Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2018, №2, Том 10 / 2018, No 2, Vol 10 https://esj.today/issue-2-2018.html

URL статьи: https://esj.today/PDF/12ITVN218.pdf

Статья поступила в редакцию 15.03.2018; опубликована 08.05.2018

Ссылка для цитирования этой статьи:
Краснов Ф.В., Докука С.В. Вероятностная модель скрытых тем на основе архива журнала «Нефтяное Хозяйство» // Вестник Евразийской науки, 2018 №2, https://esj.today/PDF/12ITVN218.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Krasnov F.V., Dokuka S.V. (2018). Probabilistic model of latent topics based on the archive of the journal "Oil Industry". *The Eurasian Scientific Journal*, [online] 2(10). Available at: https://esj.today/PDF/12ITVN218.pdf (in Russian)

УДК 316.452

Краснов Федор Владимирович
ООО «Газпромнефть НТЦ», Санкт-Петербург, Россия
Эксперт
Кандидат технических наук
E-mail: Krasnov.FV@Gazprom-Neft.ru
ORCID: http://orcid.org/0000-0002-9881-7371
РИНЦ: https://elibrary.ru/author_profile.asp?id=855886
Researcher ID: http://www.researcherid.com/rid/C-6518-2018

Докука София Владимировна
ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
Научный сотрудник
Кандидат социологических наук
E-mail: sdokuka@hse.ru
ORCID: http://orcid.org/0000-0002-0847-5129
РИНЦ: https://elibrary.ru/author_profile.asp?id=661868
Researcher ID: http://www.researcherid.com/rid/H-7306-2015

Вероятностная модель скрытых тем на основе архива журнала «Нефтяное Хозяйство»

Аннотация. Вопрос о том, по какому пути движется прикладная наука и технологии, является ключевым для любой научно-технической области. Традиционно определение векторов развития производилось и производится экспертами по предметному направлению, однако значительный рост объемов информации и увеличение числа направлений развития свидетельствуют о необходимости доработки и совершенствования этого инструментария и выработке дополнительных методов исследования индустриальных трендов. В данном исследовании мы проанализировали проанализировать тренды нефтяной индустрии посредством автоматизированной обработки текстов научных статей в отраслевом журнале «Нефтяное хозяйство». Выявляя наиболее часто встречаемые темы в журнале за период с 2008 по 2016 годы, мы сделали вывод об увеличении значимости трудноизвлекаемых запасов и росте интереса к методам разработки подобных месторождений. Вместе с тем мы выявляем ограничения используемого подхода и предлагаем возможные варианты для его совершенствования.

Ключевые слова: тематическое моделирование; автоматизированный анализ текстов; наукометрия; трудно извлекаемые запасы; ТРИЗ; технологические тренды; BigARTM

Актуальность исследования

В последние годы вопрос о том, по какой траектории происходит развитие нефтегазового комплекса, как и всей энергетической системы, приобретает все больший интерес, как со стороны экспертов, так и со стороны широкой общественности [5, 1].

Этому способствует несколько факторов. Во-первых, темпы экономического развития приводят к значительному росту мирового энергопотребления. Как отмечается в докладе Аналитического Центра при Правительстве РФ [1], значительный рост потребления энергоресурсов происходит за счет развивающихся стран, преимущественно Азиатско-Тихоокеанского региона, в то время как в развитых странах объем выработки электроэнергии стабилен, а динамика потребления схожа с тенденциями общеэкономических приростов и спадов. Во-вторых, наблюдается изменение структуры запасов углеводородов. Как отмечается в «Энергетической стратегии России на период до 2035 года»[2] (сформулированной в 2015 году), отечественная нефтяная отрасль сталкивается с такой проблемой, как «увеличение себестоимости добычи вследствие преобладания труднодоступных запасов нефти (*далее по тексту ТРиЗ*) и большой выработанности действующих месторождений, что усложняет удержание достигнутых уровней добычи нефти». При этом одной из задач, ставящейся перед нефтяным сектором, является освоение ТРиЗ в объемах до 17 % от общей добычи, которая может быть решена путем развития добывающих технологий. Наконец, в-третьих, всё большую роль в энергетическом секторе играют источники возобновляемой энергии (т. н. ВИЭ), что сказывается на структуре энергетических рынков[3]. Эксперты, политики и граждане всё больше озабочены экологическими и климатическими вызовами, что свидетельствует о необходимости диверсификации энергоносителей. Дополнительно стоит отметить негативное влияние внешних экономических и политических ограничений на сырьевой сектор российской экономики.

