ИННОВАЦИОННЫЙ ПОДХОД К ПОИСКУ ИНФОРМАЦИИ НА ПРИМЕРЕ ПАТЕНТНОГО АНАЛИЗА ПЛАНА ИМПОРТОЗАМЕЩЕНИЯ[1]

М.А. Милкова

DOI: 10.33293/1609-1442-2020-1(88)-143-157

В настоящее время процесс накопления информации настолько стремителен, что концепция привычного итерационного поиска требует пересмотра. К методам поиска необходимо предъявлять повышенные требования, находясь в мире, перенасыщенном информацией, чтобы всесторонне охватить и проанализировать исследуемую проблему. Инновационный подход к поиску должен гибко учитывать большой объем уже накопленных знаний и априорные требования к результатам. Результаты, в свою очередь, должны сразу представлять дорожную карту исследуемого направления с возможностью сколько угодно подробной детализации. Подход к поиску на основе тематического моделирования, так называемый тематический поиск, позволяет учесть все эти требования и тем самым упорядочить характер работы с информацией, повысить эффективность добычи знаний, избежать когнитивных искажений при восприятии информации, что важно как на микро-, так и на макроуровне. С целью демонстрации примера применения тематического поиска в статье рассматривается задача анализа программы импортозамещения на основе патентных данных. Программа включает планы по 22 отраслям и содержит более 1500 товаров и технологий для предполагаемого импортозамещения. Применение патентного поиска на основе тематического моделирования позволяет осуществлять поиск сразу по блокам априорно задаваемой информации – пунктам отраслевых планов импортозамещения и на выходе получать подборку релевантных документов по каждой отрасли. Данный подход позволяет не только емко представить эффективность реализации программы в целом, но и наглядно получить более детальную информацию о том, какие именно группы продуктов и технологий получали патент.

Ключевые слова: инновационный поиск, тематический поиск, тематическое моделирование, импортозамещение, патентный поиск; патентный анализ, аддитивная регуляризация тематических моделей.

JEL: D83.

Милкова Мария Александровна, научный сотрудник, Центральный экономико-математический институт РАН, Москва, Россия; ORCID 0000-0002-9393-1044; m.a.milkova@gmail.com

[1] Статья подготовлена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 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»).

ВВЕДЕНИЕ

Новая цифровая реальность, сопровождаемая гигантскими темпами накопления информации, ставит перед нами задачу внедрения принципиально новых подходов к поиску и структурированию информации. В контексте всестороннего анализа сложных проблем и ситуации перенасыщения информации концепция привычного итерационного поиска является устаревшей и требует пересмотра. Инновационный поиск должен противопоставить итерационному подход, позволяющий не просто получать результаты в виде ранжированного списка, а представлять дорожную карту, структуру исследуемого направления с последующей возможностью сколь угодно подробной детализации.

Оправданность внедрения инновационного поиска взамен итерационного объясняется несколькими причинами. Во-первых, результатами выдачи итерационного поиска

легко манипулировать и тем самым представлять информацию в искаженном виде, склоняя пользователя к принятию неправильных решений. Результат воздействия, так называемое *когнитивное искажение* (*cognitive bias*) (Kahneman, Frederick, 2002), является свойством человеческого восприятия информации и подтверждено психологическими экспериментами (Kahneman, 2003). В качестве примера можно рассмотреть исследование, проводившееся во время парламентских выборов в Индии в 2014 г., когда избирателям предлагалось расширить свои представления о кандидатах с помощью поисковой системы в Интернете (Helbing, 2019). Однако для одних групп на первой странице выдачи появлялось больше положительных отзывов о кандидате 1, а отрицательные отзывы располагались в основном на последующих страницах. Другими группами аналогично «манипулировали» относительно других кандидатов. В результате для кандидатов, по которым выводилась положительная информация на первой странице, число голосов увеличивалось на 20%.

Другим аспектом является необходимость формулирования длинного, разнородного запроса, при этом максимально учитывающего уже имеющуюся информацию по анализируемой проблеме. Одновременно при исследовании новых областей знаний формулирование поискового запроса в виде ключевых фраз может быть затруднительным.

Таким образом, необходимость пересмотра подхода к поиску информации актуальна как на уровне получения знаний и принятий решений отдельным индивидом, так и на макроуровне. Так, неупорядоченный характер работы с информацией, отсутствие необходимых навыков и инструментария в недавней работе отмечены как одни из ключевых факторов, препятствующих распознаванию грядущих инноваций и предвидению их последствий (Миловидов, 2019).

Важно, что инструментарий для решения описанных выше проблем уже есть. Последнее десятилетие развивается концепция так называемого *тематического поиска* (*topic search*) (Grant, Clint et al., 2015; Eisenstein, 2012), в основе которого лежит *тематическое моделирование* (*topic modeling*) – одно из активно развивающихся с конца 1990-х гг. направлений анализа больших объемов текстовой информации. Тематическая модель определяет структуру коллекции текстовых документов путем выявления скрытых тем в документах, а также *термов* (*terms* – слов или словосочетаний), характеризующих каждую тему. В вероятностном тематическом моделировании документ может с определенными вероятностями относиться сразу к нескольким темам, равно как и терм может с различными вероятностями определять ту или иную тему. Каждый документ описывается дискретным распределением на множестве тем, а каждая тема – дискретным распределением на множестве термов. Представление результатов в таком виде позволяет получить дорожную карту интересуемого направления и существенно повышает точность и полноту поиска (Янина, Воронцов, 2016).

