1.4. ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ: ВОСПРИЯТИЕ НАУЧНОЙ ИНФОРМАЦИИ[1]

Милкова М. А. – научный сотрудник
лаборатории экспериментальной экономики ЦЭМИ РАН

В статье обсуждается вопрос о восприятии научной информации учеными. Ставится вопрос о формализации алгоритма тематического моделирования (в концепции АРТМ) для представления большой коллекции научных публикаций. Считается, что использование инструментария тематического моделирования будет способствовать эффективному управлению ограниченной рациональностью при получении знаний об исследуемой области.

Введение

Вопрос о деградации системного мышления в условиях перенасыщения информации, цифрового подталкивания, а также ограниченной рациональности индивидов ставился неоднократно (Helbing, et al., 2019; Милкова, 20219а). Применительно к производству научных знаний в ряде публикаций отмечается, например, снижение среднего качества подаваемых рукописей (так, в ведущих экономических журналах процент приема снизился с 15% в 1990 году до 6% в 2012 году (Card and DellaVigna, 2013)), снижение продуктивности исследовательской работы молодых ученых (Conley and Önder, 2014). Несмотря на то, что поиск и навигация по базам научных публикаций в последнее время значительно усовершенствованы (доступен семантический поиск публикаций, внедрены различные рекомендательные системы), получение общего представления о структуре изучаемого направления, ключевых публикациях, основных авторах внутри подтем является важной задачей. Кроме того, участие в междисциплинарных исследованиях всегда подразумевает поиск информации по смежным и малознакомым темам, по которым полный перечень ключевых слов для поиска может быть заранее и неизвестен. В добавление к этому, ограниченная рациональность, которая свойственна также и ученым, а также тот факт, что внимание к информации всегда ограничено тем классом событий, которые мы ожидаем увидеть, могут привести к получению сильно фрагментарных знаний.

Участие исследователей в мультидисциплинарных исследованиях, интерес к смежным областям, необходимость чтения научных публикаций по незнакомым или малознакомым темам были подтверждены в ходе небольшого пилотного исследования, проводимого среди ученых (в опросе участвовали 22 исследователя, 68% с научным стажем более 20 лет, 18% – 10-20 лет; 9% – 0-5 лет, 5% – 5-10 лет)[2]. Кроме того, было отмечено, что 64% опрошенных не всегда знают ключевые слова для поискового запроса; 95% приходится уточнять, дополнять ключевые слова для нового поискового запроса на основе уже найденной информации. Для 62% респондентов существовала необходимость представления структуры научных областей. 71% ученых считали в целом вопрос о сложностях потребления возрастающей в объеме научной информации актуальным.

Разведочный поиск информации (exploratory search) (Marchionini, 2006; White and Roth, 2009), в противовес привычному итерационному поиску по ключевым словам, способен сопутствовать познанию. Методы тематического моделирования (см. обзоры Daud et al., 2010; Милкова, 2019б) в данном контексте выступают инструментом для выявления структуры больших коллекций научных публикаций. Под структурой в данном случае мы будем понимать набор основных подтем изучаемого направления, по которым выделяются наиболее цитируемые публикации, основные авторы, ключевые слова.
Цель данной статьи – формализовать алгоритм тематического моделирования для представления структуры научных публикаций.

Некоторые аспекты применения тематического моделирования

Отметим, что основная сфера применения методов тематического моделирования – исследования в социальной сфере: анализ новостей, публикаций в социальных сетях и блогах с целью анализа дискурса, основных тем, настроений. Применению тематического моделирования для анализа научных публикаций посвящено меньше работ (см., например, Asmussen and Møller, 2019). Помимо классификации работ по сфере применения, их также целесообразно делить по применяемому методу. В настоящее время лидирующим является подход на основе метода латентного размещения Дирихле (LDA), однако его недостатком является существенная сложность включения в модель иных параметров, помимо текста (к которым может относиться различная мета-информация, например, авторы, теги, ссылки и т.п.). Альтернативным подходом является аддитивная регуляризация тематических моделей (АРТМ) (Воронцов, Потапенко, 2014а). Сравнение методов LDA и АРТМ можно посмотреть в работах

[1] Работа выполнена при поддержке Гранта РФФИ № 19-010-00293 «Разработка методологии, экономико-математических моделей, методик и систем поддержки принятия решений для проведения поисковых исследований по выявлению возможностей импортозамещения высокотехнологичной продукции на основе мировых патентных и финансовых информационных ресурсов».