Таким образом, энергетическая сфера находится в процессе постоянной трансформации, а одним из важных вопросов повестки дня нефтяного сообщества является оптимизация методов геологоразведки, добычи и использования энергоносителей.

Анализировать, по какой траектории движется изменение научно-технических и технологических процессов нефтедобычи, можно несколькими способами. Наиболее очевидным видится опрос экспертов, специализирующихся на вопросах добычи. Методы экспертных опросов (также называемые *методами экспертных оценок*) широко используются в различных исследованиях, в которых невозможны или труднодоступны другие формы исследований ввиду отсутствия объективных данных. Таким образом реализуется подавляющее большинство форсайт-исследований. К достоинствам экспертного опроса можно отнести их относительную простоту и доступность, а также возможность применения в случае отсутствия информации об изучаемом явлении.

В то же время очевидным недостатком экспертного опроса являются возможный субъективизм и ограниченность экспертов, их приверженность определенной точке зрения. Как отмечается в работе Бахтина с соавторами [1], в течение последних лет объемы экспертно-аналитической и научной литературы, а также информации в целом, стремительно растут (по некоторым оценкам объемы информации удваиваются каждые два года), так что задача получения, фильтрации, переработки и рефлексивного восприятия всей информации становится фактически невозможной. При этом эксперту необходимо развиваться и

[1] http://ac.gov.ru/files/publication/a/7945.pdf (Дата обращения 10.04.2017 г.).

[2] http://www.energystrategy.ru/ab_ins/source/ES-2035_09_2015.pdf (Дата обращения 17.05.2017 г.).

[3] https://issek.hse.ru/news/197481895.html (Дата обращения 17.05.2017 г.).

совершенствоваться в различных содержательных направлениях, что требует ещё больших трудовых и временных инвестиций. Это свидетельствует о необходимости разработки и формирования дополнительной обратной связи, которая призвана помочь экспертному и профессиональному сообществу анализировать огромные объемы информации и выделять из нее наиболее релевантные аспекты, в частности – выявлять технологические тренды.

С развитием автоматизированных методов обработки неструктурированных данных, в частности текстовых данных, популярность набирает тематическое моделирование научных текстов [3]. Как было продемонстрировано в работе Блея и Лафферти, тематическое моделирование оказывается перспективным инструментов отслеживания трендов в таких научных направлениях, как ядерная физика и нейронауки [3], технологии агропромышленного комплекса [1] и так далее. Изучение автоматически выделенных тематик во временной перспективе иллюстрирует изменение интереса научного сообщества к различным объектам и предметам исследования. Достоинством этого метода является возможность автоматизированной обработки огромных массивов информации и выявления латентных (скрытых) тематик текстов. При этом тематическое моделирование нельзя назвать исключительно автоматизированным методом, так как полученные в результате машинной классификации тематики в дальнейшем должны быть вручную просмотрены и проработаны экспертами-специалистами предметной области. Таким образом, тематическое моделирование может рассматриваться как метод, заключающий в себе достоинства и автоматизированной обработки текста, и экспертной оценки. Реализация подобного метода в приложении к различным содержательным задачам позволит формировать диалог между наукой и стратегией на принципиально новом уровне.

Данная статья ставит перед собой задачу проследить эволюцию тематик, рассматриваемых в российском нефтегазовом сообществе и выявить перспективы инструментов автоматизированного анализа текстов в предсказании технологических трендов.

