технологических отраслях, не рассматривая решение задачи прогноза множества различных технологических трендов. Кроме того, точность оценки предложенных методов в существующих исследованиях либо довольно не высокая (максимальная метрика F1 оценки точности прогноза ~ 74%), либо отсутствует (не была осуществлена оценка качества метода). Таким образом, в данной статье предложен новый метод для анализа и прогнозирования технологических тенденций на основе обработки разнородных данных (научные статьи, патенты) из открытых источников путем разработки алгоритма извлечения значимых ключевых слов и методов создания матриц совместного появления элементов (ключевых слов, кодов CPC).**

***Ключевые слова*—технологический прогноз, извлечение ключевых слов, Гауссовский процесс, матрица совпадения, кластеризация сети совпадения, VOSviewer.**

I. Введение

В последние годы быстро развиваются и появляются новые технологии с прорывными характеристиками, которые не только изменило существующие отрасли, но и привело к созданию новых отраслей, оказавших значительное влияние на социально-экономическую структуру [1,2]. Многие руководители и исследователи осознают важность понимания пути появления и определения будущих тенденций развития новых технологий для конкурентоспособности и устойчивого развития своей организации [3–5]. В то же время каждая появляющаяся технология может предоставить множество возможностей для бизнеса и привлечь венчурные инвестиции. По этой причине, с учетом множества новых технологий, правильная оценка и определение тенденции технологий как можно раньше имеет решающее значение для любого подразделения управления и промышленного бизнеса.

Изучение технически подходящих исторических знаний позволяет выяснить, как на вариации технологических достижений влияют прошлые и существующие сдвиги в связанных технологиях. Важным аспектом выявления технологических тенденций являются данные [4,6,7]. Новости, социальные сети, блоги и отчеты компаний раскрывают в основном технологии, которые недавно достигли пика своего развития или были доступны на рынке до сих пор. Однако о преждевременных технологических тенденциях обычно сначала сообщают в научных публикациях. Таким образом, эти данные являются важнейшими источниками знаний для первичных тенденций и знаков [4,8].

Известно, патенты предоставляют источник современной и надежной информации для раскрытия технологической информации и развития. Благодаря анализу технологической информации, имеющейся в патентах, можно лучше выявить, понять путь технологической эволюции, и определить тенденции развития технологии с помощью экспертов в данной области. Следовательно, исследователи начали использовать патентные данные для анализа и изучения технологических тенденций [9–11]. Однако, из-за неопределенности и неоднозначности, присущей этим появляющимся технологиям, исследования тенденций развития были сосредоточены на патентных данных, которые не только чувствительны ко времени, но и имеют

Статья получена 19 июля 2022. Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 19-07-01200 и 20-37-90092.

Нгуен Тхань Вьет – Волгоградский государственный технический университет, г. Волгоград; Фам Ван Донг университет, г. Куанг Нгай, Вьетнам (email: vietqn1987@gmail.com).

А. Г. Кравец – Волгоградский государственный технический университет, г. Волгоград; Государственный университет «Дубна», г. Дубна, Московская область (email: agk@gde.ru).

ограниченные перспективы многогранного проявления новых технологий [12]. Поэтому, анализа только патентов недостаточно для полного понимания эволюционного пути и тенденций развития новых технологий.

Некоторые ученые отмечали, что различные типы источников информации предоставляют разнообразные знания о пути эволюции технологического развития, и комплексное использование этих источников данных совершенно даст более полную картину технологической тенденции. Однако, несмотря на все усилия, не были найдены методологические и технологические результаты, дающие возможность осуществить создание моделей и методов для целостного восприятия вычислительной системой разнородной информации – научных публикаций и патентов, которая содержится в открытых источниках. При этом большинство существующих исследований предназначено для анализа и раннее выявления новых технологий или мониторинга трендов в одних конкретных технологических отраслях, не рассматривая решение задачи прогноза множества различных технологических трендов. Кроме того, точность оценки предложенных методов в существующих исследованиях либо довольно не высокая (максимальные метрики F1 ~ 74%, precision ~ 76%), либо отсутствует (не была осуществлена оценка качества метода).

Таким образом, в данной статье предложен новый метод для анализа и прогнозирования технологических тенденций (АПТТ) на основе обработки разнородных данных (научные статьи, патенты) из открытых источников.

II. Предложенный метод

Технологический тренд оценивается как неуклонно развивающаяся практическая технологическая область с определенной закономерностью, которая существует в течение определенного периода времени [13]. В последнее время было предложено множество подходов для изучения закономерностей, анализа и прогнозирования технологических тенденций. Одним из таких подходов, принятым различными учеными, является исследование модели взаимоотношений между наукой и технологией, которая существует уже много десятилетий и продолжает оставаться предметом длительных дискуссий в академическом мире. Большое количество эмпирических результатов доказало, что наука и технология взаимодействуют по-разному, более того, взаимозависимости между ними в последние годы усиливаются [14].

В частности, фундаментальные научные исследования обеспечивают основу для технического прогресса, поскольку в инновационной модели наука рассматривается как источник технологий. Следовательно, существенные и передовые знания о технологических разработках могут быть получены путем анализа научных публикаций и патентов [15].

С другой стороны, за последние 20 лет количество патентных ссылок на научные статьи быстро увеличивается. Это указывает на все более сильный поток знаний от академической науки к промышленным инновациям. Многие исследования показали, что склонность патентов к цитированию академических публикаций в последнее время возросла, даже с учетом изменений в объеме и распределении патентов по областям [16].

В то же время связи между наукой и технологией широко изучались с использованием ссылок на непатентную литературу или сравнений авторов и изобретателей. В статье [17] было принято латентное распределение Дирихле (latent Dirichlet distribution – LDA) для создания тематических связей между публикациями и патентами на основе семантического содержания документов. Этот подход позволяет обнаруживать тематические совпадения между патентными и научными публикациями, выделяя тематические области, используемые для исследований и технологий.

Кроме того, в исследовании [18] Ли и др. предложили фреймворк, в котором использовались научные статьи и патенты в качестве источников данных и интегрированный анализ цитирования с анализом текста, чтобы проанализировать эволюционный путь технологии наногенераторов, а затем предсказать ее тенденцию. Авторы начали с применения анализа цитирования для изучения технических знаний, отраженных в научных статьях, и наблюдения за эволюционным путем технологии наногенераторов. Кроме того, метод тематической модели иерархического процесса Дирихле (Hierarchical Dirichlet Process – HDP) был применен для выявления технических тем в собранных научных статьях. Аналогично метод HDP был также принят для изучения технических тем в собранных патентах. Наконец, авторы проанализировали отставание (delay) между наукой и технологией, а затем объединили их с экспертными знаниями и эволюционным путем технологии наногенератора для прогноза тенденции ее развития.

Из этого можно сделать вывод, что жизненный цикл технологии начинается с научных публикаций, за которыми следуют заявки на патенты, а затем другие технологические новости. Следовательно, сначала анализируются научные публикации для выявления основных технологических направлений, начиная с известной наукометрической базы данных Web of Science (WoS), а затем патентные заявки будут использованы для АПТТ.

На рисунке 1 показаны общие процедуры предложенного метода и функциональный анализ автоматизированной системы для АПТТ по методологии IDEF0. Иными словами, на этом рисунке представлена декомпозиция первого уровня основного

функционального блока A_0 разработанной системы.

