Систематический обзор научной литературы на основе анализа данных и тематического моделирования по цифровым двойникам: статья Алексея Боровкова, Кузьмы Кукушкина и Юрия Рябова опубликована в журнале Data (MDPI)

В декабрьском номере журнала *Data* одного из крупнейших в мире издательств MDPI опубликована статья **Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling** «Цифровой двойник. Систематический обзор научной литературы на основе анализа данных и тематического моделирования».

Авторами являются проректор по цифровой трансформации СПбПУ, руководитель Передовой инженерной школы СПбПУ «Цифровой инжиниринг», Научного центра мирового уровня СПбПУ «Передовые цифровые технологии», Центра компетенций НТИ СПбПУ «Новые производственные технологии» и Инжинирингового центра (CompMechLab®) СПбПУ **Алексей Боровков**, генеральный директор Ассоциации «Технет» **Кузьма Кукушкин** и начальник отдела технологического и промышленного форсайта Инжинирингового центра «Центр компьютерного инжиниринга» (CompMechLab®) СПбПУ **Юрий Рябов**.

Статья, подготовленная в рамках программы «Приоритет-2030», посвящена

разработке собственной методики анализа большого объема публикаций о цифровых двойниках (ЦД). В ее основе лежит интеллектуальный анализ текстов (text mining) с применением методов машинного обучения и обработки естественного языка. Были использованы метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) и тематическое моделирование BERTopic (кластеризация семантических векторов).

Как отметил **Алексей Боровков**, цифровые двойники выступают инструментом для решения высокотехнологичных задач в самых разных областях: от промышленного производства до медицины. Это, в свою очередь приводит к разнообразию подходов к определению понятия ЦД. С быстрым увеличением числа тематических статей растет количество новых интерпретаций.

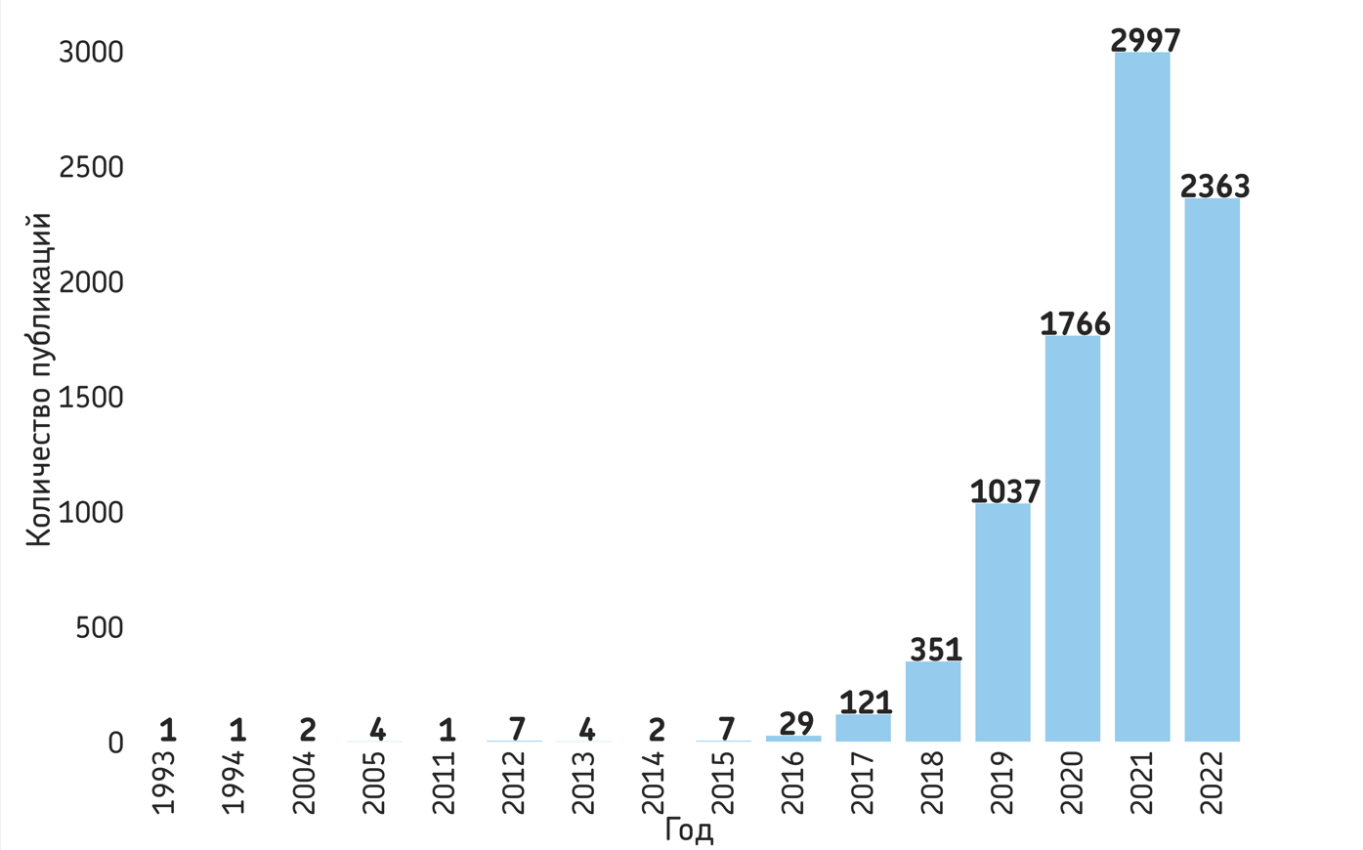
«Стандартизация определений ЦД является важной задачей. Во-первых, это позволит определить, что такое цифровой двойник, а что цифровым двойником не является. Во-вторых, это может помочь в создании более совершенных и универсальных методов разработки ЦД. Еще одна важная задача — разработать четкую типологию ЦД для разных применений. Необходимо обобщить всю информацию из разных исследований, что может помочь разработать четкие определения ЦД и основные области применения концепции», – подчеркнул **Алексей Иванович**.

