DOI: 10.17323/2587-814X.2021.1.30.46

Динамика исследований в области интеллектуального анализа данных: тематический анализ публикаций за 20 лет

Ю.А. Зеленков
E-mail: yzelenkov@hse.ru

Е.А. Анисичкина
E-mail: eaanisichkina@edu.hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

Аннотация

Цель работы состоит в выявлении тенденций развития интеллектуального анализа данных (data mining) как научной дисциплины. На основе метода латентного размещения Дирихле (latent Dirichlet allocation, LDA) построена тематическая модель трудов Международной конференции по интеллектуальному анализу данных (International Conference on Data Mining, ICDM) за 2001–2019 годы и выделено девять основных направлений (тем) исследований. Для каждой темы исследованы динамика ее популярности (количество публикаций) и влияния (количество цитирований). Центральная тема, которая объединяет все прочие направления, – это общие вопросы обучения (general learning), включающее алгоритмы машинного обучения. За рассматриваемый период 20% усилий научного сообщества было потрачено на развитие именно этого направления, однако в последнее время его влияние снижается. Также снижается внимание к таким темам, как обнаружение ассоциаций (pattern mining) и разделение объектов (segmentation), включая кластеризацию. В то же время растет популярность исследований, связанных с рекомендательными системами (recommender systems), анализ сетей различной природы, в том числе социальных (network analysis) и анализ поведения человека, в частности поведения потребителей (human behavior analysis), что, скорее всего, связано с увеличением доступности данных и практической ориентацией этих тем. Направление исследований, связанных с приложениями интеллектуального анализа данных (applications), также имеет тенденцию к росту. Две последние темы – анализ текстовой информации (text mining) и прогнозирование потоков данных (data streams) привлекают относительно постоянный интерес исследователей. Полученные результаты проливают свет на структуру и динамику интеллектуального анализа данных как научной дисциплины за последние двадцать лет. Они также свидетельствуют, что за последние пять лет сформировалась новая повестка, характеризующаяся сдвигом интереса от алгоритмов к практическим приложениям, влияющим на все аспекты человеческой деятельности.

Ключевые слова: интеллектуальный анализ данных; датамайнинг; тематический анализ; наукометрия.

Цитирование: Зеленков Ю.А., Анисичкина Е.А. Динамика исследований в области интеллектуального анализа данных: тематический анализ публикаций за 20 лет // Бизнес-информатика. 2021. Т. 15. № 1. С. 30–46. DOI: 10.17323/2587-814X.2021.1.30.46

Введение

Термин «интеллектуальный анализ данных» (data mining, DM) используется с 1960-х годов для описания поиска корреляций без выдвижения априорных гипотез [1]. Согласно широко принятому определению, которое сейчас используется во многих учебниках, интеллектуальный анализ данных – это извлечение неявной, ранее неизвестной и потенциально полезной информации из данных [2, 3]. В учебнике [4] интеллектуальный анализ данных определяется как комбинация трех концепций, к числу которых относятся:

✦ статистика, которая включает классические описательные модели и инструменты, например, степени свободы, *F*-статистики и *p*-значения, но не рассматривает вывод гипотез;

✦ большие данные, как общий термин для наборов данных любого размера, но с акцентом на большой объем, поскольку доступность данных влияет практически на все аспекты нашей жизни;

✦ машинное обучение, то есть инструменты для создания компьютерных программ, которые анализируют базы данных в поисках закономерностей или шаблонов [2].

Статистика и машинное обучение создают техническую основу интеллектуального анализа данных. Некоторые авторы также рассматривают DM как часть процесса извлечения знаний (knowledge data discovery). Этот процесс может включать такие методы, как предварительная обработка данных (очистка и интеграция), хранение данных, аналитическая обработка данных в режиме онлайн, кубы данных и т.д. [3].

Как следует из приведенных определений, интеллектуальный анализ данных – это научная дисциплина, сочетающая достижения в нескольких областях исследований. Структура любой научной дисциплины может быть представлена как набор эволюционирующих тем, то есть значимых неявных ассоциаций, скрытых во фрагментированных областях знаний. Динамика этих тем (например, изменение количества публикаций и их цитирования) отражает сдвиг интересов научного сообщества. В частности, изучение этой динамики позволяет определить наиболее актуальные направления исследований, имеющие место в настоящее время, и экстраполировать их на ближайшее будущее. Кроме того, понимание фундаментальной динамики интересов исследователей помогает определить место изучаемой дисциплины в общей совокупности человеческих знаний, ее взаимодействие с другими дисциплинами и вклад в научно-технический прогресс.

Традиционным методом изучения структуры научной дисциплины является обзор литературы. Однако из-за междисциплинарного характера интеллектуального анализа данных практически нет обзоров, рассматривающих DM как отдельную дисциплину. Тем не менее, следует отметить, что обзоры более узких тем, например, таких как рекомендательные системы, публикуются постоянно.

К. Янг и С. Ву в обзорной статье [5], опубликованной в 2006 году, констатировали, что интеллектуальный анализ данных достиг огромных успехов. Однако, они также отметили проблемы с эффективным и своевременным обменом важными темами в научном сообществе. Кроме того, авторы выделили десять наиболее важных задач DM [5]:

✦ разработка единой теории интеллектуального анализа данных;
✦ масштабирование для данных высокой размерности и высокоскоростных потоков данных;
✦ анализ последовательностей данных и временных рядов;
✦ извлечение сложных знаний из сложных данных;
✦ анализ графов;
✦ распределенный и мультиагентный анализ данных;
✦ анализ данных в биологии и экологии;
✦ процесс интеллектуального анализа данных;
✦ безопасность, конфиденциальность и целостность данных;
✦ работа с нестационарными, несбалансированными и чувствительными к стоимости ошибки данными.

Перечисленные задачи разделяют общий поток исследований DM на более мелкие и более сфокусированные сегменты. В 2010 году С. Ву предоставил дополнительные комментарии по этим вопросам [6], и они стали предметом обсуждения на специальной панели на 10-й Международной конференции по интеллектуальному анализу данных (ICDM).

К. Янг и С. Ву [5] рассматривали разработку единой теории DM как наиболее важную цель. Это должна быть теоретическая основа, объединяющая различные методы, разработанные для отдельных задач, включая кластеризацию, классификацию, ассоциативные правила и т.д., а также различные технологии интеллектуального анализа данных, такие как статистика, машинное обучение и системы баз данных. По мнению авторов, такая теория должна обеспечить основу для будущих исследований.

Отметим, что большинство перечисленных проблем относится к алгоритмам работы с новыми типами данных, которые стали актуальными в 2000-х годах (данные сверхвысокой размерности, высокоскоростные потоки данных, временные ряды, сети и другие сложные структуры). Кроме того, авторы работ [5, 6] отмечают экологическую информатику как важнейшую область приложений DM.

Помимо анализа критических проблем, в работе [6] представлен список наиболее важных тем интеллектуального анализа данных (*таблица 1*). Этот список составлен на основании опроса экспертов; следовательно, он может служить ссылкой на структуру научной дисциплины. Однако экспертный подход не позволяет определить количественные показатели, которые определяют относительную важность различных тем и их изменение с течением времени.

Таблица 1.

Десять главных тем интеллектуального анализа данных [6]

№	Тема
1	Классификация (включая C4.5, CART, kNN и Naive Bayes)
2	Статистическое обучение (SVM и mixture models)
3	Анализ ассоциаций
4	Обнаружение связей (например, алгоритм PageRank)
5	Кластерный анализ
6	Бэджинг и бустинг
7	Шаблоны последовательностей
8	Интегрированный анализ (например, объединение классификации и ассоциативных правил)
9	Приближенные множества (rough sets)
10	Анализ графов

В работе [7] представлен обзор литературы по методам и приложениям интеллектуального анализа данных с января 2000 г. по август 2011 г. Авторы выбрали 216 статей из 159 научных журналов, используя такие ключевые слова, как “data mining”, “decision tree”, “artificial neural network”, “clustering” и т.д. На основании выбранных документов они определили девять категорий методов DM (оптимизация систем, системы, основанные на знаниях, моделирование, архитектура алгоритмов, нейронные сети и т.д.). Также авторы [7] выявили основные тенденции в области интеллектуального анализа данных. Согласно представленным результатам, наиболее важной тенденцией являются ассоциативные правила (ранг 5), за ними следуют нейронные сети (ранг 4), а затем – классификация и метод опорных векторов (оба направления имеют ранг 3). Авторы не описывают метод ранжирования, однако можно предположить, что он основан на подсчете количества ссылок на каждую тему в анализируемом корпусе публикаций.

Насколько нам известно, процитированные выше публикации – единственные, в которых исследуется динамика интеллектуального анализа данных как единой научной дисциплины. Как уже отмечалось, они основаны на субъективных (экспертных) оценках.