В настоящее время существуют различные подходы к построению тематических моделей (Милкова, 2019; Boyd-Graber, Hu et al., 2017; Daud, Li et al., 2010). В нашей работе использована *концепция аддитивной регуляризации тематических моделей* (Additive Regularization of Topic Models, ARTM) (Vorontsov, Potapenko, 2014), позволяющей расширить стандартные алгоритмы путем добавления различного рода априорной информации и знаний по исследуемой проблеме.

Данная статья демонстрирует пример применения тематического поиска при проведении патентного анализа – неотъемлемой части как форсайт-исследований, так отдельных исследований перспектив инновационного развития, технологических трендов в различных областях и др. В последние годы набирает популярность применение тематического моделирования для анализа патентных данных. Научно-исследовательские публикации в области патентного анализа показывают эффективность как использования методов интеллектуального анализа текстов в целом

(Tseng, Lin, 2007), так и оправданность применения тематического моделирования (Chen, Shang et al., 2016; Tang, Wang et al., 2012). Построение тематических моделей используется для получения общего представления о патентовании в интересующей отрасли (Suominen, Toivanen et al., 2017), выявления технологических трендов (Choi, Song, 2018), разработки отдельных специализированных программных продуктов для проведения тематического патентного анализа (Tang, Wang et al., 2012).

ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

В связи с реализацией государственной программы достижения технологического суверенитета России в 2015 г. был законодательно утвержден ряд положений, регламентирующих планы мероприятий по импортозамещению в 22 отраслях (далее – Программа). В каждой отрасли составлен перечень товаров и технологий, по которым установлен свой показатель доли импорта к 2020 г.

В преддверии окончания установленного для реализации Программы срока, важно представить некоторые результаты импортозамещения, основанные на анализе патентных данных. В период реализации Программы в научном сообществе появился ряд публикаций, оценивающих возможности импортозамещения по тем или иным товарам (Эриванцева, 2016, 2017; Андрейчиков, Тевелева и др., 2019). Несмотря на исключительную важность подробного анализа по каждому направления развития, полезно иметь и общую структуру результатов. Подход, охватывающий сразу все отрасли, позволит продемонстрировать результаты выполнения программы в целом и даст общее представление о состоянии различных отраслей экономики (на основе патентных данных).

Ключевыми документами Программы являются планы импортозамещения (далее – Планы), содержащие перечень продуктов/технологий для импортозамещения, фактические показатели доли импорта до начала реализации Программы и планируемые к 2020 г. показатели[2].

Данные показатели по каждой из 22 отраслей визуализированы на рис. 1 (диаграммы Box-Whiskers). Наиболее «амбициозными» отраслями, заявляющими снижение доли импорта на 80% и более (медианные значения), являются: автомобильная промышленность, транспортное машиностроение, фармацевтическая промышленность, гражданское авиастроение. К отраслям, обозначившим минимальные показатели снижения доли импорта (не более 30%, медианные значения) относятся: детские товары, станкоинструментальная промышленность, энергетическое машиностроение, кабельная и электротехническая промышленность, нефтегазовое машиностроение.

Общий план импортозамещения подразумевал снижение доли импорта с фактического медианного значения 90% до планового медианного значения в 15% (рис. 2).

Анализируя суверенитет России, в работе (Стрельцова, Фурсов и др., 2016) был проведен патентный анализ (2000–2015 гг.), в ходе которого на фоне роста патентной активности отечественных заявителей было отмечено усиление технологической зависимости России от иностранных разработок: если в 2000 г. на долю зарубежных заявителей приходилось 19% патентных заявок на изобретения, поданных в Роспатент, то в 2015 г. – 36%. Реализация планов импортозамещения напрямую связана с выданными патентами по конкретным продуктам/технологиям в соответствующих отраслях. Таким образом, первым этапом в анализе результатов Программы является осуществление патентного поиска по всем позициям каждого Плана.

[2] Отраслевые планы импортозамещения по двадцати двум отраслям промышленности. Минпромторг России, 2015–2018 гг. URL: https://gisp.gov.ru/plan-import-change.