В рамках настоящего исследования нами была предпринята попытка выявить ключевые технологические тенденции в нефтяном комплексе путем анализа текстов журнала «Нефтяное хозяйство» за период с 2008 по 2016 годы. «Нефтяное хозяйство» является одним из ключевых научно-исследовательских журналов, посвященных нефтяной промышленности. Из полных текстов статей для каждого года были выделены наиболее релевантные тематики, которые впоследствии были проанализированы и интерпретированы экспертом. Результаты исследования продемонстрировали важность разработки методов для получения трудно извлекаемых запасов нефти в экспертном дискурсе, однако нам не удалось выявить каких-то конкретных точных и детальных методик и технологий, которые бы появлялись для решения данной задачи и были бы актуальны в течение длительного периода времени. Подобные результаты могут быть интерпретированы в различных направлениях. С одной стороны, нами показано, что в анализируемом журнале за выбранный период наибольший акцент ставится на исследование новейших технологий, при этом не прослеживается систематического обзора используемости существующих технологических трендов в области добычи. С другой стороны, анализ эволюции технологий разработки трудноизвлекаемых запасов может быть осуществлен в рамках более целевой текстовой выборки, например, на материалах тезисов специализированных конференций. Но даже узкоспециализированные конференции не будут отражать падение интереса к тем или иным технологиям. Сложно представить себе материал конференции, констатирующий список технологий, которые авторы не применили с объяснением причин. Гораздо чаще в докладах представляют только истории успеха.

Структура статьи устроена следующим образом. Во второй части мы описываем используемый метод. В третьей части описан алгоритм исследования. В четвертой

представлены результаты. Пятая часть подводит итоги работы и описывает перспективы данного метода.

Метод исследования

Наиболее современным подходом к задаче выявления тематик на взгляд авторов является метод аддитивной регуляризации для стохастических матриц, изложенный в работе [15]. Тематическое моделирование – это одно из направлений автоматизированного анализа текстов, призванное выявить к каким темам относится каждый из документов коллекции и какими терминами определяется та или иная тема. Таким образом, на входе исследователем подается коллекция текстов, а на выходе мы получаем тематики (с набором слов для каждой из тематик) на основе которых может быть проведена кластеризация коллекции текстов. Тема в данном случае – это результат би-кластеризации, то есть одновременной кластеризации слов и документов по их семантической близости [14]. Важной задачей, реализуемой в рамках тематического моделирования, является также задача снижения размерности, потому что число тем значительно меньше числа слов в документах. Тематическое моделирование – это популярный инструмент анализа текстовой информации, реализованный в таких широко используемых библиотеках как Gensim [6].

Инструментарий тематического моделирования используется для изучения массивов данных научных публикаций и определения трендов развития того или иного научного направления [3], анализа новостных потоков [12], информационного поиска [16] и так далее. Как отмечается Коршуновым и Гомзиным [14], переход из пространства терминов (так называют преобразованные определенным образом слова) в пространство найденных тематик позволяет разрешать проблемы синонимии и полисинонимии терминов. Благодаря автоматическому анализу текстов становится решаемой задача реферирования и систематизации сверхбольших коллекций текстов, состоящих из миллиардов документов.

Большинство моделей разрабатывается на основе латентного размещения Дирихле (*Latent Dirichlet Allocation*, далее LDA) [7]. Эта модель объясняет присутствие одинаковых слов в различных документах тем, что документы посвящены одной тематике (то есть могут быть отнесены к латентной группе). На взгляд, авторов более перспективным подходом к моделированию тем является PLSA [9] за счет использования регуляризации при построении модели. Развитием PLSA, но уже с множественной регуляризацией является ARTM [15], примененный авторами данного исследования.

Среди недостатков тематического моделирования стоит отметить тот факт, что априорные распределения Дирихле имеют крайне слабые лингвистические обоснования [13] а также не учитывают порядок слов в предложении, порядок документов в корпусе и т. д. [14]. В случае подобных исследований текст воспринимается как «мешок слов» (bag of words, BoW), в котором порядок слов и частота встречаемости словосочетаний не играют значимой роли, а темы выявляются на основе частотности отдельных слов. Таким образом, несмотря на то, что традиционные подходы к автоматизированному анализу текстов и выявлению ключевых тематических направлений, и позволяют обработать большие объемы информации, недоступные ручной экспертизе, содержательная интерпретация полученных результатов не всегда оказывается полной и однозначной.

Одним из решений данной проблемы видится использование моделей тематического моделирования, учитывающих порядок слов в предложении и позволяющей прослеживать совместную встречаемость двух и более слов. Такие модели называются n-граммными моделями, где n – это число учитываемых слов [8] (в случае, когда n = 1, говорят об исследовании *униграмма*). В эмпирических исследованиях чаще всего используют модели

биграммов, при которых отслеживаются словосочетания из двух слов [10]. Совместное использование униграммов и биграммов позволяет выявлять ключевые темы, принимая во внимание и частотности слов и порядок слов в предложении.