Идея нашей работы – применить формальные методы тематического анализа к публикациям в области DM. В качестве объекта анализа мы используем материалы Международной конференции по интеллектуальному анализу данных (International Conference on Data Mining, ICDM), которая проводится ежегодно с 2001 года.

1. Данные

Международная конференция по интеллектуальному анализу данных (International Conference on Data Mining, ICDM) – это ведущая конференция, которая, наряду с SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM International Conference on Web Search and Data Mining (WSDM) и некоторыми другими, образует сеть основных форумов в области интеллектуального анализа данных. База данных Web of Science (WoS) содержит информацию о 5120 публикациях основных треков ICDM и связанных с ними се-

минаров за 2001–2019 годы. На *рисунке 1* показано распределение этих публикаций по времени.

База данных WoS содержит информацию, необходимую для нашего исследования, включая сведения об авторах, названия публикаций, аннотации и количество цитирований.

2. Метод исследования

Одним из наиболее популярных методов анализа библиографических сетей является установление связей между термами пропорционально частоте их совместного появления в документах (term-level coupling), реализованное в программном пакете VOSviewer [8]. Этот подход позволяет идентифицировать кластеры слов, которые можно рассматривать как более или менее устойчивые неявные структуры, формирующие научную дисциплину. Авторы обзора [9] перечисляют основные вычислительные методы, автоматизирующие процесс обнаружения знаний в публикациях. Они отмечают, что тематическое моделирование, которое позволяет наблюдать как информация на уровне темы распространяется среди документов, обеспечивает более глубокое понимание корпуса документов, чем анализ на уровне терминов. Однако тематическое моделирование все еще относительно редко используется при анализе научной литературы [9, 10].

В статье [11] используется сочетание тематического моделирования и анализа цитирования для оценки динамики влияния тем и тематического разнообразия публикаций по информатике. Авторы статьи [12] использовали тематическое моделирование для анализа содержания материалов Международной конференции по принципам и практике многоагентных систем (International Conference on Principles and Practice of Multi-Agent Systems, PRIMA). Среди недавних публикаций, в работе [10] тематический анализ применяется к области управления знаниями. В последней статье особое внимание уделяется динамике тем, то есть тому, как количество публикаций и цитирований по каждой теме меняется во времени. Это позволяет пролить свет на изменение исследовательского интереса и выявить критические тенденции настоящего времени.

Еще одно приложение тематического анализа представлено в работе [13], где он используется для количественной оценки сходства и эволюции научных дисциплин. В работе [14] авторы предлагают эволюционную тематическую модель на базе деревьев, генерируемых на основе неоднородной библиографической сети.

Определим тему как множество слов, которые часто совместно встречаются в текстах, относящихся к определенной предметной области. Вероятностное тематическое моделирование основывается на идее, что документы представляют собой смесь тем, а каждая тема представляет собой распределение вероятностей по терминам.

Пусть корпус документов D содержит множество тем T, которое неизвестно. Каждое использование

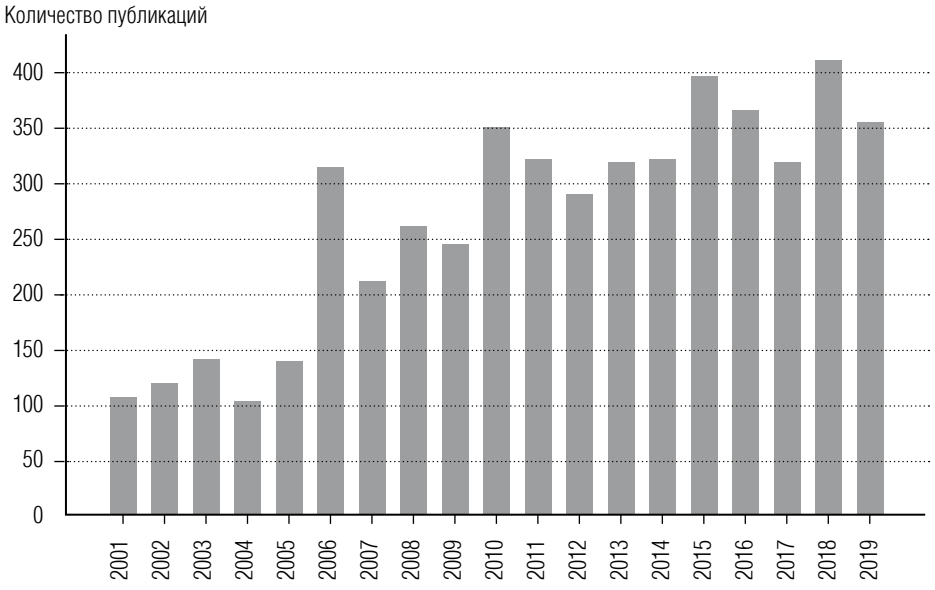

Рис. 1. Количество публикаций в трудах ICDM

термина w в документе d ассоциируется с некоторой темой $t \in T$. Таким образом, коллекцию документов можно рассматривать как множество триплетов (d, w, t), выбранных случайно и независимо из распределения, определенного на конечном множестве $D \times W \times T$. Документы $d \in D$ и термины $w \in W$ – наблюдаемые переменные. Темы $t \in T$ – это скрытые переменные, которые необходимо определить.

Тематическая модель автоматически определяет скрытые темы по наблюдаемым частотам слов в документах:

$$p(w|d) = \sum_{t \in T} p(t|d)\, p(w|t).$$

Входными данными алгоритма является матрица $D \times W$, ячейки которой содержат количества слов w в документе d.

Для того, чтобы построить матрицу $D \times W$, мы использовали аннотации 5120 статей, загруженных из базы данных Web of Science и описанных в предыдущем разделе. Согласно [15], различия между данными, подготовленными на основе аннотаций и полных текстов статей, более очевидны в небольших (размером несколько сотен) коллекциях документов. Поэтому в качестве объекта анализа мы выбрали аннотации.

В соответствии с общей методикой интеллектуального анализа текста, аннотации были токенизированы, а полученные термы преобразованы в стандартную форму. Затем были удалены слова, принадлежащие расширенному списку стоп-слов. Расширенный список стоп-слов включает стандартные английские стоп-слова и специфические для корпуса слова, которые встречаются менее чем в 5% и более чем в 60% документов. Мы также создали биграммы для объединения термов, которые часто встречаются рядом. В результате мы получили разреженную матрицу $D \times W$ размером 5120 × 1000, только 1,62% ячеек которой содержат значения, отличные нуля.

Для вычисления тем мы использовали алгоритм латентного размещения Дирихле (latent Dirichlet allocation, LDA), который основан на дополнительном предположении, что распределение Θ документов θ_d и распределение Φ тем φ_t порождаются распределениями Дирихле [16]. Для построения модели следует определить количество тем $|T|$; алгоритм LDA вычисляет распределения Θ и Φ. В результате каждая тема представлена взвешенным списком слов, а вес слова соответствует его важности в определении темы. Каждый документ представляется взвешенным списком тем, а вес темы соответствует ее значимости в документе.

Определение количества тем – ключевой вопрос тематического анализа. Многие авторы используют различные виды поиска по сетке для оптимизации определенной метрики [10]. В нашем исследовании использовалась байесовская оптимизация [17]. Этот подход позволяет оптимизировать одновременно не только количество тем, но и параметры распределений Θ и Φ, а также другие параметры алгоритма. Оптимизируемая метрика при этом – перплексия $P(D)$, которая измеряет сходимость модели с заданным словарем W:

$$P(D) = \exp\left[-\frac{1}{N} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \right].$$

Перплексия коллекции D является мерой качества языка и часто используется в компьютерной лингвистике. В нашем случае язык – это распределение слов в документах $p(w|d)$. Чем меньше перплексия, тем меньше вероятность, что это распределение является случайным.

В целом представленный подход удовлетворяет рекомендациям для пользователей LDA, представленным в работе [18].

Дополнительным показателем, который мы использовали для оценки качества модели, является разнообразие, то есть энтропия распределения слов, характеризующих тему:

$$H_t = -\frac{1}{\ln n_w} \sum_{i}^{n_w} p_t(w_i) \ln p_t(w_i), \qquad (1)$$

где n_w – количество слов, описывающих тему;

$p_t(w_i)$ – вес i-го слова в теме t.

Поскольку значение этой метрики нормализовано по количеству признаков (слов), ее возможные значения находятся в интервале [0; 1]. Значение 0 соответствует максимальной сфокусированности, когда тема описывается единственным словом. Значение 1 определяет ситуацию, когда все признаки, представленные в описании темы, имеют одинаковые веса, т.е. она не идентифицирована. Очевидно, что в более-менее пригодной для использования модели значения этой метрики должны быть небольшими и примерно одинаковыми для всех тем.