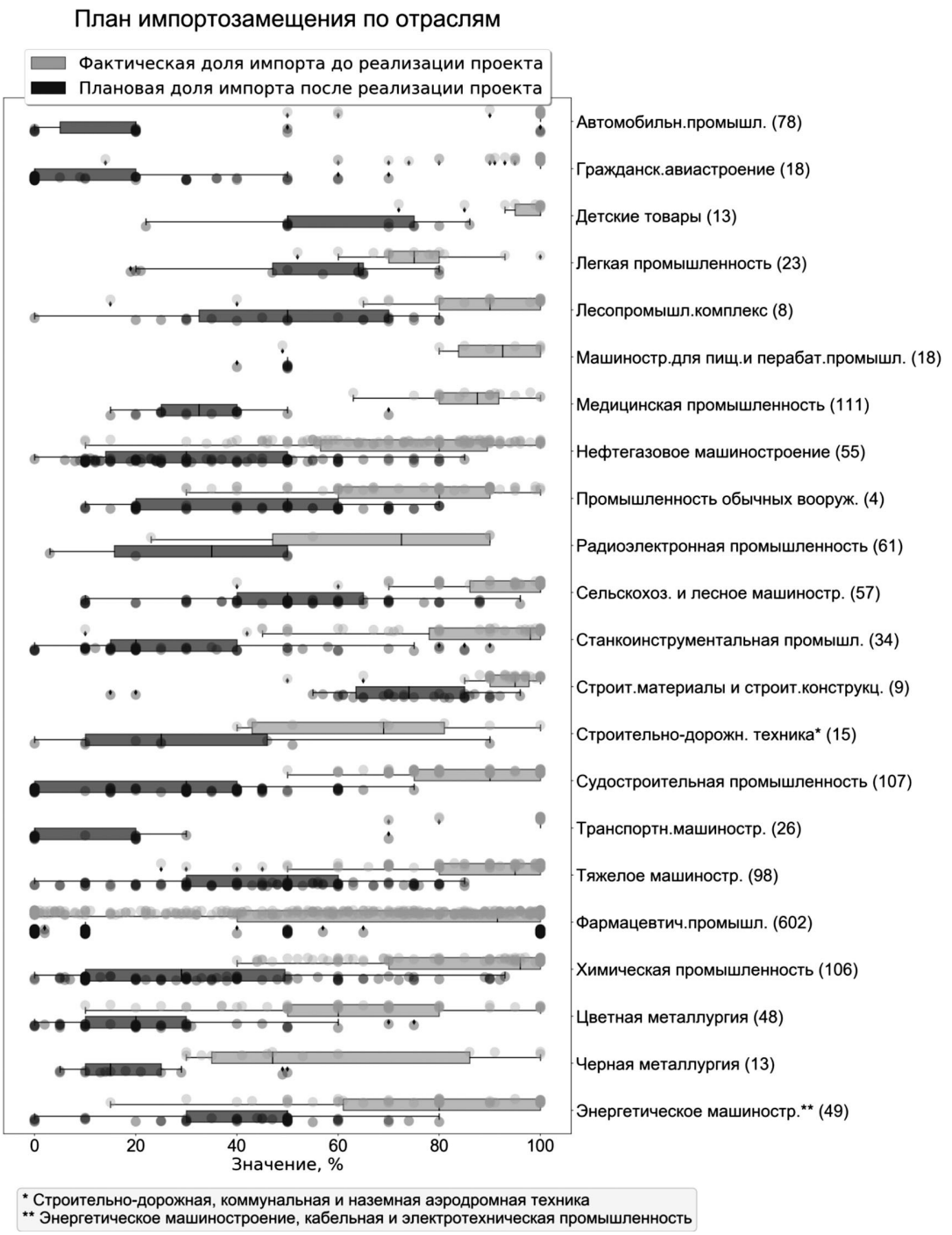

Рис. 1. Диаграмма Box-Whiskers по отраслям для фактической доли импорта до реализации плана по импортозамещению и плановой доли импорта после реализации программы. В скобках приведено число пунктов в каждом отраслевом плане

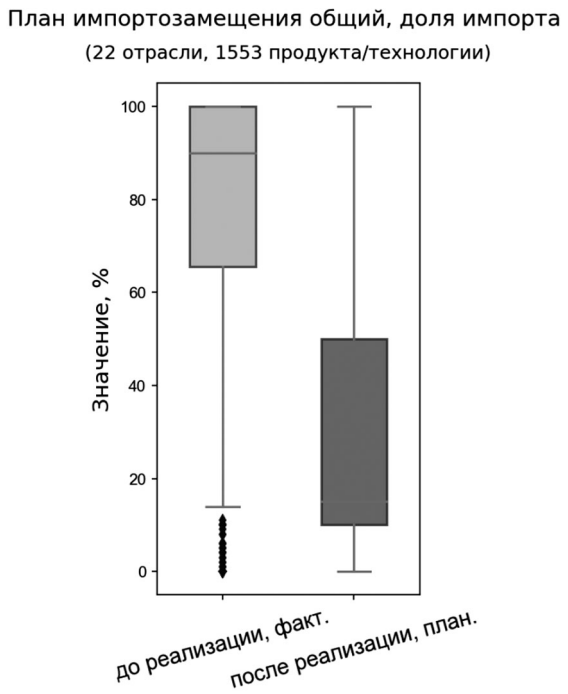

Рис. 2. Диаграмма Box-Whiskers для общей доли импорта до реализации Программы и плановой доли – после ее реализации

Еще пять лет назад в основе большинства систем патентного поиска и анализа лежал консервативный булев поиск (т.е. поиск на основе ключевых слов, с поддержкой языка запросов). Для учета различных написаний ключевых слов применялись нечеткий поиск (т.е. приблизительное совпадение строк) либо лемматизация/стемминг (лемматизация – приведение слов к нормальной форме, например, словоформа «цифровые» преобразовывается к лемме «цифровой»; стемминг – нормализация словоформы к ее квазиоснове, например, вышеприведенная словоформа будет усечена до формы «цифров»). Подобный инструментарий обязывал пользователя конструировать сложные поисковые запросы, что часто не обеспечивало релевантных результатов поиска, особенно для русского языка (Девяткин, Смирнов и др., 2016).