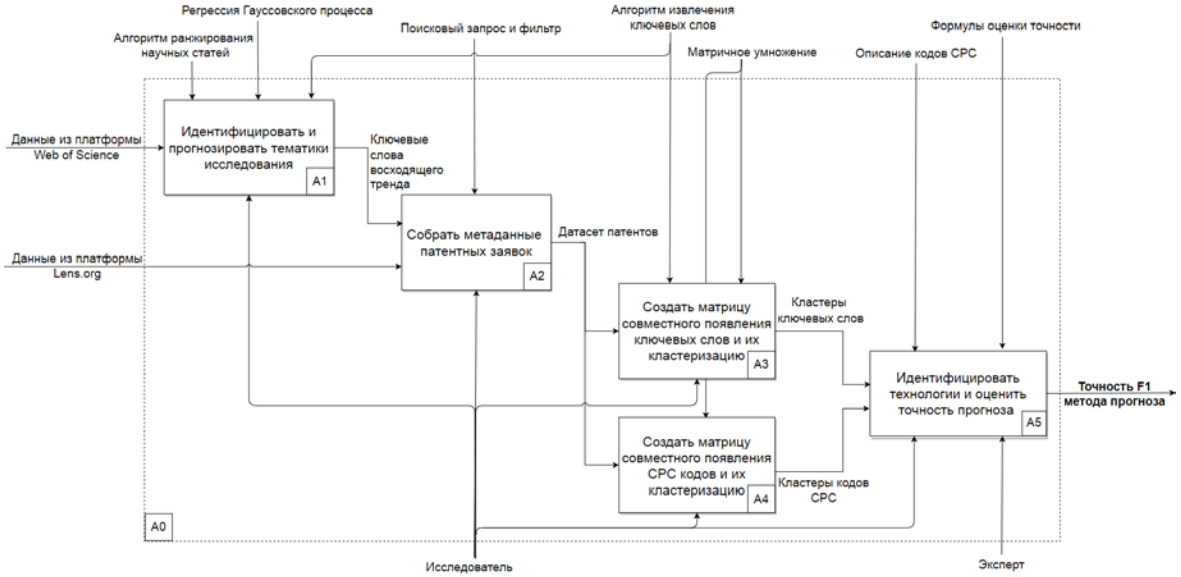


Рис. 1 – Функциональный анализ автоматизированной системы для АПТТ

Программный код системы написан на языке Python, с поддержкой популярных библиотек машинного обучения, таких как Scikit-learn, Scipy, Pandas, Numpy, и др. Все эмпирические запуски (empirical running) проводились в интерактивной среде Google Colab.

III. Анализ научных статей

В нашей недавней статье [19] предложен алгоритм для анализа тематической эволюции исследований в области Искусственного Интеллекта (ИИ) за период 2005-2016гг из наукометрических данных платформы WoS. Следует отметить, что обоснование выбора области ИИ в качестве главной целевой области для анализа было приведено заранее в нашей публикации [8]. В результате можно выявить самые влиятельные научные статьи (топ-20 лучших статей). После этого разработан новый метод извлечения самых значимых ключевых слов из каждого текста топ-20 статей, который состоит из 4 шагов: извлечение всех возможных ключевых слов, преобразование извлеченных ключевых слов в векторы встраивания (Embedding vectors), ранжирование извлеченных ключевых слов по релевантности и разнообразию, и извлечение значимых ключевых слов из документов.

A. Метод извлечения самых значимых ключевых слов

Шаг 1. Извлечение всех возможных ключевых слов:

Этот шаг включает в себя 3 этапа:

– тегирование частей речи (Part-of-speech – POS tagging) с использованием библиотеки spaCy [20];

– идентификация ключевых слов: максимально длинная последовательность именных словосочетаний из нескольких (≥ 2) последовательных слов в предложении, при которой последнее слово в последовательности является существительным, а каждое из остальных слов является либо существительным, либо прилагательным;

– устранение форм множественного числа с помощью платформы Natural Language Toolkit (NLTK) [21].

Псевдокод алгоритма извлечения всех ключевых слов представлен на рисунке 2.

```
    Алгоритм извлечения всех ключевых слов из текста
ВХОД: Текст и разрешенные теги части речи allowed_postags:
ADJ (прилагательное), NOUN (существительное), PROPN (местоимение)
ВЫХОД: Список всех именных словосочетаний в форме единственного числа

masked_words = []
while word in document do
    get POS tag of word
    if word POS tag is in allowed_postags or word == '-' do
        append word to masked_words
    else if previous word or next word == '-' do
        append word to masked_words
    else do
        append '.' to masked_words
    end
end
extracted_phrases = []
while index < masked_words length do
    get token word
    start_index = index
    if word != '.' doc
        start_index = last index that word not in {'.', 'ADJ'}
        if end_index > start_index + 1
            get phrase between (index and start_index)
            append phrase to extracted_phrases
        end
    end
    index = start_index + 1
end
lemmatized_phrases = []
while phrase in extracted_phrases do
    while token in phrase do
        get lemmatized word
    end
    join lemmatized words into phrase
    append pharase to lemmatized_phrases
end
```

Рис. 2 – Псевдокод алгоритма извлечения всех ключевых слов из текста

Шаг 2. Преобразование извлеченных ключевых слов в векторы встраивания (Embedding vectors):

Фреймворк Python SentenceTransformers [22] используется для преобразования извлеченных ключевых слов в векторы встраивания 768 размерностей.

Шаг 3. Ранжирование извлеченных ключевых слов по

релевантности и разнообразию:

В этом шаге используется метод Максимальной предельной релевантности (Maximal marginal relevance), предложенный автором Maarten Grootendorst [23]. Идея этого метода заключается в минимизации избыточности (появление похожих ключевых слов) и максимальном разнообразии результатов в задачах реферирования текста (уменьшение данных при увеличении полезных знаний). Конкретно метод начинается с выбора ключевого слова, наиболее похожего на содержание документа. Затем он итеративно выбирает новые многообещающие фразы, которые одновременно похожи на документ и отличаются от ранее выбранных ключевых слов.

Шаг 4. Извлечение значимых ключевых слов из документов:

При этом выбираются 10 наиболее значимых ключевых слов из каждого текста статьи, состоящего из Заголовка, Аннотации, Авторских ключевых слов и Ключевых слов плюс. В окончательном списке после устранения дубликатов получены 163 уникальных ключевых слова из топ-20 лучших статей. Далее, начиная с набора идентифицированных фраз, выполняется выбор значимых ключевых слов, связанных с методом, алгоритмом, областью или разделом исследования, и исключаются все общие или неинформативные фразы, такие как «new method», «data feature», «few decade», и т.д. Кроме того, некоторые общие по значению ключевые слова были группированы для дальнейшего точного вычисления. Например, ключевые слова «artificial bee colony», «bee colony algorithm», «abc algorithm» будут считаться одной фразой «artificial bee colony algorithm». В результате получен список из 117 значимых уникальных ключевых слов.

На следующем этапе, для каждого ключевого слова суммируя рассчитанные по предложенному алгоритму импакт-оценки статей, которые содержат рассмотренное ключевое слово. Далее вычислим итоговые полученные оценки по годам. Таким образом, можно продемонстрировать множество ключевых слов во временных рядах по годам и наблюдать, какие ключевые слова сохраняют восходящий тренд. Для визуализации рассматриваем 47 довольно частых ключевых слов (присутствуют в статьях за минимум 3 года), первые 16 из 47 выбранных ключевых слов продемонстрированы на рисунке 3.

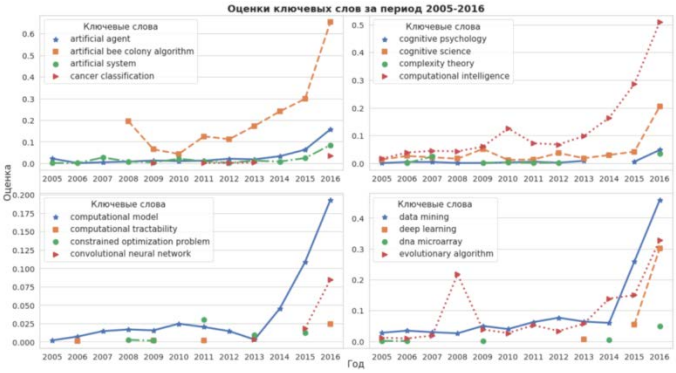


Рис. 3 – Визуализация импакт-оценки ключевых слов по годам

Затем для прогнозирования тенденции импакт-оценки ключевых слов были выбраны 42 из 47 ключевых слов, которые имеют восходящий тренд. При осуществлении прогноза используется метод регрессии Гауссовского процесса, который будет описан подробнее в последующем подразделе.