[2] Опрос проводился лабораторией экспериментальной экономики ЦЭМИ РАН, 8-15 июня 2021. Ссылка на вопросы опросника:
https://docs.google.com/forms/d/e/1FAIpQLSfBpzrilkrA91DE79jEPTo7b_F99hNRLeCMaUHFEssjbketag/viewform

(Potapenko and Vorontsov, 2013). Обзор работ на основе LDA приведен в Asmussen and Møller (2019), в таблице 1 мы приводим работы на основе метода АРТМ.

Таблица 1. Работы о применении и тестировании подхода АРТМ

Статья	Данные	Объем данных	Описание. Стратегия регуляризации, коэффициенты
Янина, Воронцов (2016).	Определение тематик статей на habrahabr.ru	132157 статей	Использование модальностей слов (1.0), авторов (0.5), комментаторов (0.75), тегов (15.0), хабов (10.0). Использование регуляризаторов декоррелирования (1e+8), разреживания (-1.5), сглаживания (0.5)
Apishev et al. (2016).	Выявление этнически обусловленного контента на платформе LiveJournal	1.38 млн. документов	Использование только модальности слов. Выделение основных и фоновых тем. Использование экспертного словаря этнонимов. Сравнительный анализ 8 различных моделей.
Chirkova, Vorontsov (2016)	Построение иерархической модели (hARTM).	Тестирование на коллекции Wikipedia (3665223 статей); статей postnauka (1728 статей)	Разработка концепции иерархической аддитивной регуляризации.
Милкова (2020)	Разведочный поиск патентных документов (из базы Роспатента), соответствующих пунктам плана импортозамещения по 22 отраслям промышленности	152718 патентных документов	Использование модальности слов (1.0), наиболее частотных биграмм (5.0). Сглаживание по словарям с наименованиями товаров для импортозамещения (1e+8)
Ianina, Vorontsov (2020)	Разработка и тестирование системы для разведочного поиска схожих по тематике к задаваемым пользователем документам	Тестирование на основе данных habrahabr.ru (175143 статей); TechCrunch (759324 статей); триплетов (статья – похожие статьи – непохожие статьи) от Dai et al. (на основе 963564 статей arxiv.org)	Тестирование различных моделей: tf-idf, bm-25, GloVe, fasttext, CNN, MaLSTM, BERT, ARTM, hARTM
Gorshkov et al. (2021).	Определение тематик сообществ в сети Vkontakte	6967 сообществ.	Модель строится на основе одиночных слов. Приводится сравнение моделей LDA и ARTM. Применение регуляризаторов разреживания матриц Φ и θ. Значения коэффициентов регуляризации не приведены.

Подчеркнем, что цель нашего подхода – использовать методы тематического моделирования не просто для получения информации о структуре коллекции, а для конфигурирования среды так, чтобы она способствовала познанию, в частности, в области науки.

Схожую цель преследуют авторы в работе (Ianina, Vorontsov, 2020) в рамках которой разработана система для разведочного поиска научных публикаций в базе arxiv. Система позволяет осуществлять поиск тематически похожих документов к загружаемым пользователем подборкам статей.