Однако методы компьютерной лингвистики и машинного обучения развиваются стремительно, равно как и растут требования, предъявляемые к современным системам патентного анализа. В настоящее время многие поисковые и аналитические системы (например, Яндекс.Патенты, Google Patents, Patseer) используют передовые достижения компьютерной лингвистики, в том числе методы семантического анализа текстов. Современные поисковые системы способны находить схожие патенты, смежные патенты (в которых упоминается тот или иной интересующий пользователя документ, или иные документы, на которые он ссылается). Поиск похожих патентов осуществляется не только по ключевым словам, но и по смыслу.

Стоит отметить, что современные системы патентного поиска и аналитики удобны, если необходимо получить информацию о небольшом числе объектов. В случае если интересуемых объектов оказывается много, осуществление поиска требует значительных временных вложений.

Для анализа реализации плана по импортозамещению необходимо получить информацию по всем 1553 пунктам Плана, что, очевидно, требует принципиально иного подхода к осуществлению патентного поиска.

МЕТОДИКА ИССЛЕДОВАНИЯ

В качестве метода, позволяющего решить данную проблему, было выбрано *тематическое моделирование* – современное направление, находящееся на стыке компьютерной лингвистики и машинного обучения. Результатом построенной тематической модели является, с одной стороны, мягкая кластеризация патентных документов по заданному числу тем, а с другой, – выявленный набор слов или словосочетаний, характеризующий с определенными вероятностями каждую из тем. Кроме того, в модель можно добавлять некоторую *априорную информацию* об интересующих термах, на основе которых необхо-

димо сформировать *темы* (концепция ARTM). Таким образом, для нашей задачи ядром каждой из тем были слова и словосочетания из отраслевых планов импортозамещения.

Далее опишем ключевые моменты тематического моделирования в целом и ARTM в частности. Пусть есть коллекция документов $\mathcal{D} = \{d_1, \ldots, d_N\}$, каждый из которых представляет собой последовательность n_d терминов ($w_1, \ldots, w_{n_d}$) из словаря $\mathcal{W}$. Термин может повторяться в документе несколько раз.

Предполагается, что существует конечное множество тем T, и каждое употребление терма w в каждом документе d связано с некоторой темой $t \in T$. Термы w и документы d являются наблюдаемыми переменными, тема $t \in T$ является латентной (скрытой) переменной. Коллекция документов рассматривается как случайная и независимая выборка троек (d_i, w_i, t_i), $i = 1, \ldots, N$, из дискретного распределения $p(d, w, t)$ на конечном множестве $\Omega = \mathcal{D} \times \mathcal{W} \times T$ (Воронцов, Потапенко, 2014). Вероятности событий в пространстве Ω совпадают с частотными оценками этих вероятностей.

Совместная вероятностная модель над $\mathcal{D} \times \mathcal{W}$ определяется вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w \mid t)$ и тем в документах $\theta_{td} = p(t \mid d)$:

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

Вероятностная модель (1) описывает процесс порождения коллекции по известным распределениям $p(w|t)$, $p(t|d)$. Задача тематического моделирования – обратная задача: по заданной коллекции D требуется восстановить породившие ее распределения $p(t|d)$, $p(w|t)$ (Воронцов, Потапенко, 2014).

В настоящее время существуют различные подходы к определению оптимальных значений скрытых параметров (Милкова, 2019; Boyd-Graber, Hu et al., 2017; Daud, Li et al., 2010). Доминирующим подходом является *байесовское обучение* – большинство моделей разрабатываются на основе *модели латентного размещения Дирихле* (Blei, Ng et al., 2003). Однако развивается и альтернативный, многокритериальный подход, получивший название «*Аддитивная регуляризация тематических моделей*» (ARTM) (Воронцов, Потапенко, 2014), в основе которого лежит модель PLSA (Probabilistic Latent Semantic Analysis), предложенная Хофмангом (Hofmann, 1999), и в котором модель оптимизируется по взвешенной сумме критериев. Данная работа основывается именно на подходе аддитивной регуляризации.

Подход аддитивной регуляризации представляет задачу тематического моделирования как задачу стохастического матричного разложения заданной матрицы частот слов (термов) в документах:

$$F = (p_{wd})_{\mathcal{W} \times \mathcal{D}}, \quad p_{wd} = p(w|d) = \frac{n_{dw}}{n_d},$$

где n_{dw} – число вхождений терма w в документ d; n_d – длина документа d.

Задача сводится к поиску приближенного представления заданной матрицы частот F в виде произведения двух матриц меньшего размера:

- матрицы термов в темах Φ

$$\Phi = (\varphi_{wt})_{W \times T}, \quad \varphi_{wt} = p(w|t) = \frac{n_{wt}}{n_t};$$

- матрицы тем в документах Θ

$$\Theta = (\theta_{td})_{T \times D}, \quad \theta_{td} = p(t|d) = \frac{n_{td}}{n_d}, \quad F \approx \Phi\Theta.$$

Матрицы F, Φ, Θ – стохастические, т.е. имеют неотрицательные нормированные столбцы, представляющие дискретные распределения.