B. Метод регрессии Гауссовского процесса

Гауссовский процесс (GP) [24] – это случайный процесс, в котором любой точке $\boldsymbol{x} \in \mathbb{R}^d$ (d – размерность) ставится в соответствие случайная величина $f(\boldsymbol{x})$, а совместное распределение конечного числа этих переменных $p(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N))$ является Гауссовским.

$$p(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \mathbf{K}) \quad (1)$$

где $\boldsymbol{f} = \big(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)\big), \boldsymbol{\mu} = (m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_N))$, $\mathrm{m}(\cdot)$ представляет собой среднее значение функции. $K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\kappa(\cdot)$ – положительная определенная ядерная или ковариационная функция.

Таким образом, Гауссовский процесс – это распределение по функциям, форма (гладкость) которых определяется матрицей $\boldsymbol{K}$. Если точки $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$ рассматриваются ядерной функцией как подобные, то значения функций в этих точках $f(\boldsymbol{x}_i)$ и $f\big(\boldsymbol{x}_j\big)$ также можно ожидать, что они будут похожи.

Учитывая обучающий набор данных без шума (noise-free) значениями функции $\boldsymbol{f}(\boldsymbol{X})$, априорная GP может быть преобразована в апостериорную GP $p(\boldsymbol{f}_*|\boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f})$, которую затем можно использовать для прогнозирования $\boldsymbol{f}_*$ на новых входных данных $\boldsymbol{X}_*$.

По определению GP совместное распределение наблюдаемых значений $\boldsymbol{f}$ и прогнозов $\boldsymbol{f}_*$ снова является Гауссовским, которое можно представлять следующим образом:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\boldsymbol{T}} & \boldsymbol{K}_{**} \end{pmatrix}\right) \quad (2)$$

где $\boldsymbol{K}_* = \kappa(\boldsymbol{X}, \boldsymbol{X}_*)$, $\boldsymbol{K}_*^{\boldsymbol{T}}$ есть транспонированная матрица относительно $\boldsymbol{K}_*$, $\boldsymbol{K}_{**} = \kappa(\boldsymbol{X}_*, \boldsymbol{X}_*)$, N – количество выборок обучения, N_* – количество выборок для прогноза. Тогда размерность матриц $\boldsymbol{K} \in \mathbb{R}^{N \times N}, \boldsymbol{K}_* \in \mathbb{R}^{N \times N_*}, \boldsymbol{K}_{**} \in \mathbb{R}^{N_* \times N_*}$. Используя стандартные правила формирования Гауссовых функций, прогностическое распределение

определяется выражением:

$$p(\boldsymbol{f}_*|\boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \quad (3)$$

$$\boldsymbol{\mu}_* = \boldsymbol{K}_*^T \boldsymbol{K}^{-1} \boldsymbol{f} \quad (4)$$

$$\boldsymbol{\Sigma}_* = \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}^{-1} \boldsymbol{K}_* \quad (5)$$

Следует отметить, что в данной работе используется квадратичная экспоненциальная ядерная функция (радиальная базисная функция ядра – RBF kernel/Gaussian kernel):

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right) \quad (6)$$

Параметр длины l управляет гладкостью функции. Более высокие значения l приводят к более гладким функциям, следовательно, к более грубым приближениям обучающих данных. Более низкие значения l делают функции более волнистыми с широкими областями неопределенности между точками обучающих данных. Тем времени параметр σ_f управляет вертикальным изменением функций, взятых из GP. Это будет видно по широким областям неопределенности за пределами области обучающих данных. Оптимальные значения для этих параметров могут быть оценены путем максимизации логарифмической предельной вероятности, которая определяется выражением:

$$\log p(\boldsymbol{f}|\boldsymbol{X}) = \log \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}) = -\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{N}{2}\log(2\pi) \quad (7)$$

C. Прогноз ключевых слов восходящего тренда

При применении вышеописанного метода регрессии Гауссовского процесса получены результаты приспособления и прогноза импакт-оценки выбранных ключевых слов, визуализация которых показана на рисунке 4 (представлены 6 из 42 ключевых слов).

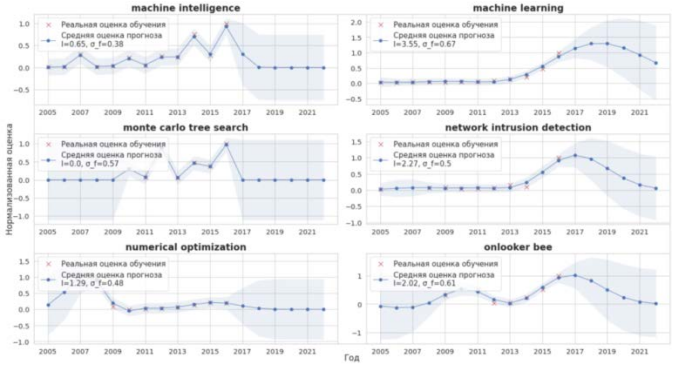


Рис. 4 – Результаты приспособления и прогноза импакт-оценки с помощью метода регрессии GP

Из рисунков видно, что следующие ключевые слова по прогнозу сохраняют восходящие тенденции с 2017 года: «machine learning», «network intrusion detection», «onlooker bee». Однако неинформативное ключевое слово «onlooker bee» будет исключено от окончательного прогноза из за общего значения.

Аналогичным способом выполнена визуализация результатов приспособления и прогноза импакт-оценки для всех остальных ключевых слов. В итоге собраны 20 явно восходящих по ключевым словам исследовательских тенденций, и это также прогнозируемые тенденции в периоде 2017-2019. Эти ключевые слова представляют собой: «artificial bee colony algorithm», «cancer classification», «cognitive science», «computational intelligence», «convolutional neural network», «data mining», «deep learning», «evolutionary algorithm», «fuzzy cognitive map», «game theory», «genetic algorithm», «intelligent tutoring system», «machine learning», «network intrusion detection», «particle swarm optimization», «pattern recognition problem», «radial basis function neural network», «recurrent neural network», «rough set theory», и «video game».

IV. Анализ патентных заявок

The Lens – всемирный открытый бесплатный полнотекстовый ресурс патентной информации [25]. The Lens, флагманский проект общественной организации Cambia, направлен на поиск, объединение и связывание различных наборов открытых знаний, включая научные работы и патенты, для информирования об открытии, анализе, принятии решений и партнерстве с удобным пользовательским интерфейсом, основанном на открытой веб-платформе Lens.org. В течение более 20 лет развития, The Lens собирает, объединяет, нормализует и обслуживает более 225 миллионов научных работ и 127 миллионов глобальных патентных записей с обширными метаданными, взятыми из различных источников данных.

Таким образом, в данной работе платформа The Lens использована для сбора данных патентов. Изложенные предсказанные ключевые слова восходящего тренда за период 2017-2019 были использованы в качестве создания поискового запроса в платформе The Lens, который показан на рисунке 5. Это означает, что ключевые слова должны появляться или в Заголовке, или в Аннотации, или в Утверждении (Claim) всех патентных заявок с самой ранней датой приоритета (самая ранняя дата подачи в семействе патентных заявок) в периоде с 01.01.2017 по 31.12.2019. Кроме того, измененные патентные заявки (Amended application), измененные патенты (Amended patent), патенты на право дизайна (design right), неизвестные патенты (Unknown), и абстрактные патенты (Abstract) были исключены. Тем более рассмотрены только патенты из юрисдикции таких ведомств, как USPTO (Ведомство по патентам и товарным знакам США), WIPO (мировое патентохранилище), EPO (Европейская патентная организация). В следующем подразделе «Анализ по кодам CPC патентов» будет приведено более подробное обоснование того, почему были выбраны эти ведомства патентования.