Когда оптимальное количество тем и соответствующее распределение тем для каждого документа найдено, мы можем изучать динамику тем. Пусть θ_{dt} – это вес темы t в документе d ($0 \le \theta_{dt} \le 1$). Таким

образом, общую популярность темы в корпусе документов можно определить следующим образом [10]:

$$\hat{\theta}_t = \frac{1}{|D|}\sum_{d\in D}\theta_{dt}. \quad (2)$$

Для вычисления популярности темы в заданный год y достаточно в формуле (2) положить $D = D_y$, где D_y – множество статей, опубликованных в год y.

Пусть C_d – количество цитирований документа d и $C = \sum_{d\in D} C_d$. Тогда влияние темы можно определить следующим образом [10]:

$$\hat{i}_t = \frac{1}{C}\sum_{d\in D}\theta_{dt}C_d \ . \quad (3)$$

Аналогично, чтобы вычислить влияние темы в год необходимо установить $D = D_y$ в уравнении (3).

3. Результаты и обсуждение

После выполнения всех операций предварительной обработки, описанных в предыдущем разделе, и 100 итераций байесовской оптимизации модели LDA, было найдено оптимальное количество тем, равное 9. Соответствующее значение перплексии составило 568,75.

Анализируя доминирующие термины по темам (*рисунок 2*), можно сделать вывод, что каждая тема представляет собой некоторую согласованную область исследования. При этом веса тем в документах либо велики (т.е. документ сильно связан с темой), либо близки к нулю (т.е. документ не относится к данной теме).

Чтобы определить краткие наименования тем, мы проанализировали распределение терминов и наиболее репрезентативные статьи по каждой теме. Чтобы выбрать наиболее репрезентативные статьи, мы отсортировали публикации по весу темы, а затем по количеству цитирований, обе сортировки в порядке убывания. На *рисунке 2* представлены присвоенные краткие наименования, а в *таблице 2* приведено описание тем.

В *таблице 2* также представлены значения разнообразия, популярности и влияния для каждой темы во всей коллекции D, рассчитанные по формулам (1), (2) и (3) соответственно. Поскольку сумма как

Рис. 2. Визуализация тематической модели с помощью облака слов
(каждое облако представляет одну тему, размер шрифта слова соответствует его весу в данной теме)

Таблица 2.

Темы интеллектуального анализа данных

Тема	Комментарии	Разнообразие	Популярность	Влияние
Text mining	Обнаружение паттернов в текстах	0,779	0,107	0,110
General learning	Алгоритмы машинного обучения и связанные методы, такие как отбор признаков, разметка классов и др.	0,826	0,213	0,211
Segmentation	Методы разделения объектов: кластеризация, обнаружение выбросов и т.д.	0,777	0,084	0,080
Applications	Практическое использование методов DM	0,826	0,097	0,095
Data streams	Модели данных, зависящих от времени	0,805	0,097	0,102
Recommender systems	Рекомендательные системы	0,799	0,076	0,079
Pattern mining	Общие вопросы поиска корреляций между элементами данных	0,750	0,110	0,114
Network analysis	Обнаружение сообществ и потоков влияния в различных сетях	0,762	0,093	0,111
Human behavior analysis	Выявление и прогнозирование закономерностей в поведении людей: отток клиентов, сегментация рынка, мошенничество, угрозы безопасности и т.д.	0,844	0,121	0,096

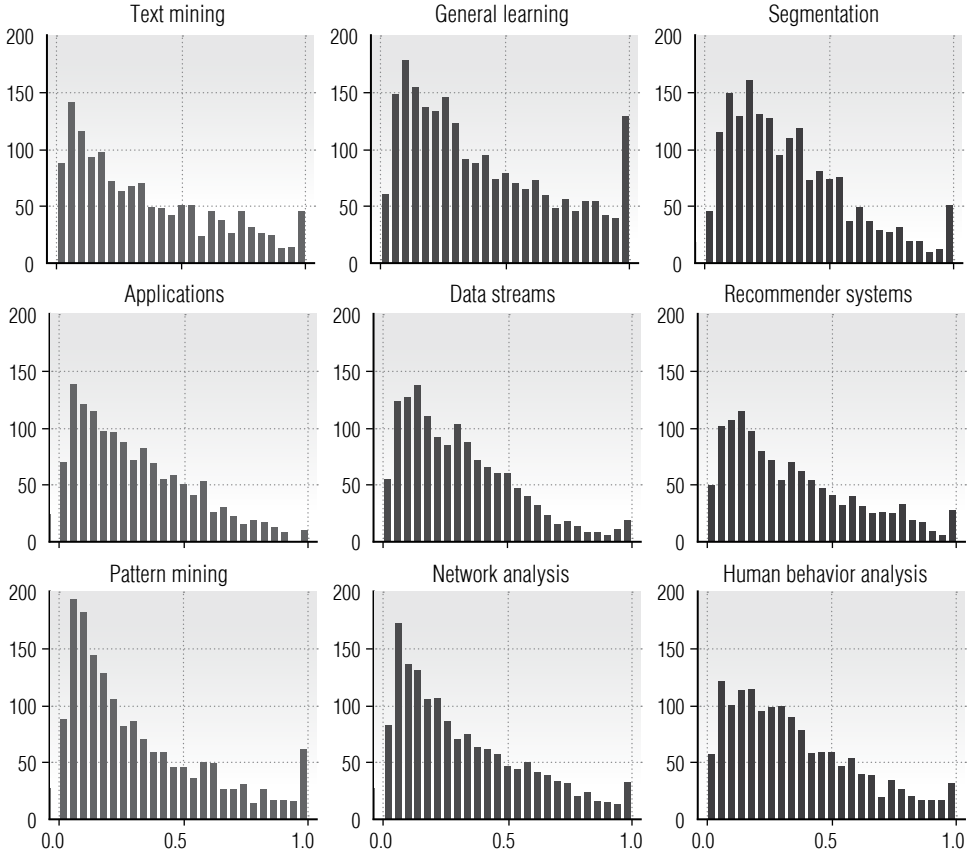

Рис. 3. Распределение весов тем по корпусу документов
(горизонтальная ось – веса тем, вертикальная – соответствующее количество документов)

популярности, так и воздействия равна единице, представленные значения можно рассматривать как долю определенной темы в общем потоке исследований интеллектуального анализа данных, то есть ее общий вес.

Дополнительную информацию о структуре DM можно получить из анализа распределения весов тем в корпусе документов (*рисунок 3*). Для более наглядного представления мы исключили из графика для каждой темы документы, в которых вес этой темы равен нулю.

Как следует из *таблицы 2*, тема, которая привлекала наибольшее внимание за последние 20 лет, – это общие вопросы обучения (general learning). На развитие этого направления потрачено более 20% усилий исследователей в области интеллектуального анализа данных ($\hat{\theta}_t = 0{,}213$). Работы в этой области охватывают широчайший спектр проблем машинного обучения, например, отбор признаков [19] (вес доминирующей темы в этом документе $\theta_{dt} = 0{,}974$), мульти-классовую классификацию с использованием ансамблей [20] ($\theta_{dt} = 0{,}963$), градиентные методы [21] ($\theta_{dt} = 0{,}925$) и т.д. Эти примеры статей выбраны, поскольку все они имеют большое количество цитирований (более 90, согласно WoS). Интересно, что работы с максимальным весом этой темы в основном посвящены методам на основе ядер (например, [22]) с $\theta_{dt} = 0{,}990$. Согласно *рисунку 3*, эта тема имеет вес, близкий к единице в наибольшем количестве документов (более 100). Эти статьи посвящены исключительно методам машинного обучения и не пересекаются с другими тематиками DM.

Мы определили вторую по важности тему ($\hat{\theta}_t = 0{,}121$) как «Анализ поведения человека» (human behavior analysis). Она фокусируется на обнаружении и прогнозировании закономерностей в деятельности групп людей и систем, на которые эти группы влияют. В этой области изучаются такие вопросы, как продвижение за счет снижения цен [23] ($\theta_{dt} = 0{,}988$), обнаружение подозрительных финансовых транзакций [24] ($\theta_{dt} = 0{,}985$), волатильность биткойна [25] ($\theta_{dt} = 0{,}985$) и другие. Отметим, что значительная часть этих работ представлена на семинарах, сопровождающих основную конференцию.

Следующая тема, – обнаружение ассоциаций (pattern mining), – фокусируется на извлечении ассоциаций (правил), то есть на задаче поиска корреляций между элементами в наборе данных. Исследователи изучают как практическое применение ассоциативных правил (например, данных о рыночной корзине), так и общие принципы обнаружения связей в больших базах данных. С одной стороны, это может быть идентификация наиболее часто встречающихся комбинаций элементов данных [26] ($\theta_{dt} = 0{,}960$). С другой стороны, это может быть паттерн, состоящий из редких, но сильно коррелированных элементов [27] ($\theta_{dt} = 0{,}987$). Отметим также, что это наиболее сфокусированная тема с наименьшим значением H_t.