В качестве источника информации о научных публикациях в проведенном исследовании использовалась библиографическая база данных рецензируемой научной литературы Scopus. В выборку попали **8693** статьи по теме ЦД, размещенные в период с января 1993 года по сентябрь 2022 года. Авторы подготовили подробный обзор публикаций, проанализировали подходы к концепции ЦД, используемые российскими и зарубежными учеными.

«Вопрос об определениях и трактовках термина ЦД стоит наиболее остро: единство мнений существует только по совсем базовым основам трактовки понятия "цифровой двойник". Поэтому многие коллеги изучают подходы к определению и систематизируют основные концепции. Например, делают выборки из нескольких десятков статей и анализируют их вручную или выбирают несколько сотен наиболее цитированных статей и проводят автоматизированный анализ. Главное отличие нашего исследования – фокусировка сразу на всем массиве статей по тематике цифровых двойников, выгружаемых системой Scopus», – прокомментировал **Юрий Рябов**.

Отмечено как общее увеличение числа публикаций, посвященных ЦД, так и рост сложности обзоров в статьях. Зафиксировано появление метаобзоров по тематике ЦД, что указывает на поиск исследователями новых подходов и методов наращивания

объема анализируемых выборок и снижения сложности обобщения большого количества публикаций.



Источник: Scopus, 1993–2022 (8 мес.) гг.