Рис. 5 – Поисковый запрос и фильтр для сбора патентных заявок

На момент написания данной статьи (07.06.2022) были собраны 45973 патентные заявки, данные которых продемонстрированы на рисунке 6.

	Title	Abstract	CPC Classifications
0	METHOD FOR TRAINING A CONVOLUTIONAL RECURRENT ...	A method for training a convolutional recurren...	G06N3/04;;G06N3/08;;G06K9/6271;;G06N3/0454;;G0...
1	Method and apparatus for detecting road lane	A method and an apparatus for detecting a road...	G06K9/6274;;G06N3/0454;;G06N3/0445;;G06V20/588...
2	Spatio-temporal anomaly detection in computer ...	In one embodiment, a device receives sensor da...	G06N3/0445;;G06N3/0454;;G06N3/049;;G06N3/08;;G...
3	TRAINING OF A CONVOLUTIONAL NEURAL NETWORK	The present invention is related to a method, ...	B60W60/001;;B60W2050/0075;;B60W2050/0082;;B60W...
4	MACHINE LEARNING ANALYSIS OF NANOPORE MEASUREM...	A series of measurements taken from a polymer ...	G06N3/0445;;G06N3/0454;;C12Q1/6869;;G01N33/487...
...	...	...	...
45968	METHODS AND SYSTEMS FOR DETERMINING THE CELLUL...	Provided herein are methods for determining th...	G16B20/00;;G16B40/00;;G16B40/20;;G16B30/10;;G1...
45969	METHODS, SYSTEMS, KITS AND APPARATUSES FOR MON...	A variety of kits are provided that are config...	H04L41/0803;;H04L67/12;;H04L41/0886;;Y02P90/02...
45970	POINT CLOUD ENCODING METHOD, POINT CLOUD DECOD...	Embodiments of this application disclose a poi...	G06T9/001
45971	Safety and Stability Control Method against Ve...	A safety and stability control method against ...	B60W10/04;;B60W30/02;;B60C23/04;;B60T8/17558;;...
45972	Safety and Stability Control System against Ve...	Disclosed is a car flat tire safety and stabil...	B60W10/18;;B60W30/02;;B60C23/00;;B60T8/17558;;...

45973 rows × 3 columns

Рис. 6 – Собранные данные патентных заявок (датасет) из Lens.org для анализа

Далее описан метод создания матрицы совместного появления (совпадения, co-occurrence) для ключевых слов и кодов CPC (Cooperative Patent Classification –совместная патентная классификация).

A. Анализ и создание матрицы совпадений

Были разработаны различные подходы к анализу совпадений для извлечения сетей с использованием разных единиц анализа. Например, при анализе совпадения слов используются наиболее важные ключевые слова документов (патентов) для изучения концептуальной структуры области исследования или большой коллекции различных документов. Это единственный метод, который использует фактическое содержание документов для построения меры подобия (similarity measure); другие связывают документы косвенно через коды патентной классификации (или цитаты). Тем самым целью анализа совместного совпадений ключевых слов (или кодов CPC) является построение концептуальной структуры с использованием сети совпадения фраз для сопоставления и их кластеризации, извлеченных из Заголовков и Аннотаций в наборе данных патентов.

В частности, матрица (сеть) совпадения ключевых слов может быть получена по общей формуле:

$$W = A^T \times A \quad (8)$$

где $\boldsymbol{A}$ представляет собой [Patent × Word] матрицу, Word – это ключевые слова, извлеченные из Заголовков и Аннотаций. Матрица $\boldsymbol{A}$ представляет собой прямоугольную бинарную матрицу, где каждая строка является патентом, а каждый столбец относится к извлеченному ключевому слову из текстовых данных патентов. Общий элемент A_{ij} принимает значение по следующей принципе:

$$A_{ij} = \begin{cases} 1, \text{если патент } i \text{ содержит слово } j \\ 0, \text{если патент } i \text{ не содержит слово } j \end{cases} \quad (9)$$

Причем, сумма j-го столбца A_{+j} представляет собой количество патентов, содержащих слово j. Сумма i-й строки A_{i+} – это количество ключевых слов, появившихся в патенте i.

Следовательно, элемент W_{ij} указывает, сколько совпадений существует между ключевыми словами i и j. Диагональный элемент W_{ii} – это количество патентов, содержащих ключевое слово i. Элемент W_{ij} можно рассчитан по следующей формуле (где n – количество патентов):

$$W_{ij} = \sum_{k=1}^{n} A^T{}_{ik} A_{kj} \quad (10)$$

Аналогичным образом, матрица совпадения (совместного появления) кодов CPC также можно получена по формуле:

$$\mathrm{C} = \mathrm{B}^T \times \mathrm{B} \quad (11)$$

где $\boldsymbol{B}$ представляет собой [Patent × CPC] матрицу, CPC – это коды классификации, извлеченные из столбцы «CPC Classifications» данного датасета. Матрица $\boldsymbol{B}$ представляет собой прямоугольную бинарную матрицу, где каждая строка является патентом, а каждый столбец относится к извлеченному коду CPC из набора кодов всей коллекции патентов. Общий элемент B_{ij} принимает значение по следующей принципе:

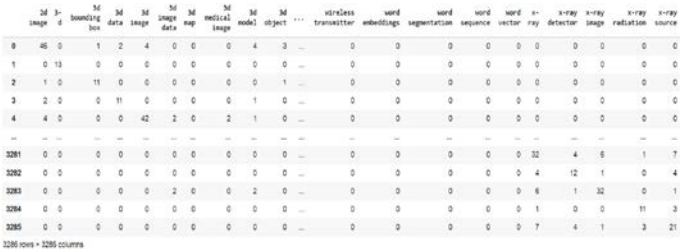
$$B_{ij} = \begin{cases} 1, \text{если патент } i \text{ содержит } CPC \text{ код } j \\ 0, \text{если патент } i \text{ не содержит } CPC \text{ код } j \end{cases} \quad (12)$$

Таким образом, элемент C_{ij} указывает, сколько совпадений существует между кодами CPC i и j. Диагональный элемент C_{ii} – это количество патентов, содержащих коды CPC i. Элемент C_{ij} можно рассчитан по следующей формуле (где n – количество патентов):

$$C_{ij} = \sum_{k=1}^{n} B^T{}_{ik} B_{kj} \quad (13)$$

B. Анализ по извлеченным из патентов ключевым словам

Здесь, вышеизложенный алгоритм извлечения значимых ключевых слов применен для извлечения 10 наиболее значимых ключевых слов из каждой патентной заявки, состоящей из Заголовка и Аннотации. В результате с порогом минимального появления ключевого слова 10 получены 3286 ключевых слов из 45973 патентов. Тогда матрица $\boldsymbol{A}$ ([Patent × Word]) имеет размерность [45954 × 3286] (здесь некоторые патенты были исключены из-за отсутствия ключевого слова, удовлетворенного изложенному условию), и матрица совпадения $\boldsymbol{W}$ имеет размерность [3286 × 3286], которая представлена на рисунке 7.



	3d image	3- d	3d bounding box	3d data	3d image	3d image data	3d map	3d medical image	3d model	3d object	...	wireless transmitter	word embeddings	word segmentation	word sequence	word vector	x-ray	x-ray detector	x-ray image	x-ray radiation	x-ray source
0	46	0	1	2	4	0	0	0	4	3	...	0	0	0	0	0	0	0	0	0	0
1	0	13	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	1	0	11	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
3	2	0	0	11	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
4	4	0	0	0	42	2	0	2	1	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3281	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	32	4	6	1	7
3282	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	4	12	1	0	4
3283	0	0	0	0	0	2	0	0	2	0	...	0	0	0	0	0	6	1	32	0	1
3284	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	11	3
3285	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	7	4	1	3	21

3286 rows × 3286 columns

Рис. 7 – Полученная матрица совпадения ***W***

Затем матрица совпадения ***W*** преобразована в тип разреженной матрицы (sparse matrix) для уменьшения занимаемой памяти при вычислении кластеризации. После этого полученная матрица подана в программу VOSviewer [26] для выполнения кластеризации сети совпадения (co-occurrence network) и визуализации (рисунок 8). При этом минимальная общая сила связи элемента (minimum total link strength of an item) была установлена равной 100. В результате получены 1574 ключевых слов для кластеризации (90 кластеров). Кроме того, следующие параметры были настроены: резолюция (resolution) = 3, минимальный размер кластера = 5.