Для оценивания параметров Φ, Θ тематической модели (1) по коллекции документов $\mathcal{D}$ максимизируется логарифм правдоподобия выборки при ограничениях неотрицательности и нормированности столбцов матриц Φ, Θ:

$$L(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} =$$
$$= \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta};$$

$$\sum_{w\in W}\varphi_{wt}=1,\quad \varphi_{wt}\geq 0; \tag{2}$$

$$\sum_{t\in T}\theta_{td}=1,\quad \theta_{td}\geq 0.$$

Однако искомое стохастическое матричное разложение $\Phi\Theta$ определено не единственным образом, а с точностью до невырожденного преобразования $\Phi\Theta=(\Phi S)(S^{-1}\Theta)$, т.е. задача является некорректно поставленной. Решение такой задачи можно доопределить и сделать устойчивым путем добавления дополнительного критерия – регуляризатора[3], учитывающего специфические особенности данной задачи и знания предметной области. Таким образом, наряду с правдоподобием (2), требуется максимизировать r критериев-регуляризаторов $R_i(\Phi,\Theta)$, $i = 1,\ldots r$, с неотрицательными коэффициентами регуляризации τ_i. Таким образом, оптимизационная задача примет вид

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}+$$
$$+R(\Phi,\Theta)\to\max_{\Phi,\Theta};$$

$$R(\Phi,\Theta)=\sum_{i=1}^{r}\tau_i R_i(\Phi,\Theta);$$

$$\sum_{w\in W}\varphi_{wt}=1,\quad \varphi_{wt}\geq 0; \tag{3}$$

$$\sum_{t\in T}\theta_{td}=1,\quad \theta_{td}\geq 0.$$

Решение задачи (3) строится на основе так называемого регуляризованного *EM*-алгоритма. На первом этапе выбирается начальное приближение для φ_{wt}, θ_{td}. На *E*-шаге вычисляются вспомогательные переменные p_{tdw}:

$$p_{tdw}=norm_{t\in T}(\varphi_{wt}\theta_{td}), \tag{4}$$

где оператор нормировки *norm* преобразует произвольный вектор в вектор вероятностей дискретного распределения путем обнуления отрицательных элементов и нормировки.

На *M*-шаге вычисляются частотные оценки максимального правдоподобия для искомых условных вероятностей φ_{wt}, θ_{td}:

$$\varphi_{wt}=norm_{w\in W}\left(n_{wt}+\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\right),$$

$$n_{wt}=\sum_{d}p_{tdw}n_{dw}; \tag{5}$$

$$\theta_{td}=norm_{t\in T}\left(n_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right),$$

$$n_{td}=\sum_{w}n_{dw}p_{tdw}. \tag{6}$$

Вычисления (4)–(6) продолжаются в цикле до сходимости.

Задача адекватного описания текстов на естественном языке и выделения интерпретируемых тем накладывает на вероятностную тематическую модель большое число требований. Регуляризаторы направлены на учет лингвистических особенностей текста и повышение интерпретируемости тем. Наиболее распространенными являются регуляризаторы разрежения, сглаживания и декоррелирования тем (Воронцов, Потапенко, 2014).

Применительно к решению нашей задачи регуляризатор необходимо построить так, чтобы сгруппировать термы каждого из Планов в своей теме, для чего подходит регуляризатор сглаживания. Опуская некоторые преобразования, укажем, что регуляризатор сглаживания приводит к модифицированному *M*-шагу:

$$\varphi_{wt}=norm_{w\in W}(n_{wt}+\beta_w); \tag{7}$$

$$\theta_{td}=norm_{t\in T}(n_{td}+\alpha_t). \tag{8}$$

Для нашей задачи данная модификация *M*-шага привела к тому, что на каждой итерации *EM*-алгоритма к частотам термов, относящихся к термам из «белого списка» (термы

[3] Регуляризация – общепринятый метод добавления некоторых дополнительных ограничений к условию с целью решения некорректно поставленной задачи (см., например, (Тихонов, Арсенин, 1986)).

Планов), добавлялся параметр β_w. Значения параметров β_w, α_t выбираются экспериментально.

Помимо построения модели на основе текста, ARTM-подход позволяет строить так называемую мультимодальную тематическую модель. Под модальностями понимаются метаданные, так или иначе характеризующие тематику текста. К модальностям могут относиться: биграммы (*n*-граммы), тэги, классы, авторы, цитируемые или цитирующие документы и т.п.

Каждая модальность имеет свой словарь термов W^m, $m = 1, \ldots, M$. В мультимодальной ARTM функция логарифма правдоподобия вводится для каждой модальности (Vorontsov, Frei et al., 2015):

$$L_m\left(\Phi^m, \Theta\right) = \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p(w \mid d) \rightarrow$$
$$\rightarrow \max_{\Phi^m, \Theta}, \qquad (9)$$

где n_{dw} – число вхождений терма $w \in W^m$ в документ d.

Таким образом, в мультимодальной модели матрица Φ определяется для каждой модальности, а матрица Θ является общей. Оптимизационная задача представляет собой максимизацию взвешенной суммы логправдоподобий и *r*-регуляризаторов при условиях нормировки и неотрицательности столбцов матриц Φ^m, Θ (Янина, Воронцов, 2016):

$$\sum_{m \in M} \frac{\tau_m}{n_m} \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} +$$
$$+ \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta};$$
$$\sum_{w \in W^m} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0;$$
$$\sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0,$$

где τ_m – вес модальности m; $n_m = \sum_{d \in D} \sum_{w \in W^m} n_{dw}$ – нормировочный множитель.