В наиболее представительных публикациях по теме анализа текстовой информации (text mining) рассматриваются такие вопросы, как идентификация и ранжирование авторов [28] ($\theta_{dt} = 0{,}974$), тематическое моделирование [29] ($\theta_{dt} = 0{,}969$) и кластеризация текста с использованием моделей на основе семантики [30] ($\theta_{dt} = 0{,}988$). Это область исследований с четкими границами, которая включает обнаружение закономерностей только в текстах и не рассматривает другие типы неструктурированных данных.

Направление исследований, который мы определили как потоки данных (data streams), касается моделей, зависящих от времени. Он включает в себя как более или менее традиционный анализ и прогнозирование временных рядов [31] ($\theta_{dt} = 0{,}981$), так и более экзотические модели, например, основанные на причинности по Грейнджеру [32] ($\theta_{dt} = 0{,}987$).

Тема «Приложения» (applications) объединяет работы, в основном посвященные практическому использованию методов интеллектуального анализа данных, которые не относятся к другим направлениям, выделенным выше. Примерами являются обнаружение событий с использованием со-локации мобильных пользователей [33] ($\theta_{dt} = 0{,}987$) и биометрическая модель безопасности для медицинского интернета вещей [34] ($\theta_{dt} = 0{,}983$).

Направление исследований «Сетевой анализ» (network analysis) посвящено моделям на основе графов, позволяющим восстановить пространственную структуру или топологию исследуемого объекта. Наиболее популярной темой такого рода исследований является обнаружение сообществ с использованием различных методов сетевого анализа [35] ($\theta_{dt} = 0{,}983$). Следующим вопросом, внимание к которому в последнее время возрастает, является анализ потоков влияния [36] ($\theta_{dt} = 0{,}985$), в том числе прогнозирование популярности сообщений в социальных сетях.

Мы определили следующую тему как разделение объектов (segmentation), поскольку она включает не

только широкий спектр алгоритмов кластеризации [37] (θ_{dt} = 0,977), но и приложения, основанные на методах разделения объектов, например, обнаружение выбросов [38] (θ_{dt} = 0,981). Согласно нашей модели, эта тема преобладает в работах, связанных с неструктурированными данными (изображения, видео, звук). Однако в большинстве случаев вес этой темы в таких приложениях не превышает веса других тем.

Наконец, последняя, но не менее важная тема, обнаруженная нашей моделью, – это рекомендательные системы (recommender systems). Это направление в дополнительных комментариях не нуждается. Стоит только отметить, что в последнее время исследователи уделяют особое внимание генерации интерпретируемых рекомендаций [39] (θ_{dt} = 0,986).

Как следует из *рисунка 3*, помимо "general learning", только в трех других областях существует относительно большое количество статей с весом темы близким к единице: это "segmentation", "pattern mining" и "text mining". Эти направления также относятся к машинному обучению, поэтому здесь публикуется относительно большое количество статей, посвященных исключительно алгоритмам. С другой стороны, такая тема, как "applications", практически не содержит статей с весом темы близким к единице. Данная тема также имеет высокое значение метрики разнообразия (1). Этот факт можно объяснить тем, что в работах, посвященных практическим приложениям DM, как правило, представляются и новые модификации алгоритмов.

Наша модель не выделяет искусственные нейронные сети (ИНС) как отдельное направление интеллектуального анализа данных. Это противоречит [7], но согласуется с работами [5, 6]. Согласно нашим результатам, публикации, в которых используются модели ИНС, чаще всего относятся к областям "general learning" и "segmentation".

Следующий вопрос, который необходимо рассмотреть, это взаимодействие тем, которое можно рассматривать как частоту совместного появления тем в документах. Пусть θ_{di} и θ_{dj} – веса тем i и j в документе d. Таким образом, мы можем определить взаимодействие тем в этом документе как произведение $\theta_{di}\theta_{dj}$. Максимальное возможное значение взаимодействия двух тем в документе $\theta_{di}\theta_{dj}$ = 0,25 при $\theta_{di} = \theta_{dj}$ = 0,5. Отсюда максимальное возможное значение взаимодействия тем в коллекции документов $0{,}25 \cdot |D|$.

Таким образом, взаимодействие двух тем i и j по корпусу документов может быть вычислено как

$$c_{ij} = \sum_{d \in D} \theta_{di}\theta_{dj}.$$

Соответствующие данные представлены на *рисунке 4*. "General learning" можно рассматривать

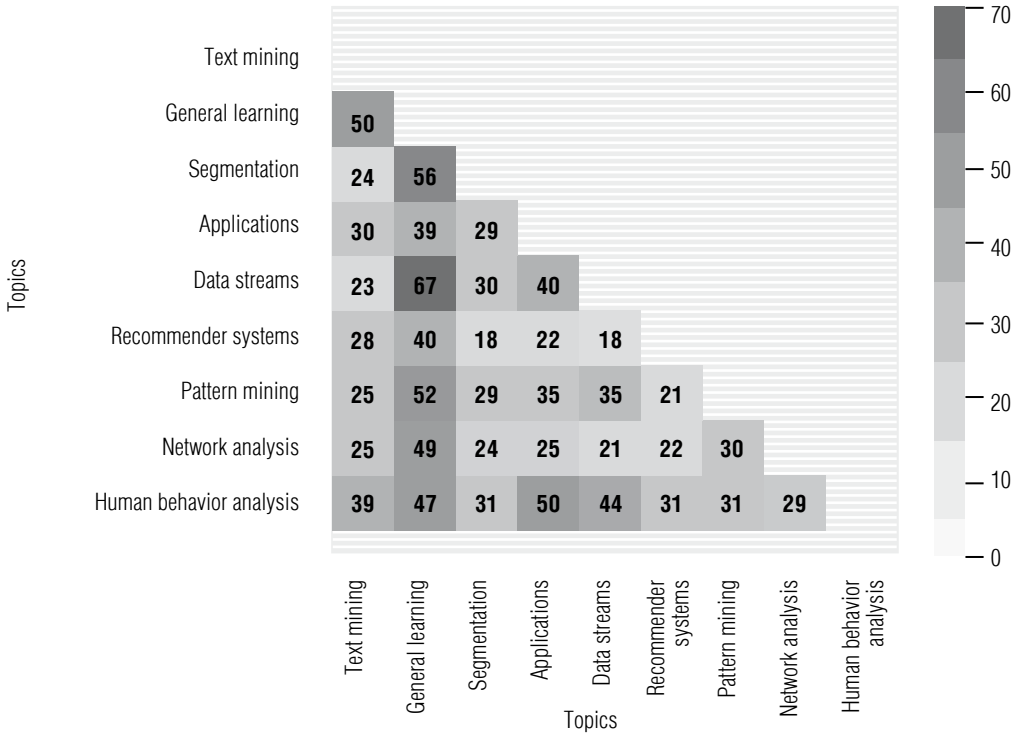

Рис. 4. Взаимодействие тем

как центральную тему, поскольку она наиболее тесно связана с другими областями исследований. “Human behavior analysis” также имеет относительно сильные связи с другими направлениями. “Recommender systems”, “network analysis” и “text mining” – более изолированные темы в нашей модели, поскольку они основаны на специализированных алгоритмах.

Следующий этап анализа – исследование динамики популярности и влияния выявленных тем. Популярность является производной от количества публикаций, а влияние рассчитывается с использованием количества цитирований. На *рисунке 5* представлена динамика популярности тем (сплошная линия) согласно выражению (2). Пунктирная линия показывает тренд. На *рисунке 6* представлены те же данные по влиянию тем, рассчитанному по формуле (3). Эти данные отражают дрейф интересов сообщества интеллектуального анализа данных в каждой области исследований.

Отметим, что популярность и влияние многих тем подвержены значительным флуктуациям. С одной стороны, это можно объяснить кратковременным смещением внимания исследователей. С другой стороны, ICDM, хотя и является одним из наиболее представительных форумов в области интеллектуального анализа данных, может не полностью отражать реальную динамику этой дисциплины. Например, формулировка названий треков конференции программным комитетом (call for papers) может повлиять на исследователей, представляющих работы. Однако мы считаем, что, несмотря на обнаруженные флуктуации, анализ публикаций за 2001–2019 гг. позволяет выявить глобальные тенденции, что и является целью нашего исследования.

Представленные данные показывают, что внимание исследователей к “pattern mining” резко снижается как с точки зрения популярности, так и воздействия. Такая же, но менее выраженная тенденция характерна и для “general learning” и “segmentation”.