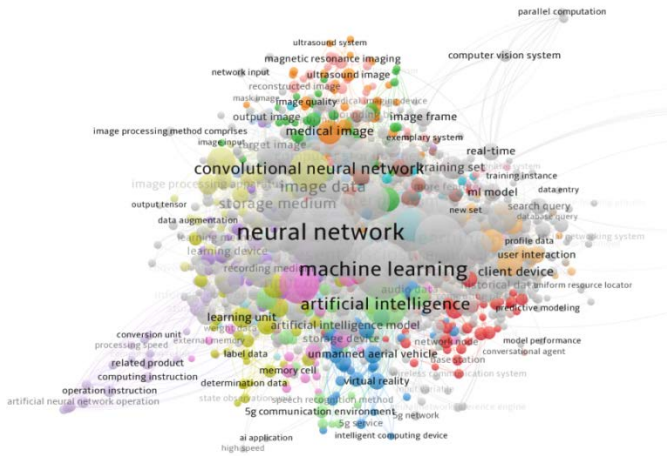


Рис. 8 – Кластеризации сети совпадения ключевых слов на VOSviewer

Следует отметить, что на данном этапе авторы не использовали готовую функции «Создать карту на основе текстовых данных» (create a map based on text data) программы VOSviewer для создания (кластеризации) сети совпадения, поскольку ключевые слова, извлеченные данной программой, в основном состоят из унограмм (unigram) и имеют много шумов (noise), что не позволяют понимать и легко идентифицировать значимые технологические тенденции. Поэтому предложенный авторами алгоритм извлечения значимых ключевых слов позволяет уменьшить количество слов и извлекать представительные фразы существительного сочетания для каждой патентной заявки.

Далее на основе полученных кластеров можно вывести 18 технологических отраслей, которые рассмотрены в качестве предсказанных технологических трендов с 2020 г. и будут представлены в разделе 5.

C. Анализ по кодам CPC патентов

Патентование является показателем производственной деятельности в большей степени, чем публикации академических исследований. В то время как научная литература в основном организована в виде журналов, патенты организованы в системы патентной классификации. Существуют две основные системы классификации: используемая Ведомством по патентам и торговле США (USPTO) и Международная патентная классификация (МПК – IPC), разработанная Всемирной организацией интеллектуальной собственности (ВОИС – WIPO) в Женеве. Последний был впервые разработан для международного патентования в рамках Договора о патентной кооперации (PCT), который был подписан большинством стран мира с момента его создания в 1970 г.

Различные системы патентования предлагают фирмам и изобретателям разные способы патентования: можно патентовать на национальном, международном (в ВОИС) или на региональном уровне, таком как EPO (Европейское патентное ведомство). Известно, что склонность к патентованию и интернационализация патентования будут различаться в зависимости от страны, сектора, дисциплины и т.д. Среди национальных патентов патенты USPTO считаются наиболее ценными из-за конкурентоспособности рынка США, поскольку США является мировым лидером в большинстве технологий. В качестве технологических индикаторов можно считать патенты США наиболее надежными, поскольку фирмы хотят защитить свои права на интеллектуальную собственность на этом крупнейшем рынке.

Кроме того, подобно кодам IPC, CPC представляет собой схему классификации с иерархической структурой, которая классифицирует патентные документы на основе их технической области изобретения. Например, в случае H01L27/11582 полный код CPC состоит из раздела (H), класса (01), подкласса (L), группы (27) и подгруппы (11582). Эту схему можно рассматривать как расширение IPC, и она применяется совместно ведомствами EPO и USPTO. Примечательно, схема CPC больше не является совместному тестированию ведомствами EPO и USPTO. Другие органы, выдающие патенты, все больше начали напрямую классифицировать документы своих национальных ведомств с использованием CPC. При этом благодаря достоинству гибкости система CPC станет все более важным инструментом для поиска документов, особенно при поиске в нескольких патентных органах на нескольких языках [27].

Таким образом, был выбран CPC код среди других кодов классификации патентов (USPC, IPC) для дальнейшего анализа. Сначала осуществлено преобразование кодов CPC исходного формата в более общую классификацию, иными словами исключены номера подгруппы. Например, 3-яя патентная заявка

«Spatio-temporal anomaly detection in computer networks using graph convolutional recurrent neural networks GCRNNs» принадлежит следующим кодам классификации: G06N3/0445; G06N3/0454; G06N3/049; G06N3/08; G06N3/084; G06N5/022; G06N20/10; H04L63/1425; H04L63/1425; H04L63/1416. Тогда после текстовой обработки получены следующие коды: G06N3, G06N5, G06N20, H04L63, которые существенно упростили объем данных для дальнейшего анализа.

В результате получены 1292 кода CPC из 45973 патентов с порогом минимального появления кода CPC равным 5. Тогда матрица ***B*** ([Patent × CPC]) имеет размерность [45940 × 1292] (здесь некоторые патенты были исключены из-за отсутствия присваивания кода классификации), и матрица совпадения ***C*** имеет размерность [1292 × 1292], которая представлена на рисунке 9.

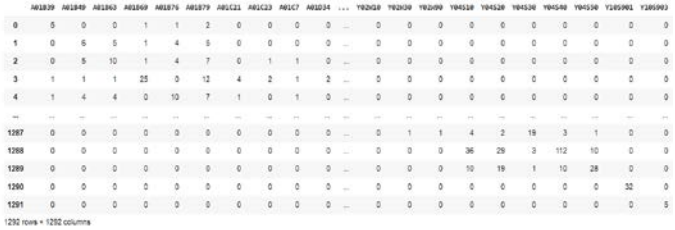


Рис. 9 – Полученная матрица совпадения ***C***

Затем матрица совпадения ***C*** преобразована в тип разреженной матрицы для уменьшения занимаемой памяти при вычислении кластеризации. После этого полученная матрица подана в программу VOSviewer для выполнения кластеризации сети совпадения и визуализации (рисунок 10). При этом минимальная общая сила связи элемента установлена равной 200. В результате получены 687 кодов CPC для кластеризации (19 кластеров). Кроме того, следующие параметры были настроены: резолюция (resolution) = 2, минимальный размер кластера = 5.

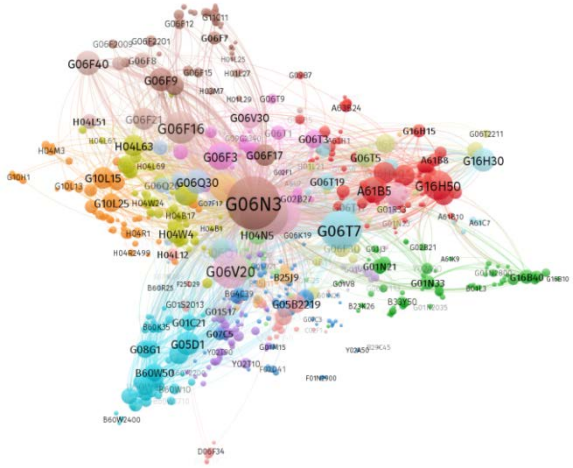


Рис. 10 – Кластеризации сети совпадения кодов CPC на VOSviewer

Таблица 1 – Идентификация технологических трендов и оценка точности метода прогноза

При этом на основе полученных кластеров можно вывести 18 технологических отраслей, которые рассмотрены в качестве предсказанных технологических трендов с 2020 г. и будут представлены в разделе 5.