Формальная постановка задачи тематического моделирования следующая. Пусть зафиксирован словарь терминов W, из элементов которого складываются документы, и дана коллекция D документов $d \subset D$. Для каждого документа d известна его длина n_d и количество n_{dw} использований каждого термина w.

Пусть $\Phi = (\phi_{wt})$ – матрица распределений терминов (w) в темах (t), а $\Theta = (\theta_{td})$ – матрица распределений тем (t) в документах (d). Тогда задача тематического моделирования состоит в том, чтобы найти такие матрицы Φ и Θ, чтобы выполнялось равенство (1).

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \quad (1)$$

где ϕ_{wt} – вероятности терминов w в каждой теме t, θ_{td} – вероятности тем t в каждом документе d, а $p(w|d)$ – вероятность появления термина w в документе d.

Эксперимент

Журнал «Нефтяное хозяйство» посвящен нефтегазовой проблематике. В нем публикуются статьи, посвященные широкому кругу вопросов нефтегазового сектора: экономических, технических, технологических, экологических и информационных. Издание насчитывает почти вековую историю и выходит каждый месяц с 1920 года. Все публикуемые статьи проходят процедуру рецензирования. Журнал включен в Российский индекс научного цитирования (РИНЦ) и международную систему индексирования Scopus. Материалы журнала находятся в закрытом доступе и распространяются по подписке. Основные рубрики журнала:

- новости нефтегазовых компаний;
- нефтяная и газовая промышленность;
- экономика, управление, право;
- геология и геологоразведочные работы;
- бурение;
- разработка и эксплуатация нефтяных месторождений;
- проектирование и обустройство месторождений;
- техника и технологии добычи нефти;
- нефтепромысловое оборудование;
- транспорт и подготовка нефти;
- экологическая и промышленная безопасность;
- информационные технологии.

Как видно, журнал подробно рассматривает практически все аспекты функционирования нефтяных компаний – от экономико-правовых вопросов, до технологических аспектов и тонкостей.

Для проведения исследования редакцией были любезно предоставлены архивы статей журнала Нефтяное хозяйство за период 2008-2016 гг. В выборке содержится 108 выпусков

журналов, со статьями от 3517 авторов. В каждом из выпусков журналов содержатся все статьи номера, таким образом, нами была получена сплошная выборка, в которой содержались материалы по самым различным содержательным направлениям. В среднем, в номере журнала «Нефтяного хозяйства» около 20-25 статей. Нами были рассмотрены именно выпуски журнала, так как они являлись единицей анализа. Авторами статей журнала являются научные сотрудники, инженеры и отраслевые эксперты, многие из них кандидаты и доктора наук. В среднем у одной статьи 3 соавтора.

Процесс исследования имел следующие этапы.

1. Изначально архивы представлены в виде файлов в формате PDF. Иногда это был единый файл (*биндер*) со статьями за весь год, а иногда разрозненные файлы с отдельными статьями. В обоих случаях файлы были предназначены для печати, то есть содержали оглавления, номера страниц, тематические вставки и другие редакторские элементы. Для анализа нужны были только тексты в виде предложений поэтому авторами был реализован программный модуль для приведения всех данных к такому формату. Отметим, что, исходя из выбранной нами методики, важно было сохранить порядок слов и разделение на предложения, при этом необходимо было сохранять принадлежность к выпуску, а не к статье, так как минимальной единицей временного анализа выбран один выпуск.

2. На втором этапе анализа происходило приведение слов к основным формам. Для анализа и сравнения слов методами частотного и вероятностного анализа необходимо сузить возможные варианты употребления словоформ. Существуют несколько алгоритмов для решения этой задачи (нормализации текста), в данном случае был использована *стемминг*. Стеммингом называют процедуру нахождения основы слова, при этом основа и корень слова могут различаться между собой.