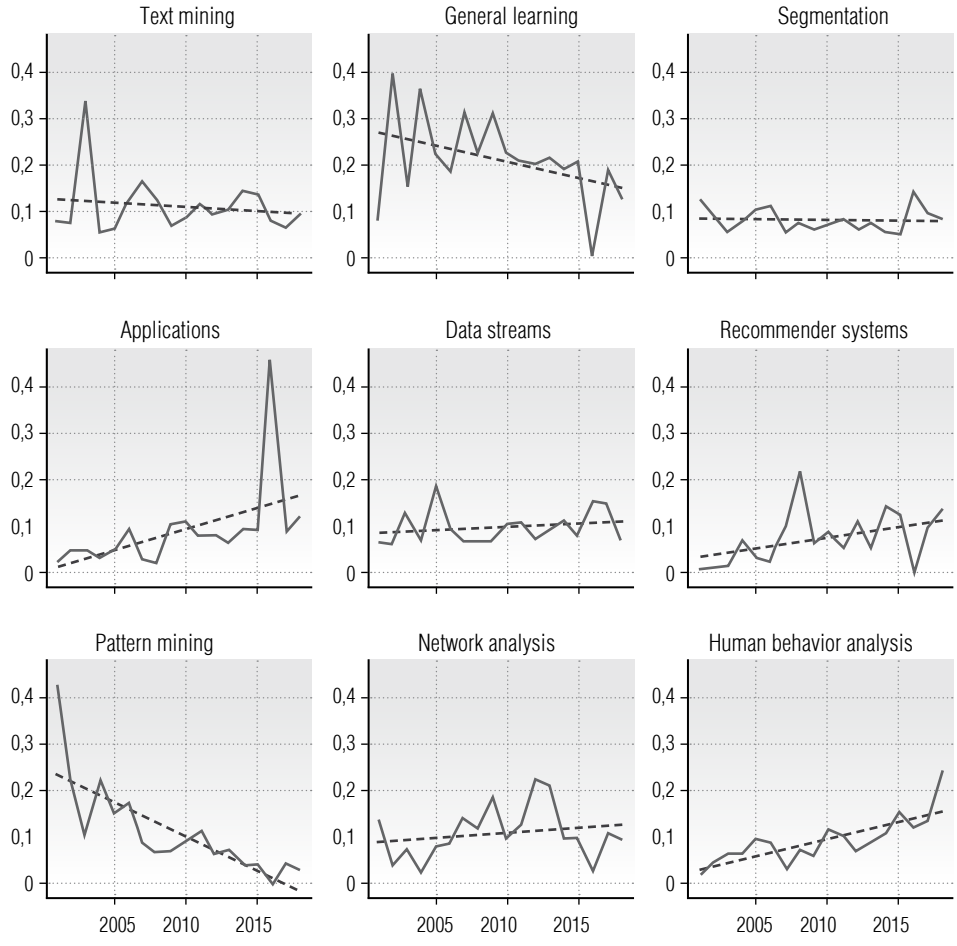

Рис. 5. Динамика популярности тем (сплошная линия) и тренд (пунктирная линия)

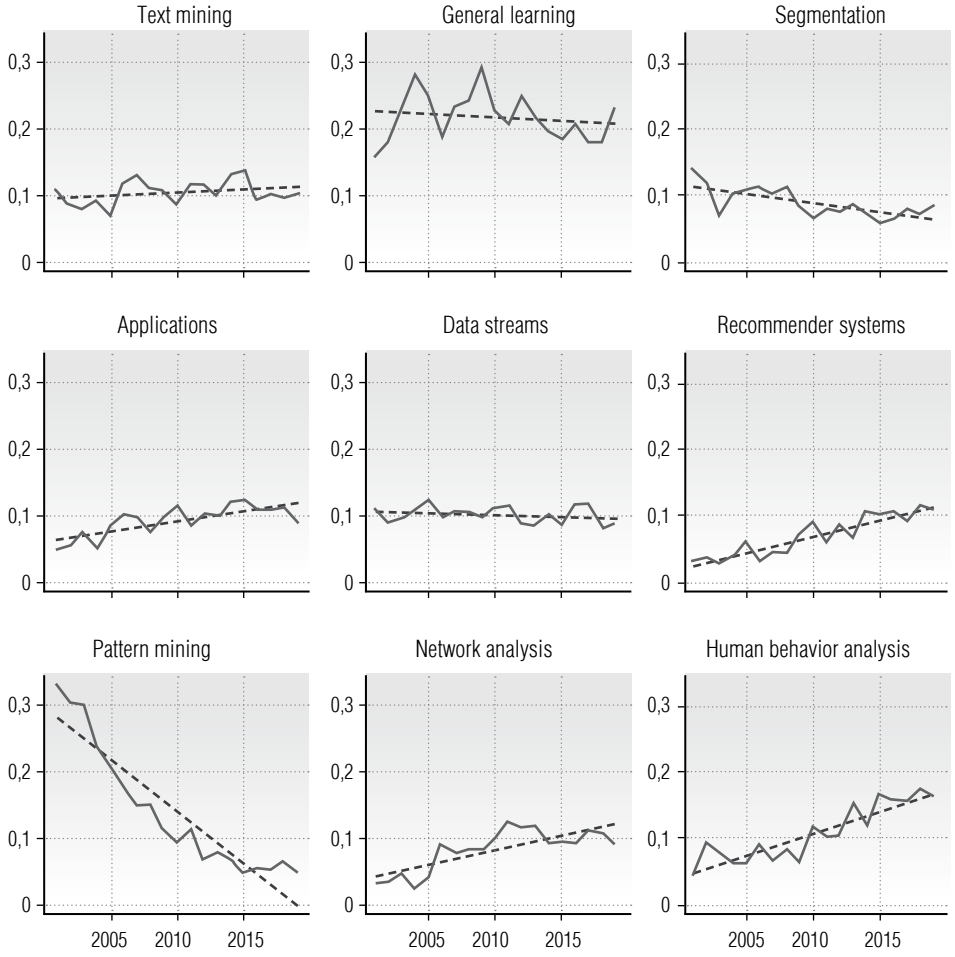

Рис. 6. Динамика влияния тем (сплошная линия) и тренд (пунктирная линия)

Эти тенденции требуют более глубокого изучения. Во-первых, можно предположить, что это связано с тем, что достижения в области алгоритмов машинного обучения и связанных с ними методов, включая анализ ассоциаций, уже настолько велики, что дальнейшее продвижение требует серьезных усилий. Сегодня наибольшая активность наблюдается в области глубокого обучения, но объем таких публикаций в трудах ICDM еще относительно невелик.

В то же время растет популярность исследований, связанных с рекомендательными системами (recommender systems), сетевым анализом (network analysis) и анализом поведения человека (human behavior analysis), что, скорее всего, связано с увеличением доступности данных и практической направленностью этих тем. Направление исследований, связанных с практическим применением интеллектуального анализа данных (applications), также имеет тенденцию к росту. Это также может объяснить снижение интереса к основополагающим алгоритмам, поскольку значительная часть исследовательского сообщества сосредоточена на более актуальных практических вопросах.

Последние две темы, “text mining” и “data streams”, вызывают относительно постоянный интерес исследователей. Представленные результаты проливают свет на структуру и динамику интеллектуального анализа данных за последние двадцать лет и позволяют расширить понимание этой научной дисциплины.

Следующий вопрос, который представляет интерес при анализе содержания публикаций, – это тематическое разнообразие документов. По аналогии с (1) мы можем определить разнообразие документа через энтропию его тем:

$$H_d = -\sum_i^T \theta_{di} \ln \theta_{di} \ ,$$

где θ_{di} – вес темы i в документе d;

T – количество тем.

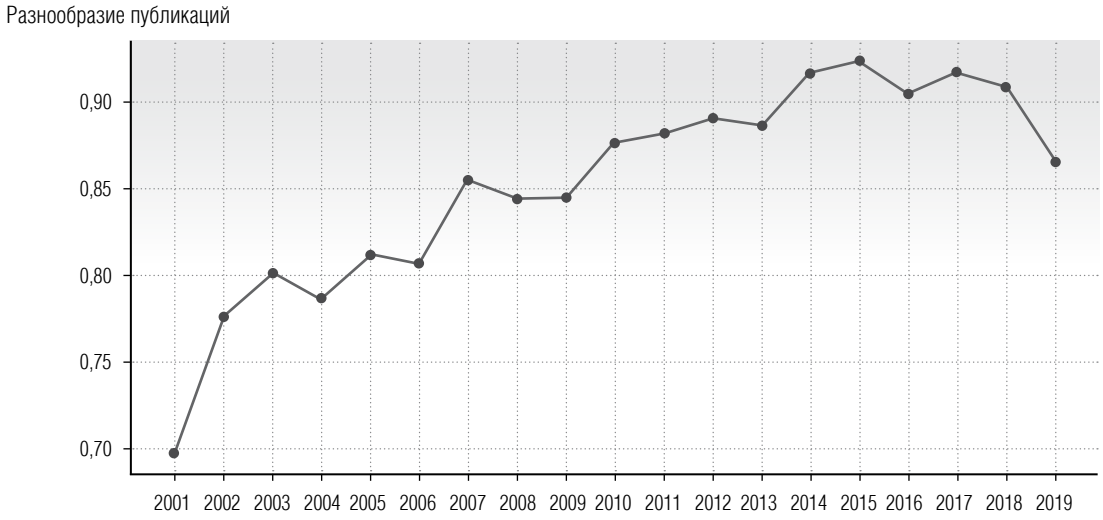

Рис. 7. Разнообразие публикаций ICDM в 2001–2019

На *рисунке 7* представлено среднее разнообразие публикаций в трудах ICDM за 2001–2019 гг. Мы видим, что тематическое разнообразие документов неуклонно росло с момента первой конференции и достигло пика в 2015 году. За последние четыре года среднее количество тем, охватываемых одним документом, сократилось.