V. Оценка предложенного метода прогноза

Чтобы оценить точность предложенного метода для прогноза технологических тенденций, мы исследуем их с теми инновациями, которые описаны в книге «Технологические тенденции будущего на практике». В этой книге представлены 25 наиболее важных технологических тенденций в мире с 2020 г., а также их потенциальный вклад в организационный успех [28,29]. При этом можно узнать, как интегрировать существующие достижения, в полной мере использовать их для развития бизнеса и планировать те, которые находятся на пути. В этой книге консультант по стратегическому бизнесу Bernard Marr объясняет роль технологий в предоставлении инновационных бизнес-решений для компаний разного размера и в разных отраслях. Также он освещает широкий спектр тенденций и дает обзор того, как компании используют эти новые и появляющиеся технологии на практике.

Bernard Marr [30] – автор всемирно известных бестселлеров, популярный основной докладчик (keynote speaker), футурист и советник по стратегическим вопросам бизнеса и технологий для правительств и крупных компаний. Он помогает организациям и их управленческим командам подготовиться к новой промышленной революции. Известный автор написал более 15 книг и сотни качественных отчетов и статей, в том числе бестселлеры «Искусственный интеллект на практике», «Большие данные на практике» и «Революция в интеллекте» [2,31].

Из предыдущих разделов был приведен анализ данных патентных заявок в периоде 2017-2019. На основании выделенных кластеров ключевых слов и кодов CPC получено всего 22 практических технологических отраслей. Рассмотрим эти отрасли как технологические тренды в 2020-2023 годах. Для проверки точности предложенного метода прогноза и полученных 22 трендов были обнаружены технологические тренды в данной известной книге. Другими словами, эксперт Bernard Marr уже подтвердил эти тренды своими публикациями, которые относятся к соответствующим группам ключевых слов и кодов (таблица 1).

Номер п/п	**Обнаруженные ключевые слова**	**Обнаруженные коды CPC**	**Соответствующий технологический тренд** [28,29]
1	(Кластер 4, 71): artificial intelligence technique, deep learning technique, machine learning technique, multiple data source, machine learning method, machine learning apparatus, machine learning program.	(Кластер 17): G06N20	01 - Machine learning, Artificial intelligence
2	(Кластер 11): home applicance, identification information, internet-of-things, mobile robot, mobile terminal, other electronic device.	(Кластер 4): H04L67, H04W4, H04W72, H04W76, H04W8, H04W84, H04W88, H04W64.	02 - The Internet of Things (IoT)
3	(Кластер 7, 10, 18, 19): anatomical feature, biologicial tissue, blood vessel, human body, imaging data, medical image data, medical image processing, tissue sample, ultrasound image, ultrasound transducer, 3d image, magnetic resonance imaging, image reconstruction, medical instrument, mr image, clinical data, electronic medical record, electronic processor, medical device, medical record, monitoring device, patient data, patient information, smart phone, blood pressure, ecg signal, emotional state, physiological data, vibration sensor, wearable device.	(Кластер 1): A61B1, A61B10, A61B18, A61B2576, A61B34, A61B5, A61H2230, A63B24, G16H10, G16H20, G16H40, G16H50, G16H80,	03 - Digital health, wearables and augmented humans
4	(Кластер 65, 73): big data, data analytics.	(Кластер 6): B60W2556	04 - Big Data and augmented analytics
5	(Кластер 57, 63): blockchain network, machine learning model, data block, computing node, data compression, data format, data structure, data management.	(Кластер 12): H04L2209, H04L9	06 - Blockchains and distributed ledgers
6	(Кластер 24, 29, 53, 65): computing device, data transfer, edge device, neural network inference engine, predictive analysis, real-time data, remote computing device, client device, digital medium environment, edge node, machine readable medium, mobile device, client device, position data, virtual environment, cloud server, mobile application, mobile phone, cloud platform.	–	07 - Cloud and edge computing
7	(Кластер 3): artificial intelligence device, artificial intelligence apparatus, augmented reality, virtual reality, intelligent computing device, iot device, intelligent device.	(Кластер 18): G06K9, G06V10, G06V20, G06V2201, G06V30	08 - Digitally extended realities
8	(Кластер 62): machine translation, natural language, natural language input, natural language processing, natural language query, natural language understanding.	(Кластер 17): G06F40, G06F8, G06N5, G06N7, H04L51	10 - Natural language processing
9	(Кластер 3, 30, 62): audio data, acoustic feature, artificial intelligence model, audio segment, automatic speech recognition, microphone array, sound data, speech recognition method, voice recognition, acoustic signal, input audio signal, signal processing device, sound source, recording medium, virtual assistant.	(Кластер 7): G10L13, G10L15, G10L17, G10L21, G10L25	11 - Voice interfaces and chatbots

10	(Кластер 12, 13, 86): face image, face detection, facial recognition, human face, image classification method, image feature, neural network training, pre-trained convolutional neural network, recognition system, video camera, image processing system, image recognition method, computer vision, computer vision system.	(Кластер 9): A63F13, A63F2300, G06F3, G06T1, G06T3, G09G5, H04N13, H04N19, H04N21	12 - Computer vision and facial recognition
11	(Кластер 2, 23): autonomous driving system, 3d point cloud, autonomous machine application, autonomous vehicle, driver assistance system, lidar sensor, ego vehicle, object recognition method, radar sensor, radar signal, semi-autonomous vehicle.	(Кластер 6): B60W10, B60W2420, B60W2720, B60W30, B60W40, B60W50, B60W60, G01C21, G05D1, G05D2201, G05D2201	14 - Autonomous vehicles
12	(Кластер 3, 11, 89): artificial intelligence algorithm, control command, communication unit, 5g communication environment, 5g communication network, 5g environment, 5g service, 5g network.	–	15 - 5G network
13	–	(Кластер 2): C12Q1, C12Q2600, G01N21, G01N33, G16B20, G16B40, G16C20	16 - Genomics and gene editing
14	(Кластер 22, 28, 65): access control, client computing device, online system, content provider, recommendation system, social network system, user data, user identity, application program, computer-implemented system, uniform resource location, web browser, web page.	(Кластер 12): G06Q10, G06Q20, G06Q30, G06Q40	18 - Digital platforms
15	(Кластер 3, 40): unmanned aerial vehicle, aerial vehicle, deep reinforcement learning, traffic data.	(Кластер 3): B64C2201, B64C39, B64F5, F02D41, G05B19, G05B23, G07C5, G08G5	19 - Drones and unmanned aerial vehicles
16	(Кластер 1, 17): anomaly detection, communication session, computer network, machine learning-based model, network device, network traffic, predictive model, remote device, wireless network, access point, data packet, cellular network, mobile communication device, sensitive data, sensitive information.	(Кластер 4): H04L1, H04L2463, H04L41, H04L43, H04L63, H04W12	20 - Cybersecurity
17	(Кластер 4, 57, 60): control device, industrial machine, learning model, learning data, machine tool, machining condition, observed state variable, robot system, servo control device, computer program product, robotic process automation, artificial intelligence system, robotic device, robotic system,	(Кластер 16): A47L2201, A47L9, B25J13, B25J19, B25J9, G06V40	22 - Robotic process automation
18	(Кластер 28, 65): interaction data, online service, personal data, profile data, server computer, data mining, electronic communication, electronic message, personal information,	–	23 - Mass personalization and micro-moments

19	–	B22F10, B29C64, B33Y50, G06F2111, G06F30, G06T11, G06T5, H01L21, Y02P90	24 - 3D, 4D printing and additive manufacturing
20	(Кластер 15, 16, 50): arithmetic circuit, arithmetic device, arithmetic operation, arithmetic processing, memory bank, multi-layer neural network, neural network processor, processing engine, computer processor, computerized system, computational resource, graphical representation, graphical user interface, manufacturing process, manufacturing system, semiconductor manufacturing process → (Технология полупроводниковых процессоров)	–	–
21	–	(Кластер 5): B60L53, B60L58, G01R31, H02J3, H02J7, Y02E10, Y02E60, Y02T90, Y04S10, Y04S40 → Технология электрических и гибридных автомобилей	–
22	–	(Кластер 14): E21B2200, E21B43, E21B44, E21B47, E21B49, G01R33, G01V1 → Технология разведки нефти и газа	–

В качестве показателей для оценки точности метода идентификации и прогнозирования тенденции технологий используются метрики Точность, Полнота, и F1. Обозначаем *a*: количество технологий, найденных системой и релевантных с точки зрения эксперта Bernard Marr [28]; *b*: количество ключевых слов, найденных системой, но не релевантных с точки зрения эксперта Bernard Marr; *c*: количество релевантных технологий, не найденных системой; то метрики вычисляются следующим образом:

Точность (precision) – это отношение найденных релевантных технологий к общему количеству найденных технологий:

$$P = \frac{a}{a+b} = \frac{19}{22} = 0.86 \quad (14)$$

Полнота (recall) – это отношение найденных релевантных технологий к общему количеству релевантных технологий:

$$R = \frac{a}{a+c} = \frac{19}{25} = 0.76 \quad (15)$$

Для наглядности, значения переменных a, b, c иллюстрированы на рисунке 11.