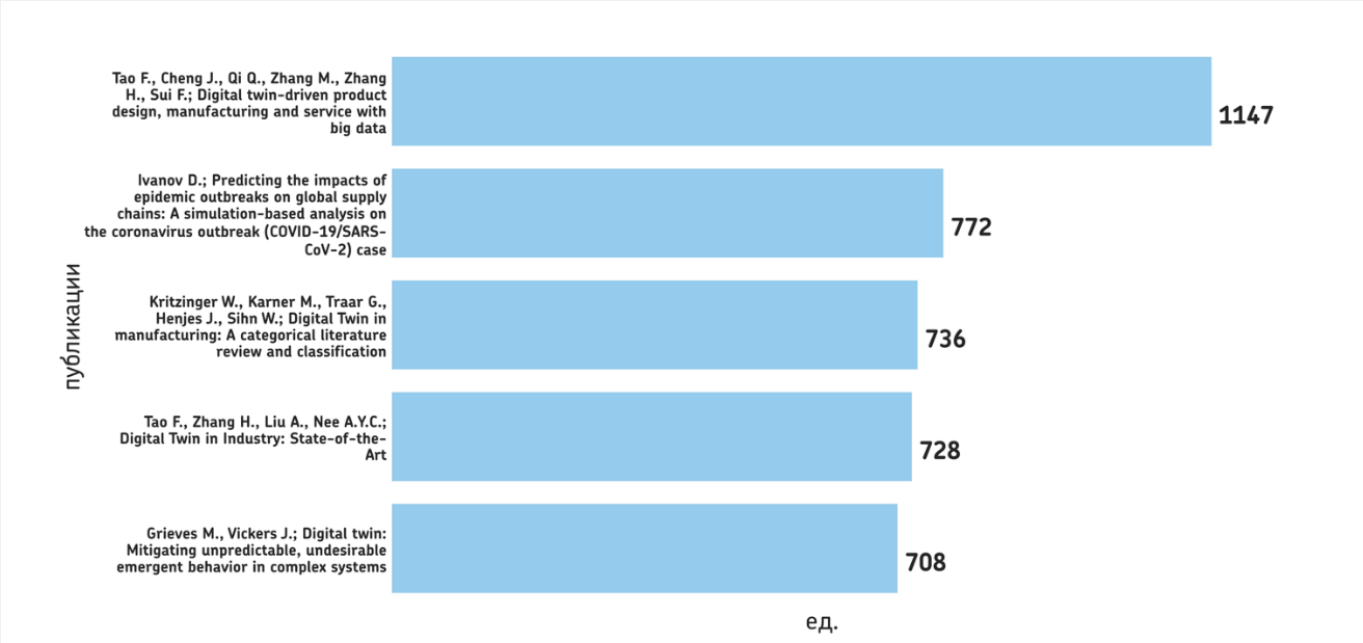
В практической части статьи приведены подробные описания каждой из выбранных методик интеллектуального анализа публикаций. В качестве инструмента обработки данных использовано приложение Jupyter Notebook, позволяющее проводить анализ данных с использованием языка программирования Python.

Как подчеркнули авторы исследования, если LDA-модель основана на статистических расчетах, то BERT-модель учитывает контекст употребления слов, то есть слова, которые встречаются в похожих контекстах и имеют близкие значения. Уточняется, что для целей данного исследования целесообразнее использовать BERTopic, однако LDA-модель была построена для более полного анализа статей по тематике ЦД.

Основное исследование проходило в три этапа:

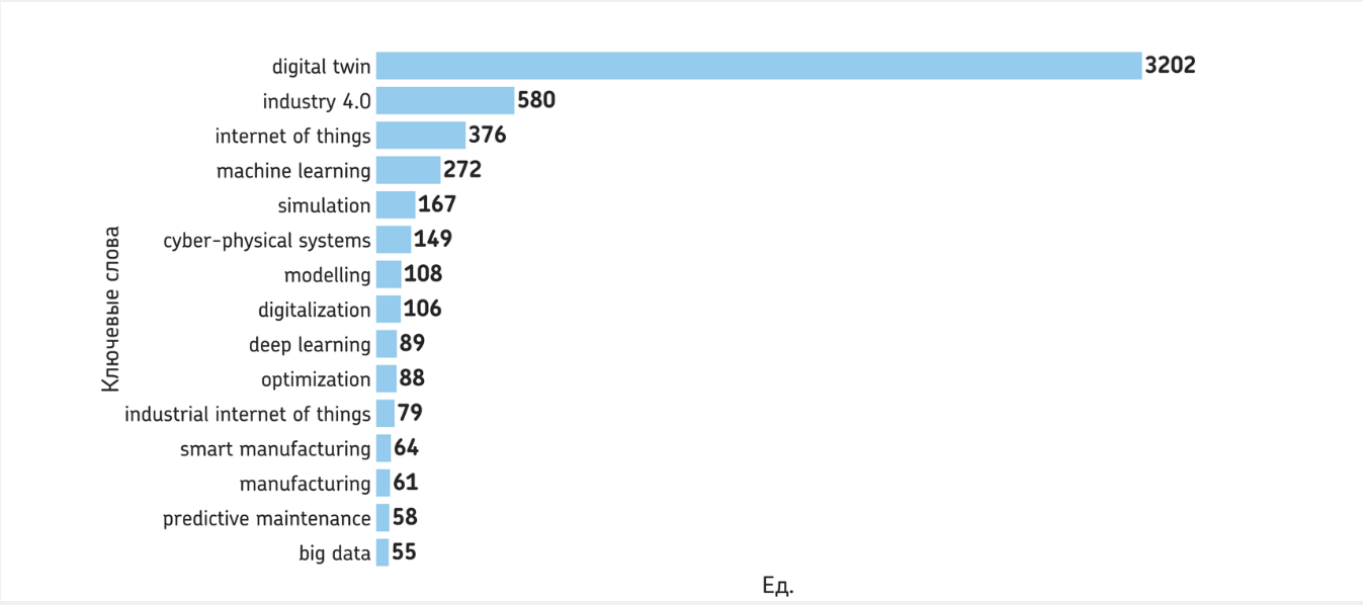
Выгрузка данных в Jupyter Notebook, расчет числа публикаций по годам, анализа соавторства, выявление наиболее цитируемых авторов, определение наиболее цитируемых публикаций.
Анализ ключевых слов публикаций.
Анализ аннотаций публикаций.