Одним из наиболее распространенных инструментов является стеммер Портера [2], который, однако, часто обрезает слово больше необходимого, что затрудняет получение правильной основы слова, например, кровать -> крова. Также стеммер Портера не справляется со всевозможными изменениями корня слова (например, выпадающие и беглые гласные), характерными для русского языка. Поэтому авторы остановились на использовании технологии стемминга компании «Яндекс» – MyStem[4]. Данная программа производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь и предлагает несколько вариантов основ слова.

Тем не менее, авторы сочли необходимым поддерживать обратный словарь для полученных словоформ, чтобы сохранять связь между изначальным словом и полученной словоформой.

Отдельной веткой обработки подвергались аббревиатуры, широко распространенные в нефтегазовой отрасли. Определение аббревиатур производилось на основе словаря аббревиатур, созданного и поддерживаемого в компании Газпром Нефть в рамках проекта Корпоративной Википедии[5].

3. На третьем этапе исследования проводилось формирование словаря. Известно, что наибольшую смысловую нагрузку несут не одиночные слова, а сочетания слов, в частности пары слов – биграммы [8]. Для выделения биграммов авторами были использованы эвристические алгоритмы. Была составлена матрица биграммов в окрестности 5 слов для

[4] https://tech.yandex.ru/mystem/ (Дата обращения 13.04.2017 г.).

[5] https://www.onepetro.org/conference-paper/SPE-181270-MS (Дата обращения 12.04.2017 г.).

каждого из предложений. Затем были рассчитаны частоты использования каждого из биграммов, после чего были зафиксированы 5 % наиболее встречаемых словосочетаний.

4. На четвертом этапе словари отдельных слов и биграммов были объединены для общей обработки алгоритмами выделения тематик. Получившийся словарь был проанализирован, на предмет выделения высоко- и низко частотных слов для их фильтрации.

Традиционно окончательное формирование словаря производится с помощью стоп-слов. Алгоритмы выделения стоп-слов [11] не использовались авторами в данной статье. Это решение было обусловлено тем, что добавление словаря стоп-слов не добавляло точности и вносило субъективный характер исследования.

5. На заключительном пятом этапе производилось построение модели тематик. Для этого был использован инструмент BigARTM [4]. На этом этапе были получены матрицы Φ (распределение тем для документа) и Θ (распределение слов для темы). Для повышения точности алгоритма авторами был применен аналитический подход, уточняющий регуляризционные параметры на основании метрик.

Результаты

Основной метрикой для выявления факта сходимости модели тем является метрика *Perplexity* вычисляемая по формуле (2).

$$\mathcal{P}(D,\Phi,\Theta) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{t\in T}\phi_{wt}\theta_{td}\right) \quad (2)$$

В уравнении (2) использованы следующие обозначения: n_{dw} – количество терминов (w) в документе d, $\Phi = (\phi_{wt})$ – матрица распределений терминов (w) в темах (t), а $\Theta = (\theta_{td})$ – матрица распределений тем (t) в документах (d).

График зависимости *Perplexity* от количества проходов по корпусу текстов отображен на (рис. 1).

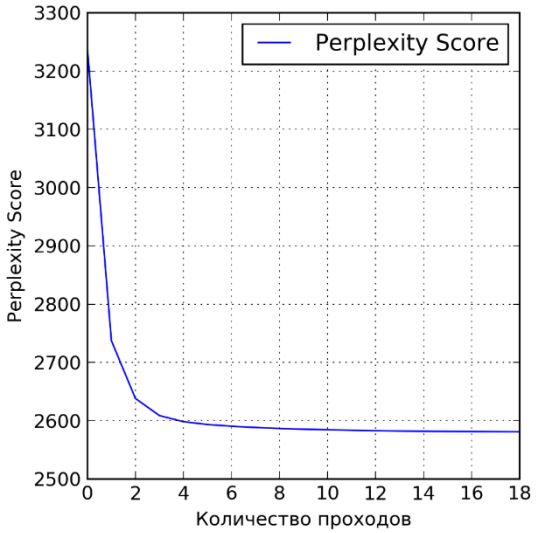

Рисунок 1. *Зависимость метрики Perplexity от количества проходов по корпусу текстов (разработано авторами)*

Из (рис. 1) видно, что за три прохода модель показала приемлемую сходимость и не нуждается в дальнейшей оптимизации.