Ключевые сложности применения АРТМ для научных исследований

Зарубежными и отечественными учеными подчеркивается сложность построения тематических моделей неспециалистами в области технических и компьютерных наук (Lee et al., 2017; Булатов, 2020). Ключевыми нерешенными проблемами являются формализация алгоритма, подбор параметров модели, обеспечение интерпретируемости результатов. В модели АРТМ исследователю необходимо не только экспериментально установить оптимальное число тем, но и выбрать оптимальную стратегию регуляризации, которая включает экспериментальный подбор коэффициентов для каждого из регуляризаторов каждой модальности. Несмотря на то, что общая стратегия регуляризации предложена в работах Воронцов, Потапенко (2014б), подбор коэффициентов «вслепую» значительно усложняет процесс построения модели. Кроме того, возможна некоторая специфика в стратегии моделирования для тех или иных типов текстовых коллекций. Имеющиеся публикации, представляющие тематическую модель для той или иной задачи, не объясняют, как именно выбирается диапазон для перебора значений коэффициентов. Выбор оптимального коэффициента делается на основе значений критериев качества модели, однако неясно, насколько сильно меняется состав топовых слов тем (для каждой модальности). Высокий барьер для входа в область тематического моделирования представителей смежных специалистов (экономистов, социологов, психологов), сложности с подбором параметров отмечаются в работах Boyd-Graber et al. (2017), Bulatov et al. (2020), Gorshkov et al. (2021).

Ряд авторов отмечает необходимость разработки программ для построения тематических моделей, предоставляющих пользователям возможности напрямую уточнять тематическую модель: разделять, объединять, удалять темы, явно указывать набор слов, которые должны встречаться в одной теме (Lee et al., 2017). Несмотря на то, что существуют задачи, которые, в соответствии со своими формулировками, требуют включения априорной информации о составе тем (см. работы Милкова, 2020; Apishev et al., 2016), участие пользователя в формировании тем не позволяет избежать влияния ограниченной рациональности на формирование новых знаний.

Стоит выделить созданную в недавнем времени систему TopicNet (Bulatov et al., 2020) – верхнеуровневую надстройку над общей библиотекой для работы с APTM (BigARTM), облегчающую работу с тематическими моделями. Однако система ориентирована на пользователей, не имеющих потребности в контроле всех параметров модели, а также больше ориентирована на решение задач в коммерческих, а не научных целях.

Таким образом, важной целью является формализация процедуры построения тематических моделей, предоставление четких инструкций и набора вспомогательных инструментов, позволяющих расширить применимость тематического моделирования.

Напомним (подробнее см. Воронцов, Потапенко, 2014a), что исходными данными для тематического моделирования является множество (коллекция) текстовых документов $\mathcal{D}$ и множество (словарь) терминов W. Каждый документ d ∈ $\mathcal{D}$ представляется последовательностью терминов $\mathcal{W} = \{w_1, \dots, w_{n_d}\}$, где n_d — длина документа. Через n_{dw} обозначается число вхождений термина w в документ d. Существует конечное множество тем T, и коллекция порождается дискретным распределением $p(d, w, t)$ на D×W ×T. Появление каждой пары (d,w) связано с некоторой неизвестной темой t. Построить тематическую модель коллекции — означает найти множество тем T, условные распределения $\varphi_{wt} = p(w|t)$ для каждой темы t ∈ T и $\theta_{td} = p(t|d)$ для каждого документа d ∈ D:

$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \qquad (1)$$

Условная вероятность в левой части оценивается по коллекции как $\hat{p}(w|d) = n_{dw}/n_d$, поэтому построение BTM можно рассматривать также как задачу поиска разложения матрицы $F = (p_{dw})_{D\times W}$ в произведение двух неотрицательных нормированных матриц меньшего размера: матрицы термов в темах $\Phi = (\varphi_{wt})_{W\times T}$, $\varphi_{wt} = p(w|t) = \frac{n_{wt}}{n_t}$; матрицы тем в документах $\Theta = (\theta_{td})_{T\times D}$, $\theta_{td} = p(t|d) = \frac{n_{td}}{n_d}$. $F \approx \Phi\Theta$.

Для определения параметров модели Φ, Θ максимизируется правдоподобие, используется EM-алгоритм. Так как построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения, множество её решений в общем случае бесконечно. Эти проблемы могут быть решены с помощью регуляризации модели, использования регуляризованного EM-алгоритма (Воронцов, Потапенко 2014a).

В недавних работах Воронцова и коллег предложен подход на основе введения относительных коэффициентов регуляризации (Булатов, 2020), позволяющих получить некоторую интерпретацию коэффициентов и тем самым облегчить их подбор.