	релевантны	не релевантны
найдено системой	a	b
не найдено системой	c	d

Рис. 11 – Иллюстрация переменных a, b, c

Тогда:

$$F1 = \frac{2PR}{P+R} \approx 0.81 \quad (16)$$

Ввиду этого точность F1 предложенного метода идентификации и прогнозирования тенденции технологий составляет около 81%, что свидетельствует о положительной надежности предлагаемого метода.

VI. Заключение

Технологические разработки оказывают существенное влияние на принятие стратегических решений. Заблаговременное осознание возможных предстоящих или возникающих технологических тенденций может привести к укреплению конкурентоспособности и рыночных позиций предприятий. Кроме того, крупные предприятия уже учитывали этот важный аспект и создали так называемые «Инновационные центры» для прогноза будущих технологических разработок и грядущих инноваций.

По этой причине, в данной статье предложен новый метод для АПТТ на основе обработки разнородных данных (научные статьи, патенты) из открытых источников. Известно, что жизненный цикл технологии начинается с научных публикаций, за которыми следуют заявки на патенты, а затем другие технологические новости. Следовательно, сначала анализированы научные публикации для выявления основных научных и технологических направлений, начиная с известной наукометрической базы данных Web of Science, а затем

патентные заявки использованы для анализа и прогноза технологических трендов.

Для этого разработан новый метод извлечения самых значимых ключевых слов из каждого текста патентной заявки, состоящей из Заголовка и Аннотации. Далее при применении метода регрессии Гауссовского процесса собраны 20 явно восходящих по ключевым словам исследовательских тенденций, и это также прогнозируемые тенденции исследования в периоде 2017-2019. После этого, эти предсказанные ключевые слова были использованы в качестве создания поискового запроса в платформе The Lens для сбора данных патентных заявок в периоде 2017-2019. Затем с помощью алгоритма извлечения самых значимых ключевых слов, метода обработки текстовых данных, операции матричного умножения разработан метод создания матриц совместного появления для ключевых слов и кодов CPC, которые позже поданы в программу VOSviewer для кластеризации сети совпадения и визуализации. Анализ совпадений таких элементов (ключевых слов, кодов CPC) позволяет создать семантические карты области исследования или коллекции документов, которые облегчают понимание их когнитивной структуры.

На основании кластеров ключевых слов и кодов CPC получено всего 22 практических технологических отраслей. Рассмотрим эти отрасли как технологические тренды с 2020 года. Для проверки точности предложенного метода прогноза и полученных 22 трендов были изучены технологические тренды в изложенной известной книге эксперта Bernard Marr. В результате проверки обнаружены 19 из 22 предсказанных технологических трендов в наборе 25 истинных технологических трендов. Ввиду этого точность F1 предложенного метода идентификации и прогнозирования тенденции технологий составляет около 81%, что свидетельствует о положительной надежности предлагаемого метода.

Библиография

1. Клименко А.Г., Зайцев К.С. Исследование подходов к разработке умных объектов // International journal of open information technologies. 2022. Т. 10, № 6. С. 141–148.
2. Marr B. The 4 Biggest Trends In Big Data And Analytics Right For 2021 [Электронный ресурс] // Forbes. 2021. URL: forbes.com/sites/bernardmarr/2021/02/22/the-4-biggest-trends-in-big-data-and-analytics-right-for-2021/ (дата обращения: 12.07.2022).
3. Кравец А.Г., Сальникова Н.А. Предсказательное моделирование трендов технологического развития // Известия Санкт-Петербургского государственного технологического института (технического университета). 2020. № 55 (81). С. 103–108.
4. Viet N.T., Kravets A., Duong Q.H.T. Data Mining Methods for Analysis and Forecast of an Emerging Technology Trend: A Systematic Mapping Study from SCOPUS Papers // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021. С. 81–101.
5. Viet N.T., Gneushev V. Analyzing and Forecasting Emerging Technology Trends by Mining Web News // Communications in Computer and Information Science. 2021. С. 55–69.
6. Lee C. A review of data analytics in technological forecasting // Technological Forecasting and Social Change. 2021.
7. Нгуен Т.В., Кравец А.Г., Щербаков М.В. Метод для анализа тенденций развития технологий управления эффективностью активов // Прикаспийский журнал: управление и высокие технологии. 2022. Т. 57, № 1. С. 39–53.
8. Viet N.T., Kravets A.G. Analyzing Recent Research Trends of Computer Science from Academic Open-access Digital Library // 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, 2019. С. 31–36.
9. Kravets A.G., Vasiliev S.S., Shabanov D. V. Research of the LDA Algorithm Results for Patents Texts Processing // 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2018. С. 1–6.
10. Zhou Y. и др. Forecasting emerging technologies using data augmentation and deep learning // Scientometrics. 2020.
11. Song K., Kim K., Lee S. Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents // Technological Forecasting and Social Change. 2018. Т. 128. С. 118–132.
12. Rotolo D., Hicks D., Martin B.R. What is an emerging technology? // Research Policy. 2015. Т. 44, № 10. С. 1827–1843.
13. Ena O. и др. A methodology for technology trend monitoring: the case of semantic technologies // Scientometrics. 2016. Т. 108, № 3. С. 1013–1041.
14. Li X. и др. Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology // Technological Forecasting and Social Change. 2019. Т. 146. С. 432–449.
15. Wang M.-Y., Fang S.-C., Chang Y.-H. Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels // Technological Forecasting and Social Change. 2015. Т. 92. С. 182–195.
16. Wei F. и др. Decreasing the noise of scientific citations in patents to measure knowledge flow // 17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings. 2019. С. 1662–1669.
17. Suominen A., Ranaei S., Dedehayir O. Exploration of Science and Technology Interaction: A Case Study on Taxol // IEEE Transactions on Engineering Management. 2021. Т. 68, № 6. С. 1786–1801.
18. Li X. и др. Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining // Nano Energy. 2020. Т. 71. С. 104636.
19. Нгуен Т. В., Кравец А. Г. Оценка и прогнозирование тенденций развития научных исследований на основе библиометрического анализа публикаций // Информационные технологии. 2021. Т. 27, № 4. С. 195–201.
20. Industrial-Strength Natural Language Processing [Электронный ресурс]. 2022. URL: https://spacy.io/ (дата обращения: 26.02.2022).