Мы полагаем, что это можно объяснить следующим образом. В начале 2000-х годов основной интерес исследователей был сосредоточен на алгоритмах обнаружения знаний, которые представлены списком критических направлений, выделенных в [5] и подтвержденных в [6] (*таблица 1*). По мере развития этих алгоритмов область их практического применения расширялась. Следовательно, и набор тем, освещаемых в одной научной публикации, становился все более широким. Это можно рассматривать как поиск в тематическом пространстве, пик которого пришелся на 2015 год. После 2015 года сформировалась новая исследовательская повестка дня. Как показано выше, алгоритмы общего обучения, а также связанные с ними области, такие как идентификация паттернов и сегментация, отодвигаются на второй план, хотя и продолжают играть важную роль. На первый план выходят более практические приложения, связанные с анализом человеческого поведения, системами рекомендаций, анализом сетевых сообществ и т.д.

В *таблице 3* представлено сравнение тем интеллектуального анализа данных, выделенных в нашей работе и в статьях [5 ,6]. Большинство тем 2010 года сосредоточено в направлении “general learning”. Мы также включили в эту категорию приближенные множества (rough sets), поскольку данная теория применяется к классическим задачам извлечения знаний, например, для обнаружения закономерностей в неполных данных [40].

Таблица 3.

Соответствие тем интеллектуального анализа данных

Темы DM в 2020 г. (настоящая работа)	Темы DM в 2005-2010 гг. [5, 6]
Text mining	Анализ связей (например, алгоритм PageRank)
General learning	Классификация
	Статистическое обучение
	Бэджинг и бустинг
	Интегрированный анализ
	Приближенные множества
Segmentation	Кластеризация
Applications	Не представлено
Data streams	Шаблоны последовательностей
Recommender systems	Не представлено
Pattern mining	Анализ ассоциаций
Network analysis	Анализ графов
Human behavior analysis	Не представлено

Новые темы, стремительный рост которых обнаружила наша модель, в 2010 году вообще не рассматривались. Отметим также, что выделенная

в нашей работе тема “text mining” включает гораздо больший спектр технологий и приложений, чем поиск взаимосвязей между документами.

В середине 2010-х годов исследователи в области экономики и управления зафиксировали рост интереса к принятию решений на основе данных (data driven decision-making, DDD) [41]. Это практика принятия решений, опирающаяся на анализ данных, а не исключительно на человеческие знания и интуицию. Авторы работы [42] сообщают, что использование DDD в производственном секторе США за период с 2005 по 2010 гг. почти утроилось (с 11% до 30% компаний). Более поздние исследования подтверждают возрастающую роль DDD как одной из лучших практик управления [43].

Определение DDD, которое приводится в литературе по менеджменту, полностью совпадает с определением “data mining”, рассмотренным в начале нашей работы. DDD также включает инжиниринг и процессинг данных, а также обнаружение полезных шаблонов. Однако, если DM рассматривает эту деятельность с технологической точки зрения, то DDD подходит к данному вопросу в контексте организационных процессов, включая деятельность, осуществляемую исключительно человеком. Тем не менее, мы можем рассматривать растущий интерес к DDD как один из ключевых факторов, способствующих сдвигу в исследованиях DM, показанному на *рисунке 7.*

Заключение

Мы представили исследование интеллектуальной структуры “data mining” как научной дисциплины, проведенное с использованием тематического анализа. Такой подход позволил выделить девять основных направлений интеллектуального анализа данных и изучить их динамику.

Главный результат работы заключается в том, что мы обнаружили смещение интересов от алгоритмов машинного обучения к более практическим приложениям. По нашим данным, такая смена фокуса наметилась в середине 2010-х годов. Мы объясняем этот сдвиг сочетанием трех факторов. Во-первых, базовые алгоритмы интеллектуального анализа данных достигли высокого уровня зрелости. Во-вторых, с развитием социальных сетей стало доступно большое количество данных. В-третьих, существует устойчивый спрос со стороны бизнеса на принятие решений на основе данных. ■

Литература

1. Piatetsky-Shapiro G., Fayyad U. An introduction to SIGKDD and a reflection on the term ‘data mining’ // ACM SIGKDD Explorations Newsletter. 2012. Vol. 13. No 2. P. 102–103. DOI: 10.1145/2207243.2207269.
2. Witten I.H., Frank E., Hall M., Pal C. Data mining: Practical machine learning tools and techniques. Cambridge, MA: Morgan Kaufmann, 2017.
3. Han J., Kamber M., Pei J. Data mining: Concepts and techniques. Waltham, MA: Morgan Kaufmann, 2012. DOI: 10.1016/C2009-0-61819-5.
4. Rather B. Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. Sound Parkway, NW: CRC Press, 2011.
5. Yang Q., Wu X. 10 challenging problems in data mining research // International Journal of Information Technology & Decision Making. 2006. Vol. 5, No 4. P. 597–604. DOI: 10.1142/S0219622006002258.
6. Wu X. 10 years of data mining research: retrospect and prospect // Proceedings of the 10th IEEE International Conference on Data Mining (ICDM). Sydney, Australia, 13–17 December 2010. P. 7. DOI: 10.1109/ICDM.2010.172.
7. Liao S.H., Chu P.H., Hsiao P.Y. Data mining techniques and applications – A decade review from 2000 to 2011 // Expert Systems with Applications. 2012. Vol. 39. No 12. P. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.
8. Van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics. 2009. Vol. 84. No 2. P. 523–538. DOI: 10.1007/s11192-009-0146-3.
9. Thilakaratne M., Falkner K., Atapattu T. A systematic review on literature-based discovery: general overview, methodology, & statistical analysis // ACM Computing Surveys. 2019. Vol. 52. No 6. Article no 129. DOI: 10.1145/3365756.
10. Zelenkov Y. The topic dynamics in knowledge management research // Proceedings of the 14th International Conference on Knowledge Management in Organizations (KMO 2019). Zamora, Spain, 15–18 July 2019. P. 324–335. DOI: 10.1007/978-3-030-21451-7_28.
11. Mann G.S., Mimno D., McCallum A. Bibliometric impact measures leveraging topic analysis // Proceedings of the 6th ACM/IEE Joint Conference on Digital Libraries (JCDL ’06). Chapel Hill, NC, USA, 11–15 June 2006. P. 65–74. DOI: 10.1145/1141753.1141765.
12. Dam H.K., Ghose A. Analyzing topics and trends in the PRIMA literature // Proceedings of the 19th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2016). Phuket, Thailand, 22–26 August 2016. P. 216–229. DOI: 10.1007/978-3-319-44832-9_13.
13. Dias L., Gerlach M., Scharloth J., Altman E.G. Using text analysis to quantify the similarity and evolution of scientific disciplines // Royal Society Open Science. 2018. Vol. 5. No 1. Article no 171545. DOI: 10.1098/rsos.171545.
14. Jensen S., Liu X., Yu Y., Milojevic S. Generation of topic evolution trees from heterogeneous bibliographic networks // Journal of Informetrics. 2016. Vol. 10. No 2. P. 606–621. DOI: 10.1016/j.joi.2016.04.002.

15. Syed S., Spruit M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation // Proceedings of the 4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017). Tokyo, Japan, 19–21 October 2017. P. 165–174. DOI: 10.1109/DSAA.2017.61.

16. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. No 3. P. 993–1022.

17. Mockus J. Bayesian approach to global optimization: Theory and applications. Heidelberg: Springer, 2012. DOI: 10.1007/978-94-009-0909-0.

18. Understanding the limiting factors of topic modeling via posterior contraction analysis / J. Tang [et al.] // Proceedings of Machine Learning Research. 2014. Vol. 32. No 1. P. 190–198.

19. Molina L.C., Belanche L., Nebot A. Feature selection algorithms – A survey and experimental evaluation // Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM 2002). Maebashi City, Japan, 9–12 December 2002. P. 306–313. DOI: 10.1109/ICDM.2002.1183917.

20. Read J., Pfahringer B., Holmes G. Multi-label classification using ensembles of pruned sets // Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). Pisa, Italy, 15–19 December 2008. P. 995–1000. DOI: 10.1109/ICDM.2008.74.

21. Chen X., Pan W., Kwok J.T., Carbonell J.G. Accelerated gradient method for multi-task sparse learning problem // Proceedings of the 9th IEEE International Conference on Data Mining (ICDM). Miami Beach, FL, USA, 6–9 December 2009. P. 746–751. DOI: 10.1109/ICDM.2009.128.