Для регуляризаторов сглаживания и разреживания формула M-шага имеет вид: $\varphi_{wt} = norm_{w\epsilon W}(n_{wt} + \tau\beta_w)$,

$\beta_w = (\frac{1}{|W|})$- равномерное распределение, $\tau > 0$ для регуляризатора сглаживания, $\tau < 0$ для регуляризатора разреживания.

$$\varphi_{wt} = \frac{n_{wt}+\tau\beta_w}{\sum_{w\in W} n_{wt}+\tau\beta_w} = \frac{n_{wt}+\tau\beta_w}{n_t+\tau} \ . \qquad (2)$$

Влияние регуляризации можно описать как притягивание (в случае сглаживания) или отдаление (в случае разреживания) распределения n_{wt}/n_t, полученного как оценка максимального правдоподобия к равномерному распределению β_w с некоторым весом λ. φ_{wt} может быть записана как выпуклая комбинация этих двух распределений:

$$\varphi_{wt} = (1-\lambda)\frac{n_{wt}}{n_t} + \lambda\beta_w, \quad 0\leq \lambda \leq 1. \qquad (3)$$

Приравнивая (2) и (3), выразим τ: $\tau = \frac{n_t\lambda}{(1-\lambda)|W|}$.

Таким образом, выражение M-шага имеет вид: $\varphi_{wt} = norm_{w\epsilon W}\left(n_{wt} + n_t\frac{\lambda}{(1-\lambda)}\beta_w\right)$.

Величина $\frac{\lambda}{(1-\lambda)}$ определяет, во сколько раз регуляризатор влияет на оценку φ_{wt} больше, чем коллекция. Однако, чем больше значение n_t (число слов в теме), тем сильнее будет регуляризация. Возможно усреднение коэффициентов регуляризации по всем темам: $\varphi_{wt} = norm_{w\epsilon W}\left(n_{wt} + \frac{n}{|T|}\frac{\lambda}{(1-\lambda)}\beta_w\right)$.

Таким образом, относительный коэффициент показывает, во сколько раз регуляризатор влияет на оценку сильнее, чем коллекция.

О формализации алгоритма для выявления структуры научных публикаций

На основе относительных коэффициентов регуляризации предложена формализация алгоритма АРТМ для выделения структуры коллекции научных публикаций. Тестирование проводилось на основе различных выборок из базы, предоставляемой ресурсом Semantic Scholar. Результаты тестирований выложены в репозитории на Github[3]. Здесь же мы приводим формализованный алгоритм (см. Таблицу 2).

Построение модели на основе следующей метаинформации (включения следующих модальностей): 1) одиночных слов; 2) двухсловных словосочетаний (биграмм) Названий и Аннотаций статей; 3) авторов статей; 4) списков использованной литературы. Веса модальностей предлагается брать равными 0.5 для одиночных слов; 1.0 – для биграмм и ссылок, 2.0 – для авторов.

Таблица 2. Формализация алгоритма построения тематических моделей

Шаг	Описание шага	Контролируемые меры качества	Критерии отбора
1	Отбор числа тем. Исходя из общего понимания задачи, определяется диапазон возможного числа тем. Внутри диапазона с равным шагом выбирается 4-5 значений числа тем. Например: 5, 10, 15, 20 тем (для определения подтем направления); число тем значительно увеличивается в случае расширения анализируемой области знаний. Делается 10 проходов (1-10) по коллекции без регуляризации и с включенным регуляризатором декоррелирования. Тестируются несколько значений относительного коэффициента: 0.005, 0.01, 0.05, 0.1, 0.15.	Ключевые показатели: Средняя когерентность (по 15 наиболее частотным биграммам) Перплексия Дополнительно контролируемые показатели: Разреженность матриц Φ и θ.	Выбираются два эксперимента, по которым были получены результаты с наибольшей когерентностью, из них выбирается эксперимент с наименьшей перплексией. Особенность: Когерентность обратно коррелирует с перплексией
2	Подбор коэффициента сглаживания для фоновой темы (или отказ от него). Перебор значений по сетке, 10 итераций. К выбранному числу тем и значению коэфф. декоррелирования добавить одну фоновую тему. Сетка для относительного коэффициента сглаживания [0.01, 0.05, 0.1, 0.15, 0.2]		
3	Подбор коэффициентов разреживания. Делается дополнительно 10 проходов (11-21 итерации), тестируется коэффициент разреживания для матриц Φ и θ. Тестируемые значения относительного коэффициента для матрицы Φ в диапазоне от -0.1 до -0.8 с шагом 0.1. Для каждого значения рассчитывается по формуле $\tau = \frac{n}{\lvert D\rvert\lvert T\rvert}\frac{\lambda}{(1-\lambda)}$ соответствующее значение абсолютного коэфф. разреживания для матрицы θ.		
4	Анализ когерентности для выбранной стратегии регуляризации	Значения когерентности по темам, интерпретируемость тем	Если существуют неинтерпретируемые и/или плохо интерпретируемые темы, проводится тестирование для близкого числа тем (±2 темы)
5	Выбранная стратегия регуляризации тестируется в окрестности выбранного числа тем (±2 темы)	Аналогично шагам 1-3.	