Основная задача при построении аддитивно регуляризованной тематической модели заключается в подборе *траектории регуляризации* – функции коэффициента регуляризации от номера итерации и критериев качества модели. Траектории регуляризации подбираются, с учетом анализа их влияния на критерии качества модели в ходе итераций. В данной работе решение о выборе оптимальной модели принималось на основе следующих автоматически вычисляемых метрик: степень разреженности матриц Φ, Θ (доля нулевых элементов в матрице), размер ядра темы $|W_t|$ (множество слов, имеющих высокую условную вероятность, $W_t = \{w \in W \mid p(t \mid w) > 0{,}25\}$), чистота темы (насколько определяющими являются термы внутри темы – вычисляется суммарная вероятность термов ядра темы $purity_t = \sum_{w \in W_t} p(w \mid t)$), контрастность темы (насколько хорошо ядро темы отличает ее от остальных тем, т.е. вероятность встретить термы ядра именно в данной теме $contr_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w)$).

ПОСТРОЕНИЕ МОДЕЛИ

В качестве инструмента построения тематической модели была выбрана библиотека с открытым кодом BigARTM (bigartm.org) (Frei, Apishev, 2016) и реализованная в среде Python.

Для построения тематической модели были собраны патенты на изобретения и полезные модели, выданные за 3,5-летний период (январь 2016 г. – июнь 2019 г.) – всего 152 718 документов: 120 768 изобретений и 31 950 полезных моделей.

Тематическая модель строилась на основе Названий и Абстрактов патентов, представленных в виде *униграмм* (т.е. одиночных слов). В отдельную модальность были выделены наиболее частотные биграммы (двусловные словосочетания с частотой встре-

чаемости в Названии и Абстракте более или равной двум), оптимальный вес модальности биграмм определялся экспериментально и был выбран равным $\tau_m = 5$.

Было введено 22 регуляризатора сглаживания, по одному для каждой из тем. Коэффициент сглаживающих регуляризаторов был выбран равным $1e + 7$.

Итоговая модель обладала следующими метриками качества: доля разреженных элементов матриц униграмм $\Phi^1 = 0{,}994$; биграмм $\Phi^2 = 0{,}998$; $\Theta = 0{,}818$. Размер ядра равен 628; средняя чистота тем 0,992; средняя контрастность тем 0,976. Общее число итераций: 40.

РЕЗУЛЬТАТЫ

Итогом построенной модели стала подборка патентных документов по каждой из 22 отраслей в соответствии с темой, характеризуемой набором слов и словосочетаний из соответствующего Плана.

Помимо стандартных автоматически вычисляемых метрик, качество модели также оценивалась с помощью асессоров, определяющих, насколько релевантным является отобранный документ. Патентному документу ставилось в соответствие значение качества

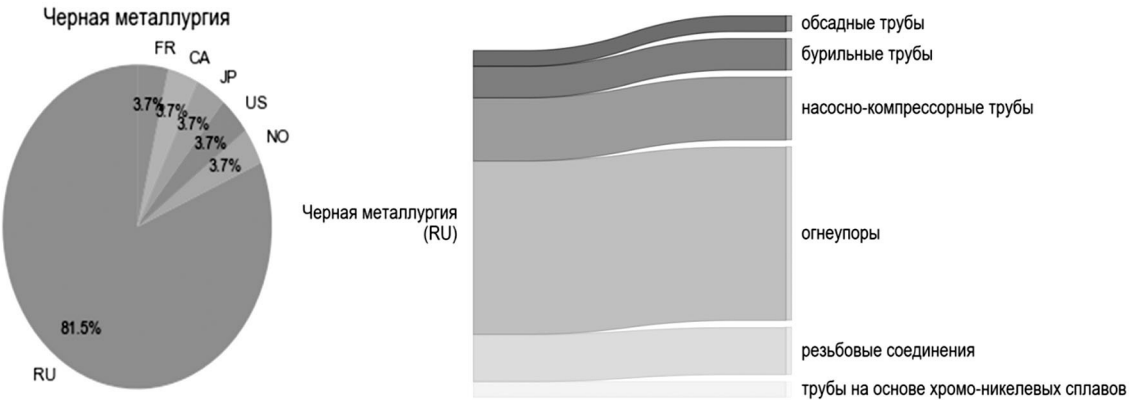

Рис. 3. Черная металлургия. Распределение по странам-патентообладателям и категориям патентования российскими патентообладателями

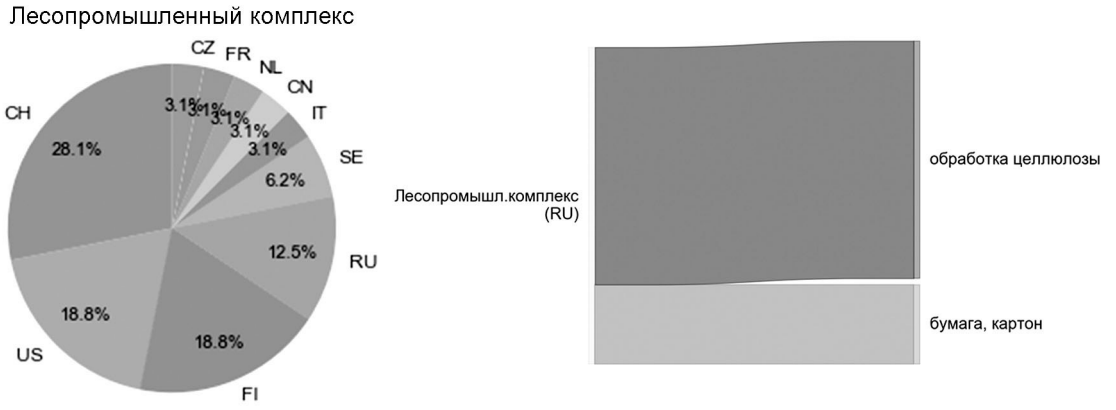

Рис. 4. Лесопромышленный комплекс. Распределение по странам-патентообладателям и категориям патентования российскими патентообладателями