Важными метриками качества модели тем являются степень разрежённости матриц Φ и Θ . Повлиять на эти метрики можно с помощью параметров τ , соответствующих регуляризаторов. На рис. 2 и рис. 3 отображены зависимости для разрежённости матриц Θ и Φ.

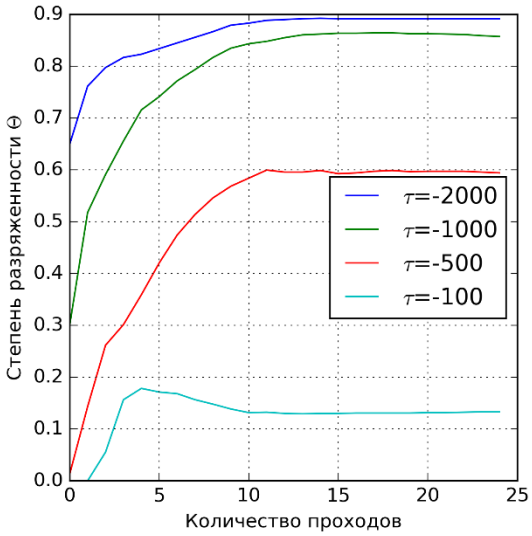

Рисунок 2. *Зависимость разрежённости матрицы Θ от параметра регуляризации τ (разработано авторами)*

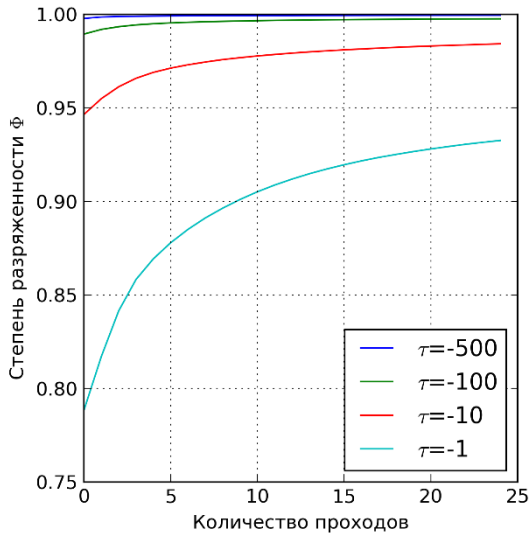

Рисунок 3. *Зависимость разрежённости матрицы Φ от параметра регуляризации τ (разработано авторами)*

На основании зависимостей рис. 2 и рис. 3 авторами были выбраны параметры регуляризации модели тем, позволяющие достичь оптимального соотношения между значимыми терминами и шумовыми.

Тематическая модель, полученная в результате данного исследования, может быть представлена в различных формах. Уровень шумовых терминов мешает интерпретировать результаты, поэтому от запланированных 12 тем содержательных осталось шесть. В таблице 1 представлены темы, выделенные с помощью модели.

Таблица 1

Фрагмент матрицы Ф для терминов с максимальными вероятностями

1	2	3	4	5	6
электронный	ЭЦН	сдвиг	почва	нефтегазоносность	ингибитор
знание	УЭЦН	сигнал	добавка	свод	разлом
автоматизация	сероводород	окисление	композиция	компания	деформация
интегрировать	фациальный	разрушение	знание	впадина	трещиноватость
пользователь	гамма	деформация	агрегат	сепарация	исследовательский
архив	доломит	реологический	загрязнение	миграция	известняк
хранение	замер	песчаный	ПЗП	прогнозный	порода
доступ	депрессия	осадки	надежность	активность	политехнический
подразделение	агент	капиллярный	камень	филиал	штанга
платформа	каротаж	сечение	окисление	цемент	приемистость

Разработано авторами

Можно с уверенностью сказать, что термины, собранные в столбце №1 характеризуют тематику управления знаниями. В столбце №2 представлена тема добычи. Остальные столбцы тоже могут быть достаточно однозначно проинтерпретированы. А для машинной обработки набор терминов важнее чем обобщающая его тема.