Литература

1. Булатов, В.Г. (2020). Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet. Диссертация на соискание ученой степени кандидата технических наук.
2. Воронцов К. В., Потапенко, А.А. (2014а). Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН, Т. 456 №3, 268—271.
3. Воронцов, К.В., Потапенко, А.А. (2014б). Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4 — 8 июня 2014 г.). С. 676-687.
4. Милкова М.А. (2019а). Теория подталкивая и ее искажения в информационной среде // Цифровая экономика, 4(8), с. 21-26. https://doi.org/10.34706/DE-2019-04-02

[3] Репозиторий GitHub: https://github.com/behavioral-econ-codes/Publications

5. Милкова М.А. (2019б). Тематические модели как инструмент «дальнего чтения» // Цифровая экономика, 1(5), с. 57—70. DOI:10.34706/DE-2019-01-06
6. Милкова М.А. (2020). Инновационный подход к поиску информации на примере патентного анализа плана импортозамещения // Экономическая наука современной России, 1(88). С. 143-157.
7. Янина, А.О., Воронцов, К.В. (2016). Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных, 2(2), 173—186.
8. Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., Vorontsov, K. (2016). Mining Ethnic Content Online with Additively Regularized Topic Models // Computación y Sistemas, Vol. 20. No. 3. P. 387–403.
9. Asmussen, C.B., Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review // Journal of Big Data, 6:93 (2019). https://doi.org/10.1186/s40537-019-0255-7
10. Boyd-Graber, J., Hu, Y., Mimno, D. (2017). Applications of Topic Models. Foundations and Trends® in Information Retrieval: Vol. 11: No. 2-3, pp 143-296. http://dx.doi.org/10.1561/1500000030
11. Bulatov, M., Egorov, E., Veselova, E., Polyudova, D., Alekseev, V., Goncharov, A., Vorontsov, K. (2020). TopicNet: Making Additive Regularisation for Topic Modelling Accessible // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp 6745–6752, Marseille, 11–16 May 2020
12. Card, D., DellaVigna, S. (2013). Nine Facts about Top Journals in Economics // Journal of Economic Literature, 51 (1), pp. 144-61, https://doi.org/10.1257/jel.51.1.144
13. Chirkova, N. A., & Vorontsov, K. V. (2016). Additive regularization for hierarchical multimodal topic modeling // Journal Machine Learning and Data Analysis, 2(2), pp. 187-200
14. Conley, J.P., Önder, A.S. (2014). The Research Productivity of New PhDs in Economics: The Surprisingly High Non-success of the Successful // Journal of Economic Perspectives, 28 (3), pp. 205-16.
15. Dai, A.M., Olah, C., Le, Q.V. (2015). Document embedding with paragraph vectors // CoRR abs/1507.07998
16. Daud, A., Li, J., Zhou, L., and Muhammad, F. (2010). Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China, 280-301. — перевод на русский К. В. Воронцов, А. В. Темлянцев и др.
17. Gorshkov S., Ilyushin E., Chernysheva A., Goiko V., Namiot D. (2021). USING TOPIC MODELING FOR COMMUNITIES CLUSTERIZATION IN THE VKONTAKTE SOCIAL NETWORK // International Journal of Open Information Technologies, Vol.9 №5, 12-17.
18. Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Hoven, J., Zicari, R.V., Zwitter, A. (2019). Will Democracy Survive Big Data and Artificial Intelligence? In: Dirk Helbing eds. Towards digital enlightenment: Essays on the dark and light sides of the digital revolution. Springer.
19. Ianina, A. Vorontsov, K. (2020). Hierarchical Interpretable Topical Embeddings for Exploratory Search and Real-Time Document Tracking // International Journal of Embedded and Real-Time Communication Systems, 11(4), pp. 134-152. https://doi.org/0.4018/IJERTCS.2020100107
20. Lee, T.Y., Smith, A., Seppi, K. Elmqvist, N., Boyd-Graber, J., Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models // International Journal of Human-Computer Studies, 105, pp. 28—42.
21. Marchionini, G. (2006). Exploratory search: From finding to understanding. Communications of the ACM, 49(4), 41–46. https://doi.org/10.1145/1121949.1121979
22. Potapenko, A. A., Vorontsov, K. V. (2013). Robust PLSA Performs Better Than LDA. 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. —Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 784–787.
23. White R.W., Roth R.A. (2009). Exploratory search: Beyond the query-response paradigm. In: G. Marchionini (ed) Synthesis Lectures on Information Concepts Retrieval and Services 1(1). Morgan & Claypool Publishers, p.98. https://doi.org/10.2200/S00174ED1V01Y200901ICR003