22. Shin K. Partitionable kernels for mapping kernels // Proceedings of the 11th IEEE International Conference on Data Mining (ICDM). Vancouver, BC, Canada, 11–14 December 2011. P. 645–654. DOI: 10.1109/ICDM.2011.115.

23. Li Z., Yada K. Why do retailers end price promotions – A study on duration and profit effects of promotion // Proceedings of the 15th IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ, USA, 14–17 November 2015. P. 328–335. DOI: 10.1109/ICDMW.2015.56.

24. Camino R.D., State R., Montero L., Valtchev P. Finding suspicious activities in financial transactions and distributed ledgers // Proceedings of the 17th IEEE International Conference on Data Mining Workshops (ICDMW). New Orleans, LA, USA, 18–21 November 2017. P. 787–796. DOI: 10.1109/ICDMW.2017.109.

25. Guo T., Bifet A., Antulov-Fantulin N. Bitcoin volatility forecasting with a glimpse into buy and sell orders // Proceedings of the 18th IEEE International Conference on Data Mining (ICDM). Singapore, 17–20 November 2018. P. 989–994. DOI: 10.1109/ICDM.2018.00123.

26. Gouda K., Zaki M.J. Efficiently mining maximal frequent itemsets // Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM). San Jose, CA, USA, 29 November – 2 December 2001. P. 163–170. DOI: 10.1109/ICDM.2001.989514.

27. Ma S., Hellerstein J.L. Mining mutually dependent patterns // Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM). San Jose, CA, USA, 29 November – 2 December 2001. P. 409–416. DOI: 10.1109/ICDM.2001.989546.

28. Zhou D., Orshanskiy S.A., Zha H., Gees C.L. Co-ranking authors and documents in a heterogeneous network // Proceedings of the 7th IEEE International Conference on Data Mining (ICDM). Omaha, NE, USA, 28–31 October 2007. P. 739-744. DOI: 10.1109/ICDM.2007.57.

29. Tang J., Jin R., Zang J. A topic modeling approach and its integration into the random walk framework for academic search // Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). Pisa, Italy, 15–19 December 2008. P. 1055–1060. DOI: 10.1109/ICDM.2008.71.

30. Shehata S., Karray F., Kamel M. Enhancing text clustering using concept-based mining model // Proceedings of the 6th IEEE International Conference on Data Mining (ICDM). Hong Kong, China, 18–22 December 2006. P. 1043–1048. DOI: 10.1109/ICDM.2006.64.

31. Yang D., Li B., Rettig L., Cudre-Mauroux P. HistoSketch: Fast similarity-preserving sketching of streaming histograms with concept drift // Proceedings of the 17th IEEE International Conference on Data Mining (ICDM). New Orleans, LA, USA, 18–21 November 2017. P. 545–554. DOI: 10.1109/ICDM.2017.64.

32. Dhurandhar A. Learning maximum lag for grouped graphical Granger models // Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDMW). Sydney, Australia, 13 December 2010. P. 217–224. DOI: 10.1109/ICDMW.2010.9.

33. Wang H., Li Z., Lee W.-C. PGT: Measuring mobility relationship using personal, global and temporal factors // Proceedings of the 14th IEEE International Conference on Data Mining (ICDM). Shenzhen, China, 14–17 December 2014. P. 570–579. DOI: 10.1109/ICDM.2014.111.

34. Pirbhulal S., Wu W., Li G. A biometric security model for wearable healthcare // Proceedings of the 18th IEEE International Conference on Data Mining Workshops (ICDMW). Singapore, 17–20 November 2018. P. 136–143. DOI: 10.1109/ICDMW.2018.00026.

35. Yang J., Leskovec J. Defining and evaluating network communities based on ground-truth // Proceedings of the 12th IEEE International Conference on Data Mining (ICDM). Brussels, Belgium, 10–13 December 2012. P. 745–754. DOI: 10.1109/ICDM.2012.138.

36. Shi L., Tong H., Tang J., Lin C. Flow-based influence graph visual summarization // Proceedings of the 14th IEEE International Conference on Data Mining (ICDM). Shenzhen, China, 14–17 December 2014. P. 983–988. DOI: 10.1109/ICDM.2014.128.

37. Hung M.-C., Yang D.-L. An efficient fuzzy c-means clustering algorithm // Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM). San Jose, CA, USA, 29 November – 2 December 2001. P. 225–232. DOI: 10.1109/ICDM.2001.989523.

38. Pei Y., Zaiane O.R., Gao Y. An efficient reference-based approach to outlier detection in large datasets // Proceedings of the 6th IEEE International Conference on Data Mining (ICDM). Hong Kong, China, 18–22 December 2006. P. 478–487. DOI: 10.1109/ICDM.2006.17.

39. A reinforcement learning framework for explainable recommendation / X. Wang [et al.] // Proceedings of the 18th IEEE International Conference on Data Mining (ICDM). Singapore, 17–20 November 2018. P. 587–596. DOI: 10.1109/ICDM.2018.00074.

40. Wang H., Wang S. Discovering patterns of missing data in survey databases: an application of rough sets // Expert Systems with Applications. 2009. Vol. 36. No 3. Part 2. P. 6256–6260. DOI: 10.1016/j.eswa.2008.07.010.

41. Provost F., Fawcett T. Data science and its relationship to big data and data-driven decision making // Big Data. 2013. Vol. 1. No 1. P. 51–59. DOI: 10.1089/big.2013.1508.

42. Brynjolfsson E., McElheran K. The rapid adoption of data-driven decision-making // American Economic Review. 2016. Vol. 106. No 5. P. 133–139. DOI: 10.1257/aer.p20161016.

43. Song P., Zheng C., Zhang C., Yu X. Data analytics and firm performance: An empirical study in an online B2C platform // Information & Management. 2018. Vol. 55, No 5. P. 633–642. DOI: 10.1016/j.im.2018.01.004.

Об авторах

Зеленков Юрий Александрович

доктор технических наук;

профессор департамента бизнес-информатики, Высшая школа бизнеса, Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Анисичкина Екатерина Алексеевна

студентка бакалавриата, образовательная программа «Бизнес-информатика», Высшая школа бизнеса,

Национальный исследовательский университет «Высшая школа экономики», 101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: eaanisichkina@edu.hse.ru

Trends in data mining research: A two-decade review using topic analysis

Yuri A. Zelenkov
E-mail: yzelenkov@hse.ru

Ekaterina A. Anisichkina
E-mail: eaanisichkina@edu.hse.ru

National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

Abstract

This work analyses the intellectual structure of data mining as a scientific discipline. To do this, we use topic analysis (namely, latent Dirichlet allocation, DLA) applied to the proceedings of the International Conference on Data Mining (ICDM) for 2001–2019. Using this technique, we identified the nine most significant research flows. For each topic, we analyse the dynamics of its popularity (number of publications) and influence (number of citations). The central topic, which unites all other direction, is General Learning, which includes machine learning algorithms. About 20% of the research efforts were spent on the development of this direction for the entire time under review, however, its influence has declined most recently. The analysis also showed that attention to topics such as Pattern Mining (detecting associations) and Segmentation (object separation algorithms such as clustering) is decreasing. At the same time, the popularity of research related to Recommender Systems, Network Analysis, and Human Behaviour Analysis is growing, which is most likely due to the increasing availability of data and the practical value of these topics. The research direction related to practical Applications of data mining also shows a tendency to grow. The last two topics, Text Mining and Data Streams have attracted steady interest from researchers. The results presented here shed light on the

structure and trends of data mining over the past twenty years and allow us to expand our understanding of this scientific discipline. We can argue that in the last five years a new research agenda has been formed, which is characterized by a shift in interest from algorithms to practical applications that affect all aspects of human activity.

Key words: data mining topics, topic analysis, scientometrics.