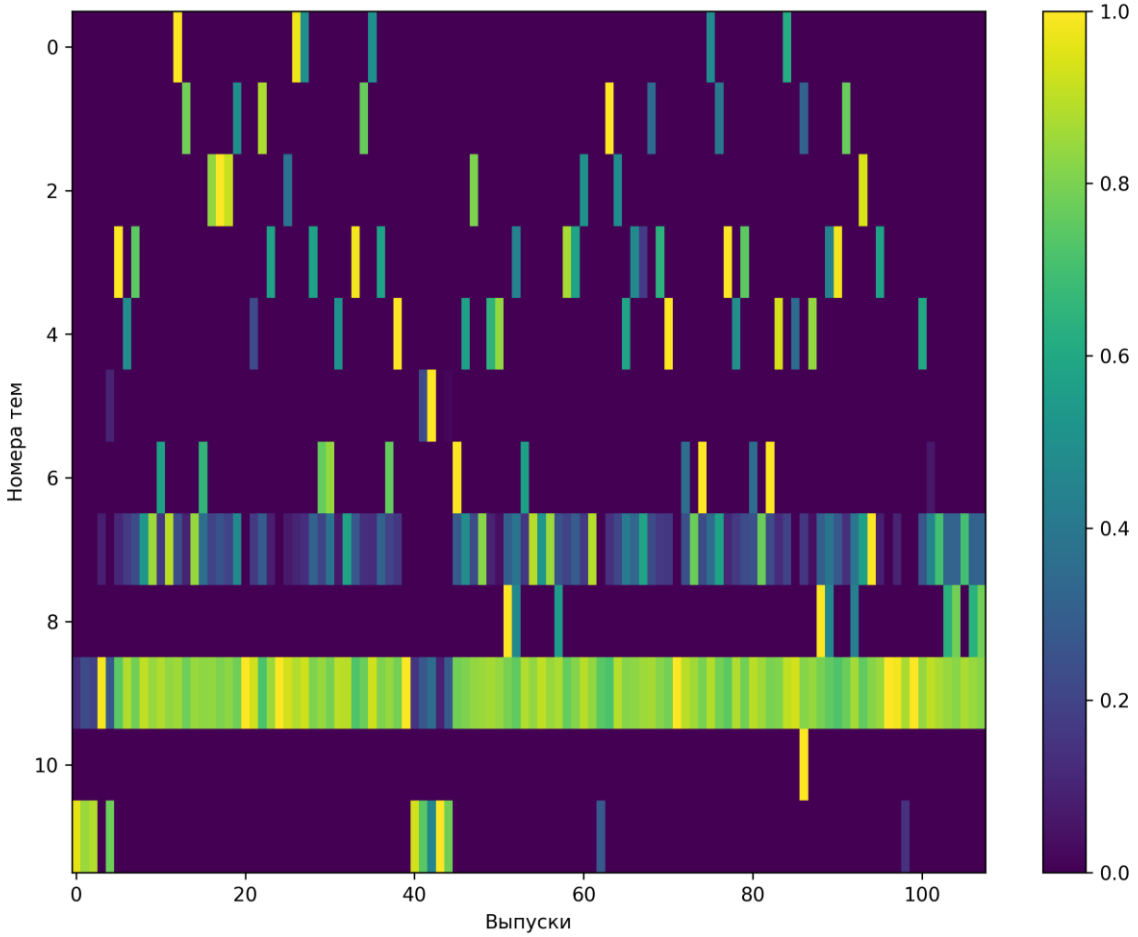

Рисунок 4. *Матриц Θ. Распределение тем для документов (разработано авторами)*

На рис. 4 представлена матрица Θ, дающая представление как полученные тематики распределены в каждом из анализируемых выпусков. Можно увидеть, что тема с №9 представлена во всех выпусках – это общая информация, поздравления и т. п. Полученное представление позволяет выбирать наиболее релевантные выпуски с определенной темой.

Заключение и выводы

В данной работе было проведено исследование, в котором из коллекции текстов, написанных в различные моменты времени, были выделены скрытые тематики. В рамках исследования тексты были сгруппированы по выпускам.

Важно отметить, что выбранный авторами метод показал высокую скорость анализа, что делает его возможным для применения в онлайновых процессах поиска. Например, на сайте издательства в качестве средства, улучшающего поиск и дающего рекомендации читателям по статьям со схожей тематикой.