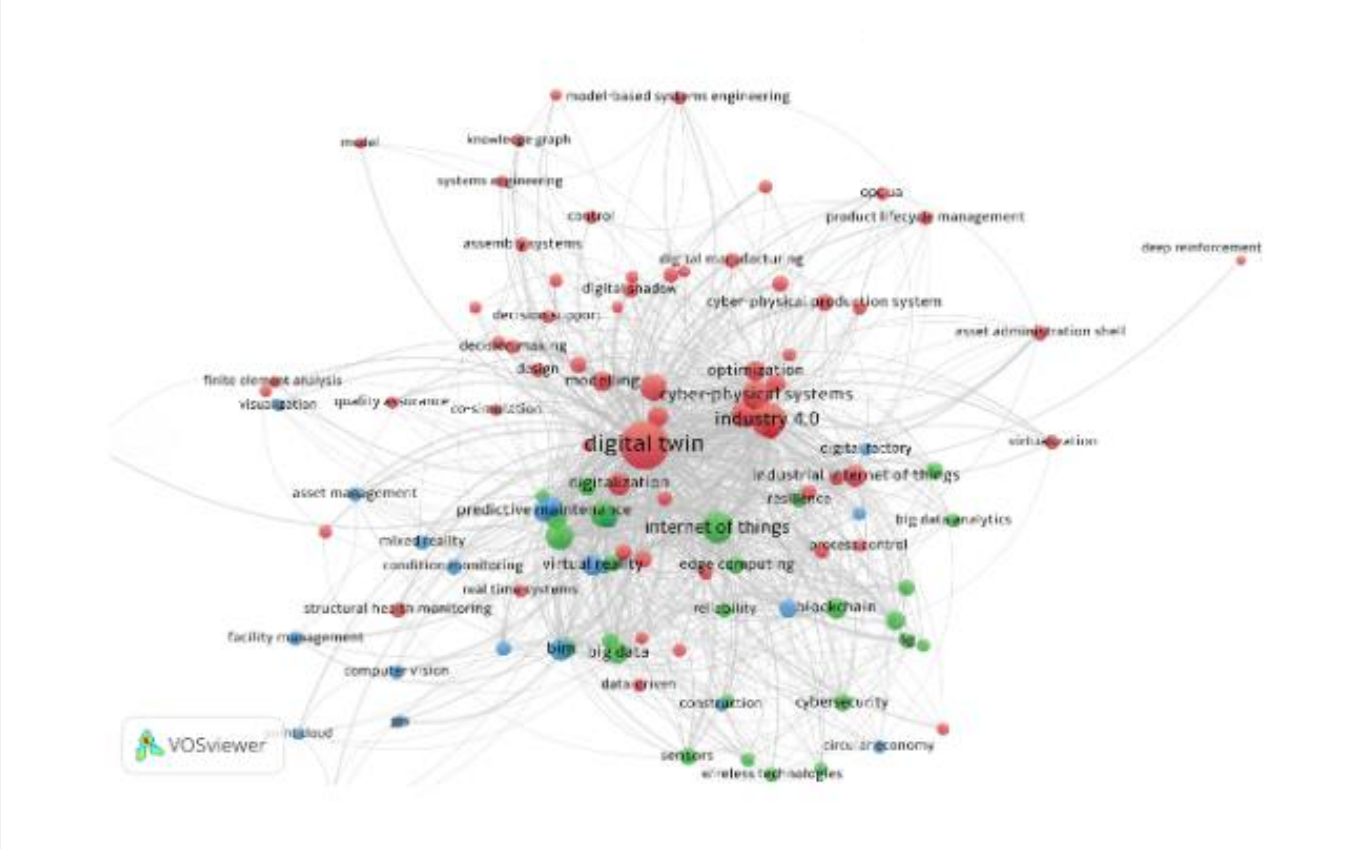
Выгруженный перечень статей оформлен в виде графика. Статьи отсортированы по количеству цитирований по убыванию.