21. Natural Language Toolkit [Электронный ресурс]. 2022. URL: https://www.nltk.org/index.html (дата обращения: 27.02.2022).
22. SentenceTransformers Documentation [Электронный ресурс]. 2022. URL: https://www.sbert.net/ (дата обращения: 26.02.2022).
23. KeyBERT - Quickstart [Электронный ресурс]. 2022. URL: https://maartengr.github.io/KeyBERT/guides/quickstart.html (дата обращения: 26.02.2022).
24. Schulz E., Speekenbrink M., Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions // Journal of Mathematical Psychology. 2018.
25. About The Lens [Электронный ресурс]. 2022. URL: https://about.lens.org/ (дата обращения: 19.06.2022).
26. van Eck N.J., Waltman L. Visualizing Bibliometric Networks // Measuring Scholarly Impact. Cham: Springer International Publishing, 2014. С. 285–320.
27. Chiba C. Life After USPC. 5 Things Patent Searchers Should Know About CPC [Электронный ресурс]. URL: lexisnexis.de/expertenbeitraege-webinare/downloads/whitepaper/life-after-uspc (дата обращения: 12.07.2022).
28. Marr B. Tech Trends in Practice : The 25 technologies that are driving the 4th Industrial Revolution // Machine Learning Process. 2020.
29. Bernard Marr. These 25 Technology Trends Will Define The Next Decade // Forbes. 2020.
30. Bernard Marr [Электронный ресурс]. URL: https://www.forbes.com/sites/bernardmarr (дата обращения: 12.07.2022).
31. Marr B. The 5 Big Problems With Blockchain Everyone Should Be Aware Of // Forbes. 2018.

A new method for predicting technological trends based on the analysis of scientific articles and patents

Nguyen Thanh Viet, A. G. Kravets

***Abstract*—To achieve competitiveness in a rapidly changing science, it is important to follow the development of existing technologies and discover new and promising technologies. Firms need to develop a technology development strategy by predicting technology trends in order to gain a competitive advantage while using limited resources. On the other hand, nowadays the number of scientific articles, patents and other miscellaneous data is growing at a rapid pace, and it becomes impossible to stay up to date with everything that is published. However, despite all efforts, none of existing methodological and technological results are able to create models and methods for the holistic perception of heterogeneous information by a computing system – scientific publications and patents, which is contained in open sources. At the same time, most of the existing studies are intended for the analysis and early detection of new technologies or monitoring trends in some specific technology industries, without considering the solution of the problem of predicting many different technology trends. In addition, the accuracy of the assessment of the proposed methods in existing studies is either rather low (the maximum metric F1 for assessing the accuracy of the forecast is ~ 74%), or is absent (the quality of the method has not been assessed). Thus, this article proposes a new method for analyzing and predicting technological trends based on the processing of heterogeneous data (scientific articles, patents) from open sources by developing an algorithm for extracting significant keywords and methods for creating co-occurrence matrices of elements (keywords, CPC codes).**

***Keywords*—technological prediction, keyword extraction, Gaussian process, coincidence matrix, coincidence network clustering, VOSviewer.**

REFERENCES

1. Klimenko A.G., Zajcev K.S. Issledovanie podhodov k razrabotke umnyh ob#ektov // International journal of open information technologies. 2022. Vol. 10, № 6. pp. 141–148.
2. Marr B. The 4 Biggest Trends In Big Data And Analytics Right For 2021 [Jelektronnyj resurs] // Forbes. 2021. URL: forbes.com/sites/bernardmarr/2021/02/22/the-4-biggest-trends-in-big-data-and-analytics-right-for-2021/ (data obrashhenija: 12.07.2022).
3. Kravec A.G., Sal'nikova N.A. Predskazatel'noe modelirovanie trendov tehnologicheskogo razvitija // Izvestija Sankt-Peterburgskogo gosudarstvennogo tehnologicheskogo instituta (tehnicheskogo universiteta). 2020. № 55 (81). pp. 103–108.
4. Viet N.T., Kravets A., Duong Q.H.T. Data Mining Methods for Analysis and Forecast of an Emerging Technology Trend: A Systematic Mapping Study from SCOPUS Papers // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021. pp. 81–101.
5. Viet N.T., Gneushev V. Analyzing and Forecasting Emerging Technology Trends by Mining Web News // Communications in Computer and Information Science. 2021. pp. 55–69.

6. Lee C. A review of data analytics in technological forecasting // Technological Forecasting and Social Change. 2021.
7. Nguen T.V., Kravec A.G., Shherbakov M.V. Metod dlja analiza tendencij razvitija tehnologij upravlenija jeffektivnost'ju aktivov // Prikaspijskij zhurnal: upravlenie i vysokie tehnologii. 2022. Vol. 57, № 1. pp. 39–53.
8. Viet N.T., Kravets A.G. Analyzing Recent Research Trends of Computer Science from Academic Open-access Digital Library // 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, 2019. pp. 31–36.
9. Kravets A.G., Vasiliev S.S., Shabanov D. V. Research of the LDA Algorithm Results for Patents Texts Processing // 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2018. pp. 1–6.
10. Zhou Y. i dr. Forecasting emerging technologies using data augmentation and deep learning // Scientometrics. 2020.
11. Song K., Kim K., Lee S. Identifying promising technologies using patents: A retrospective feature analysis and a prospective needs analysis on outlier patents // Technological Forecasting and Social Change. 2018. Vol. 128. pp. 118–132.
12. Rotolo D., Hicks D., Martin B.R. What is an emerging technology? // Research Policy. 2015. Vol. 44, № 10. pp. 1827–1843.
13. Ena O. i dr. A methodology for technology trend monitoring: the case of semantic technologies // Scientometrics. 2016. Vol. 108, № 3. pp. 1013–1041.
14. Li X. i dr. Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology // Technological Forecasting and Social Change. 2019. Vol. 146. pp. 432–449.
15. Wang M.-Y., Fang S.-C., Chang Y.-H. Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels // Technological Forecasting and Social Change. 2015. Vol. 92. pp. 182–195.
16. Wei F. i dr. Decreasing the noise of scientific citations in patents to measure knowledge flow // 17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings. 2019. pp. 1662–1669.
17. Suominen A., Ranaei S., Dedehayir O. Exploration of Science and Technology Interaction: A Case Study on Taxol // IEEE Transactions on Engineering Management. 2021. Vol. 68, № 6. pp. 1786–1801.
18. Li X. i dr. Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining // Nano Energy. 2020. Vol. 71. pp. 104636.
19. Nguen T. V., Kravec A. G. Ocenka i prognozirovanie tendencij razvitija nauchnyh issledovanij na osnove bibliometricheskogo analiza publikacij // Informacionnye tehnologii. 2021. Vol. 27, № 4. pp. 195–201.
20. Industrial-Strength Natural Language Processing [Jelektronnyj resurs]. 2022. URL: https://spacy.io/ (data obrashhenija: 26.02.2022).
21. Natural Language Toolkit [Jelektronnyj resurs]. 2022. URL: https://www.nltk.org/index.html (data obrashhenija: 27.02.2022).
22. SentenceTransformers Documentation [Jelektronnyj resurs]. 2022. URL: https://www.sbert.net/ (data obrashhenija: 26.02.2022).
23. KeyBERT - Quickstart [Jelektronnyj resurs]. 2022. URL: https://maartengr.github.io/KeyBERT/guides/quickstart.html (data obrashhenija: 26.02.2022).
24. Schulz E., Speekenbrink M., Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions // Journal of Mathematical Psychology. 2018.
25. About The Lens [Jelektronnyj resurs]. 2022. URL: https://about.lens.org/ (data obrashhenija: 19.06.2022).
26. van Eck N.J., Waltman L. Visualizing Bibliometric Networks // Measuring Scholarly Impact. Cham: Springer International Publishing, 2014. S. 285–320.
27. Chiba C. Life After USPC. 5 Things Patent Searchers Should Know About CPC [Jelektronnyj resurs]. URL: lexisnexis.de/expertenbeitraege-webinare/downloads/whitepaper/life-after-uspc (data obrashhenija: 12.07.2022).
28. Marr B. Tech Trends in Practice : The 25 technologies that are driving the 4th Industrial Revolution // Machine Learning Process. 2020.
29. Bernard Marr. These 25 Technology Trends Will Define The Next Decade // Forbes. 2020.
30. Bernard Marr [Jelektronnyj resurs]. URL: https://www.forbes.com/sites/bernardmarr (data obrashhenija: 12.07.2022).
31. Marr B. The 5 Big Problems With Blockchain Everyone Should Be Aware Of // Forbes. 2018.