(quality) $q = 1$ в случае, если патент точно соответствовал одному из заявленных в Плане пункту импортозамещения; значение $q = 0{,}5$ присваивалось в случае, если патент связан с одним из пунктов Плана; $q = 0$ – если не соответствовал ни одному пункту Плана. Данная методика успешно применялась в работах (Янина, Воронцов, 2016; Apishev, Koltsov et al., 2016) и позволяла, единожды сделав разметку результатов поиска, многократно вычислять оценки качества тематического поиска для различных реализаций модели.

Для документов со значениями $q = 1$, $q = 0{,}5$ была выделена ключевая фраза/слово, характеризующая как документ, так и его принадлежность к тому или иному пункту Плана. Какие именно позиции импорта были замещены в каждой отрасли, наглядно показано на диаграммах Sankey (для краткости приведены диаграммы для двух отраслей).

В табл. 1 агрегированы результаты всех отраслей.

Подобный вариант кластеризации удобен, так как позволяет емко представить структуру коллекции патентных документов, сформированную уже на основе только тех слов и словосочетаний, которые представляют для нас интерес.

Таблица 1

Характеристики импортозамещения на основании патентных данных

Отрасль	Доля российских патентов, %	Категории импортозамещения (*RU*)	Число категорий, *k*	Средний балл, *q*	Суммарное значение балла
Автомобильная промышленность	74,1	Двигатель внутреннего сгорания	1	0,93	18,5
Гражданское авиастроение		–	–	–	–
Детские товары	95,2	Мебель для детей; игры и игрушки; спортивные комплексы; детская одежда; детское творчество	5	0,95	19,0
Легкая промышленность	52,2	Нетканые материалы; защитная одежда; переработка шерсти	3	0,96	11,0
Лесопромышленный комплекс	12,5	Обработка целлюлозы; бумага, картон	2	0,75	3,0
Машиностроение для пищевой и перабатывающей промышленности	100,0	Обработка зерновых	1	0,83	2,5
Медицинская промышленность	41,7	Стерилизация и дезинфекция; эндоскопические аппараты; иглы инъекционные; имплантируемые насосы	4	0,90	4,5
Нефтегазовое машиностроение	78,3	Катализаторы гидроочистки; бурение скважин; катализаторы гидрокрекинга; переработка углеводородного сырья; гидроразрыв пласта; катализаторы каталитического крекинга	6	1,00	18,0
Промышленность обычных вооружений	82,4	Патроны; спортивное оружие	2	0,79	11,0
Радиоэлектронная промышленность		–	–	–	–
Сельскохозяйственное и лесное машиностроение	83,3	Подшипники; зерноуборочный комбайн; пресс подборщик	3	0,70	17,5

Окончание таблицы

Отрасль	Доля российских патентов, %	Категории импортозамещения (*RU*)	Число категорий, *k*	Средний балл, *q*	Суммарное значение балла
Станкоинструментальная промышленость	78,9	Фрезерный станок; токарный станок; расточный станок; шпиндели; финишное шлифование; гидроабразивная резка; станки ЧПУ	7	0,93	14,0
Строительные материалы и строительные конструкции	96,6	Керамическая масса для плитки; теплоизоляционные материалы; щебеночно-мастичные асфальтобетоны	3	0,68	19,0
Строительно-дорожная техника	93,3	Дорожное покрытие; гидравлическое оборудование; фронтальные погрузчики; бульдозеры; фронтальный погрузчик; экскаватор; прицеп и полуприцеп; крановое шасси; коммунальная техника	9	0,79	11,0
Судостроительная промышленность	92,9	Движитель; гребневой винт	2	0,50	6,5
Транспортное машиностроение	62,7	Вагон-цистерна; тормозная система; тележки вагона; крытый вагон	4	0,81	26,0
Тяжелое машиностроение	50,0	Крепь горная; холодильные установки	2	0,75	1,5
Фармацевтическая промышленность	56,3	Инозин + никотинамид + рибофлавин + янтарная кислота; висмут калий аммоний цитрат; дротаверин; йогексол; лопинавир + ритонавир; этилметилгидроксипиридина сукцинат; рокурония бромид; дигоксин; 1 карбамоилметил 4 фенил 2 пирролидон; фенспирид; изониазид; лаппаконитина гидробромид; иммуноглобулин стандартный; бромдигидрохлорфенил-бензодиазепин; десмопрессин; финголимод; анастрозол	17	0,94	17,0
Химическая промышленность	87,5	Лакокрасочные материалы; уплотнительные материалы; эпоксидный композит; клеевые материалы; полиэтилентерефталат; сверхвысокомолекулярный полиэтилен; полимерные композиты	7	0,79	11,0
Цветная металлургия	88,6	Алюминиевый сплав; алюминий, электролиз; алюминиевая лигатура; гидроксид алюминия; алюминиевый порошок; алюминиевая фольга; алюминиевые прутки; анодная масса	8	0,81	25,0
Черная металлургия	81,5	Огнеупоры; насосно-компрессорные трубы; резьбовые соединения; бурильные трубы; трубы на основе хромоникелевых сплавов; обсадные трубы	6	0,89	19,5
Энергетическое машиностроение	50,0	Трансформаторы тока	1	1,00	1,0