Источник: Scopus, 1993–2022 (8 мес.) гг.

Для определения трендов на основе исследований публикаций на тему ЦД использовался подсчет количества упоминаний ключевых слов, а также была проанализирована частотность соупоминаний ключевых слов. Авторы исследования пришли к выводу, что наиболее часто используются такие термины, как «Индустрия 4.0», «интернет вещей», «машинное обучение», «моделирование» и «киберфизические системы».





Соупоминание ключевых слов. Источник: по материалам Центра НТИ СПбПУ

На третьем этапе исследования выполнено подробное сравнение LDA-модели и BERTopic-модели, описаны их преимущества и недостатки, а также в сравнении представлены результаты, полученные после применения обеих методик.

Подводя итог исследованию, авторы отметили, что анализ ключевых слов в выбранных публикациях показал образование вокруг термина ЦД («Digital Twin») набора терминов, которые упоминаются только непосредственно с ним. Это может указывать на формирование собственного исследовательского поля ЦД. В свою очередь, это означает, что научные публикации по данной теме находятся на ранних стадиях развития, но уже заметны признаки отдельного научного направления.

«По результатам исследования методом LDA-анализа (латентное размещение Дирихле) мы получили 20 основных тем в изучаемых статьях, а по результатам изучения с помощью другого метода, тематического моделирования BERTopic (кластеризация семантических векторов) – более 100. Следующий этап исследований – объединить две модели для еще более точного определения ключевых тем. Уже текущие результаты показывают высокое разнообразие подходов к изучению цифровых двойников, а также основные отрасли применения технологии», – отметил **Кузьма Кукушкин**.

Проведенный анализ показал рост количества публикаций по таким темам, как:

Цифровые двойники сборочных линий и роботов.
BIM-технологии и цифровые двойники в строительстве.
Цифровые двойники электросетей и систем распределения энергии.

Как предполагают авторы исследования, такая тенденция сохранится на ближайшие годы. Также отмечено, что постепенно снижается интерес к публикациям в области концепции цифрового двойника и работам, посвященным оценке места цифрового двойника в Индустрии 4.0.

Полный текст статьи размещен по ссылке.

Data – рецензируемый журнал открытого доступа, выпускаемый издательством MDPI (Швейцария). Согласно классификатору ASJC Scopus, относится к предметным областям «Прикладная наука о компьютерах», «Информационные системы», «Информационный менеджмент и системы». В журнале публикуются статьи, посвященные методам сбора, обработки и анализа научных данных, а также описания наборов данных. Журнал выпускается ежемесячно онлайн. Индексируется в базах данных Scopus, ESCI (Web of Science), dblp, Inspec, RePEc и других. Рейтинг журнала: CiteScore – Q2 (Информационные системы и менеджмент).

Издательство Multidisciplinary Digital Publishing Institute (MDPI) входит в число самых крупных мировых издательств и объединяет более 390 академических журналов, индексируемых в международных базах цитирования Scopus и Web of Science. MDPI также является крупнейшим мировым издательством журналов с открытым доступом.