Также следует отметить, что данная методика может быть в дальнейшем усовершенствована и адаптирована для анализа существенно больших массивов динамических данных и выделения ключевых направлений технологического развития как в более широких, так и в более узких областях.

Механизм последовательной регуляризации показал себя как гибкий метод и хорошо интерпретируемый метод настройки модели тем.

Как отмечается Кузьминовым с соавторами [5], существующие прогнозы научно-технического развития (с том числе форсайт-прогнозы) в большинстве своем экстраполируют существующие тренды на долгосрочную перспективу. Таким образом, большой интерес приобретают работы, в которых становится возможным выявление новых технологических направлений, способных существенно видоизменить структуру рынков.

Сами по себе отдельные технологии не следует рассматривать как оторванные и изолированные друг от друга инициативы. В действительности многие технологические направления развиваются параллельно, что является результатом венчурной политики, технологического развития и других сопутствующих факторов. Ввиду этого важным направлениям анализа технологических трендов выглядит изучение коэволюции развития сразу нескольких технологий. Именно изучение совокупности научно-технических инициатив позволит содержательно проанализировать направление технологического развития.

ЛИТЕРАТУРА

1. Bakhtin P. et al. Trend monitoring for linking science and strategy // Scientometrics. – 2017. – С. 1-17.
2. Porter M.F. Snowball: A language for stemming algorithms. – 2001.
3. Blei D.M., Lafferty J.D. Dynamic topic models // Proceedings of the 23rd international conference on Machine learning. – ACM, 2006. – С. 113-120.
4. Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections // International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 370-381.
5. Kuzminov I.F., Bereznoy A.V., Bakhtin P. Global energy challenges and the national economy: stress scenarios for Russia // foresight. – 2017. – Т. 19. – № 2.
6. Řehůřek R., Sojka P. Gensim-Python Framework for Vector Space Modelling // NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. – 2011.
7. Blei D.M. Probabilistic topic models // Communications of the ACM. – 2012. – Т. 55. – №. 4. – С. 77-84.

8. Wallach H.M. Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning. – ACM, 2006. – С. 977-984.
9. Mei Q. et al. Topic modeling with network regularization // Proceedings of the 17th international conference on World Wide Web. – ACM, 2008. – С. 101-110.
10. Wang X., McCallum A., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. – IEEE, 2007. – С. 697-702.
11. Wilbur W.J., Sirotkin K. The automatic identification of stop words // Journal of information science. – 1992. – Т. 18. – № 1. – С. 45-55.
12. Zhao W.X. et al. Comparing twitter and traditional media using topic models // European Conference on Information Retrieval. – Springer Berlin Heidelberg, 2011. – С. 338-349.
13. Воронцов К.В. Вероятностное тематическое моделирование // Москва. – 2013.
14. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. – 2012. – Т. 23.
15. Vorontsov K., Potapenko A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // International Conference on Analysis of Images, Social Networks and Texts_x000D_. – Springer, Cham, 2014. – С. 29-46.
16. Ponte J.M., Croft W.B. A language modeling approach to information retrieval // Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 1998. – С. 275-281.

Krasnov Fedor Vladimirovich
LLC "Gazpromneft NTC", Saint-Petersburg, Russia
E-mail: Krasnov.FV@Gazprom-Neft.ru

Dokuka Sofiya Vladimirovna
National research university "Higher school of economics", Moscow, Russia
E-mail: sdokuka@hse.ru

Probabilistic model of latent topics based on the archive of the journal "Oil Industry"

Abstract. The question of which way applied science and technology are moving is key for any scientific and technological field. Traditionally, subject matter experts make the definition of vectors of development, but a significant increase in the volume of information and an increase in the number of areas of development indicate the need to improve this approach and develop additional methods for studying industrial trends.

In this study, we analyzed the trends of the oil industry through the automated processing of scientific articles in the journal "Oil industry (Neftyanoe Khozyaystvo)". Identifying the most frequently encountered topics in the journal for the period from 2008 to 2016, we concluded that the importance of hard-to-recover (HTR) reserves increased interest in the methods of development of such oilfields. However, we identify limitations to the modeling approach and propose options for improving it.

Keywords: topic modeling; automated analysis of texts; scientometrics; hard recoverable reserves; technology trends; BigARTM