Citation: Zelenkov Yu.A., Anisichkina E.A. (2021) Trends in data mining research: A two-decade review using topic analysis. *Business Informatics*, vol. 15, no 1, pp. 30–46. DOI: 10.17323/2587-814X.2021.1.30.46

References

1. Piatetsky-Shapiro G., Fayyad U. (2012) An introduction to SIGKDD and a reflection on the term 'data mining'. *ACM SIGKDD Explorations Newsletter*, vol. 13, no 2, pp. 102–103. DOI: 10.1145/2207243.2207269.
2. Witten I.H., Frank E., Hall M., Pal C. (2017) *Data mining: Practical machine learning tools and techniques.* Cambridge, MA: Morgan Kaufmann.
3. Han J., Kamber M., Pei J. (2012) *Data mining: Concepts and techniques.* Waltham, MA: Morgan Kaufmann. DOI: 10.1016/C2009-0-61819-5.
4. Rather B. (2011) *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data.* Sound Parkway, NW: CRC Press.
5. Yang Q., Wu X. (2006) 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, vol. 5, no 4, pp. 597–604. DOI: 10.1142/S0219622006002258.
6. Wu X. (2010) 10 years of data mining research: retrospect and prospect. Proceedings of the *10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010*, p. 7. DOI: 10.1109/ICDM.2010.172.
7. Liao S.H., Chu P.H., Hsiao P.Y. (2012) Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, vol. 39, no 12, pp. 11303–11311. DOI: 10.1016/j.eswa.2012.02.063.
8. Van Eck N.J., Waltman L. (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, vol. 84, no 2, pp. 523–538. DOI: 10.1007/s11192-009-0146-3.
9. Thilakaratne M., Falkner K., Atapattu T. (2019) A systematic review on literature-based discovery: general overview, methodology, & statistical analysis. *ACM Computing Surveys*, vol. 52, no 6, article no 129. DOI: 10.1145/3365756.
10. Zelenkov Y. (2019) The topic dynamics in knowledge management research. Proceedings of the *14th International Conference on Knowledge Management in Organizations (KMO 2019), Zamora, Spain, 15–18 July 2019*, pp. 324–335. DOI: 10.1007/978-3-030-21451-7_28.
11. Mann G.S., Mimno D., McCallum A. (2006) Bibliometric impact measures leveraging topic analysis. Proceedings of the *6th ACM/IEE Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, NC, USA, 11–15 June 2006*, pp. 65–74. DOI: 10.1145/1141753.1141765.
12. Dam H.K., Ghose A. (2016) Analyzing topics and trends in the PRIMA literature. Proceedings of the *19th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA* 2016*), Phuket, Thailand, 22–26 August 2016*, pp. 216–229. DOI: 10.1007/978-3-319-44832-9_13.
13. Dias L., Gerlach M., Scharloth J., Altman E.G. (2018) Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, vol. 5, no 1, article no 171545. DOI: 10.1098/rsos.171545.
14. Jensen S., Liu X., Yu Y., Milojevic S. (2016) Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics*, vol. 10, no 2, pp. 606–621. DOI: 10.1016/j.joi.2016.04.002.
15. Syed S., Spruit M. (2017) Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. Proceedings of the *4th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017), Tokyo, Japan, 19–21 October 2017*, pp. 165–174. DOI: 10.1109/DSAA.2017.61.
16 Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, no 3, pp. 993–1022.
17. Mockus J. (2012) *Bayesian approach to global optimization: Theory and applications.* Heidelberg: Springer. DOI: 10.1007/978-94-009-0909-0.
18. Tang J., Meng Z., Nguyen X., Mei Q., Zhang M. (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of Machine Learning Research*, vol. 32, no 1, pp. 190–198.
19. Molina L.C., Belanche L., Nebot A. (2002) Feature selection algorithms – A survey and experimental evaluation. Proceedings of the *2nd IEEE International Conference on Data Mining (ICDM* 2002*), Maebashi City, Japan, 9–12 December 2002*, pp. 306–313. DOI: 10.1109/ICDM.2002.1183917.
20. Read J., Pfahringer B., Holmes G. (2008) Multi-label classification using ensembles of pruned sets. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 995–1000. DOI: 10.1109/ICDM.2008.74.
21. Chen X., Pan W., Kwok J.T., Carbonell J.G. (2009) Accelerated gradient method for multi-task sparse learning problem. Proceedings of the *9th IEEE International Conference on Data Mining (ICDM), Miami Beach, FL, USA, 6–9 December 2009*, pp. 746–751. DOI: 10.1109/ICDM.2009.128.
22. Shin K. (2011) Partitionable kernels for mapping kernels. Proceedings of the *11th IEEE International Conference on Data Mining (ICDM), Vancouver, BC, Canada, 11–14 December 2011*, pp. 645–654. DOI: 10.1109/ICDM.2011.115.
23. Li Z., Yada K. (2015) Why do retailers end price promotions – A study on duration and profit effects of promotion. Proceedings of the *15th IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015*, pp. 328–335. DOI: 10.1109/ICDMW.2015.56.

24. Camino R.D., State R., Montero L., Valtchev P. (2017) Finding suspicious activities in financial transactions and distributed ledgers. Proceedings of the *17th IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017*, pp. 787–796. DOI: 10.1109/ICDMW.2017.109.

25. Guo T., Bifet A., Antulov-Fantulin N. (2018) Bitcoin volatility forecasting with a glimpse into buy and sell orders. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 989–994. DOI: 10.1109/ICDM.2018.00123.

26. Gouda K., Zaki M.J. (2001) Efficiently mining maximal frequent itemsets. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 163–170. DOI: 10.1109/ICDM.2001.989514.

27. Ma S., Hellerstein J.L. (2001) Mining mutually dependent patterns. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 409–416. DOI: 10.1109/ICDM.2001.989546.

28. Zhou D., Orshanskiy S.A., Zha H., Gees C.L. (2007) Co-ranking authors and documents in a heterogeneous network. Proceedings of the *7th IEEE International Conference on Data Mining (ICDM), Omaha, NE, USA, 28–31 October 2007*, pp. 739-744. DOI: 10.1109/ICDM.2007.57.

29. Tang J., Jin R., Zang J. (2008) A topic modeling approach and its integration into the random walk framework for academic search. Proceedings of the *8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008*, pp. 1055–1060. DOI: 10.1109/ICDM.2008.71.

30. Shehata S., Karray F., Kamel M. (2006) Enhancing text clustering using concept-based mining model. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 1043–1048. DOI: 10.1109/ICDM.2006.64.

31. Yang D., Li B., Rettig L., Cudre-Mauroux P. (2017) HistoSketch: Fast similarity-preserving sketching of streaming histograms with concept drift. Proceedings of the *17th IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017*, pp. 545–554. DOI: 10.1109/ICDM.2017.64.

32. Dhurandhar A. (2010) Learning maximum lag for grouped graphical Granger models. Proceedings of the *10th IEEE International Conference on Data Mining Workshops (ICDMW), Sydney, Australia, 13 December 2010*, pp. 217–224. DOI: 10.1109/ICDMW.2010.9.

33. Wang H., Li Z., Lee W.-C. (2014) PGT: Measuring mobility relationship using personal, global and temporal factors. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 570–579. DOI: 10.1109/ICDM.2014.111.

34. Pirbhulal S., Wu W., Li G. (2018) A biometric security model for wearable healthcare. Proceedings of the *18th IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 17–20 November 2018*, pp. 136–143. DOI: 10.1109/ICDMW.2018.00026.

35. Yang J., Leskovec J. (2012) Defining and evaluating network communities based on ground-truth. Proceedings of the *12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, 10–13 December 2012*, pp. 745–754. DOI: 10.1109/ICDM.2012.138.

36. Shi L., Tong H., Tang J., Lin C. (2014) Flow-based influence graph visual summarization. Proceedings of the *14th IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014*, pp. 983–988. DOI: 10.1109/ICDM.2014.128.

37. Hung M.-C., Yang D.-L. (2001) An efficient fuzzy c-means clustering algorithm. Proceedings of the *2001 IEEE International Conference on Data Mining (ICDM), San Jose, CA, USA, 29 November – 2 December 2001*, pp. 225–232. DOI: 10.1109/ICDM.2001.989523.

38. Pei Y., Zaiane O.R., Gao Y. (2006) An efficient reference-based approach to outlier detection in large datasets. Proceedings of the *6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 18–22 December 2006*, pp. 478–487. DOI: 10.1109/ICDM.2006.17.

39. Wang X., Chen Y., Yang J., Wu L., Wu Z., Xie X. (2018) A reinforcement learning framework for explainable recommendation. Proceedings of the *18th IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018*, pp. 587–596. DOI: 10.1109/ICDM.2018.00074.

40. Wang H., Wang S. (2009) Discovering patterns of missing data in survey databases: an application of rough sets. *Expert Systems with Applications*, vol. 36, no 3, part 2, pp. 6256–6260. DOI: 10.1016/j.eswa.2008.07.010.

41. Provost F., Fawcett T. (2013) Data science and its relationship to big data and data-driven decision making. *Big Data*, vol. 1, no 1, pp. 51–59. DOI: 10.1089/big.2013.1508.

42. Brynjolfsson E., McElheran K. (2016) The rapid adoption of data-driven decision-making. *American Economic Review*, vol. 106, no 5, pp. 133–139. DOI: 10.1257/aer.p20161016.

43. Song P., Zheng C., Zhang C., Yu X. (2018). Data analytics and firm performance: An empirical study in an online B2C platform. *Information & Management*, vol. 55, no 5, pp. 633–642. DOI: 10.1016/j.im.2018.01.004.

About the authors

Yury A. Zelenkov

Dr. Sci. (Tech.);

Professor, Department of Business Informatics, Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: yzelenkov@hse.ru

ORCID: 0000-0002-2248-1023

Ekaterina A. Anisichkina

Student, BSc Program "Business Informatics", Graduate School of Business, National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: eaanisichkina@edu.hse.ru