References in Cyrillics

1. Bulatov, V.G. (2020). Metody` ocenivaniya kachestva i mnogokriterial`noj optimizacii te-maticheskix modelej v biblioteke TopicNet. Dissertaciya na soiskanie uchenoj stepeni kandidata texnicheskix nauk.
2. Voronczov K. V., Potapenko, A.A. (2014a). Additivnaya regulyarizaciya tematicheskix modelej kollekcij tekstovy`x dokumentov // Doklady` RAN, T. 456 №3, 268-271
3. Voronczov, K.V., Potapenko, A.A. (2014b). Regulyarizaciya veroyatnostny`x tematicheskix mode-lej dlya povy`sheniya interpretiruemosti i opredeleniya chisla tem. Komp`yuternaya lingvi-stika i intellektual`ny`e texnologii: Po materialam ezhegodnoj Mezhdunarodnoj konfe-rencii «Dialog» (Bekasovo, 4 — 8 iyunya 2014 g.). S. 676-687.
4. Milkova M.A. (2019a). Teoriya podtalkivaya i ee iskazheniya v informacionnoj srede // Cifrovaya e`konomika, 4(8), s. 21-26. https://doi.org/10.34706/DE-2019-04-02

5. Milkova M.A. (2019b). Tematicheskie modeli kak instrument «dal`nego chteniya» // Cifrovaya e`konomika, 1(5), s. 57-70. DOI:10.34706/DE-2019-01-06
6. Milkova M.A. (2020). Innovacionny`j podxod k poisku informacii na primere patentnogo analiza plana importozameshheniya // E`konomicheskaya nauka sovremennoj Rossii, 1(88). S. 143-157.
7. Yanina, A.O., Voronczov, K.V. (2016). Mul`timodal`ny`e tematicheskie modeli dlya razvedoch-nogo poiska v kollektivnom bloge // Mashinnoe obuchenie i analiz danny`x, 2(2), 173-186.I

Милкова Мария Александровна –научный сотрудник
лаборатории экспериментальной экономики ЦЭМИ РАН
(m.a.milkova@gmail.com)

Ключевые слова
тематические модели, АРТМ, научная информация.

Maria Milkova, Topic modeling of scientific information perception

Keywords
topic models, ARTM, scientific information.

DOI: 10.34706/DE-2021-02-04
JEL classification: D83 – Поиск • Обучение • Информация и знания • Взаимодействие • Мнение • Неосведомленность

Abstract
The article discusses the issue of the perception of scientific information by scientists. The question is raised about the formalization of the topic modeling algorithm (in the concept of ARTM) for the presentation of a large collection of scientific publications. It is believed that the use of thematic modeling tools will contribute to the effective management of bounded rationality in gaining knowledge about the study area.