В зависимости от числа найденных релевантных патентных документов, степени их соответствия заявленному плану, а также общему количеству пунктов Плана, каждая из отраслей получила свой рейтинг:

$$\text{Score} = \sum(q) \cdot k / N,$$

где k – число категорий – различных позиций Плана, по которым были найдены релевантные патентные документы ($q = 1$, $q = 0{,}5$); N – общее число пунктов Плана.

Результаты ранжирования отраслей представлены на рис. 5.

Итак, в контексте выданных патентов на изобретения и полезные модели к отраслям, продемонстрировавшим лучшие показатели импортозамещения можно отнести черную металлургию, детские товары, строительно-дорожную, коммунальную и наземную аэродромную технику, строительные материалы и строительные конструкции, промышленность обычных вооружений, цветную металлургию, транспортное машиностроение.

Отрасли, которые на данный момент не способны соответствовать плану импортозамещения (на основании выданных патентных документов): гражданское авиастроение, радиоэлектронная промышленность, энергетическое машиностроение, кабельная и электротехническая промышленность, тяжелое машиностроение, судостроительная промышленность, машиностроение для пищевой и перерабатывающей промышленности, медицинская промышленность.

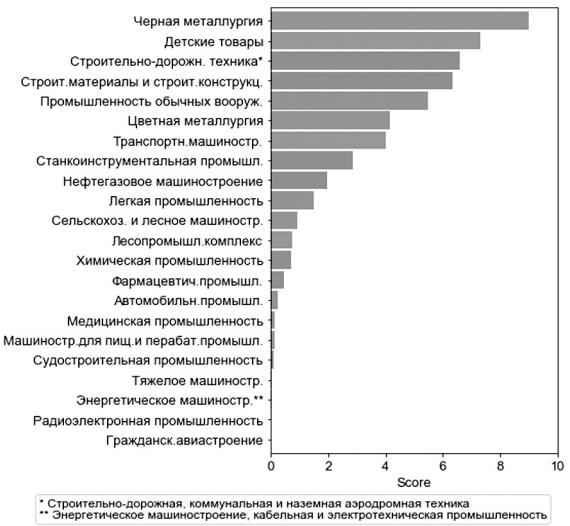

Рис. 5. Рейтинг отраслей импортозамещения на основании выданных патентов

Важно, что полученная структура позволяет при необходимости детализировать результаты. Например, выявлять долю индивидуальных патентообладателей, которые не смогут стать основными агентами захвата ниш рынка и не смогут конкурировать с крупными иностранными компаниями, долю недействующих патентов и т.д.

ОБСУЖДЕНИЕ И ВЫВОДЫ

Полученные результаты демонстрируют эффективность применения нового метода патентного поиска, основанного на тематическом моделировании. Подход позволяет осуществлять поиск сразу по блокам априорно задаваемой информации (в данном случае – пункты сразу всех 22 отраслевых планов импортозамещения) и на выходе получать подборку релевантных документов по каждой из отраслей. Данный подход является своего рода «крупным планом» патентного поиска, который может как служить конечной целью, так и являться отправной точкой для более детального анализа.

Задачи кластеризации и классификации больших объемов текстовых данных с целью получения информации о структуре коллекции документов встречаются довольно часто: это и библиометрический анализ (Гибсон, Дайм и др., 2018), и кластеризация пользователей социальных сетей (Halibas, Shaffi et al., 2018), и анализ дискурса и тональности сообщений (Krishna, Aich et al., 2018; Apishev, Koltsov et al., 2016), и анализ юридических документов (Sulea, Zampieri et al., 2017) и др. Однако в современной непрерывно меняющейся цифровой реальности темпы нако-

пления информации настолько стремительны, что требует от нас пересмотра подходов к смысловой компрессии информации. Так, например, в контексте цифровизации, по мнению К. Шваба, «характер происходящих изменений настолько фундаментален, что мировая история еще не знала подобной эпохи – времени как великих возможностей, так и потенциальных опасностей» (Шваб, 2016). Для того чтобы всесторонне охватить и проанализировать весь спектр происходящих изменений, к методам поиска информации необходимо предъявлять повышенные требования. Инновационный подход к поиску должен гибко учитывать большой объем уже накопленных знаний и априорные требования к результатам. Результаты, в свою очередь, должны сразу представлять дорожную карту исследуемого направления с возможностью сколько угодно глубокой детализации. Подход на основе тематического моделирования позволяет учесть все эти требования и тем самым упорядочить характер работы с информацией, повысить эффективность добычи знаний, избежать когнитивных искажений при восприятии информации, что важно как на микро-, так и на макроуровне.

Список литературы / References

*Андрейчиков А.В., Тевелева О.В., Неволин И.В., Милкова М.А., Кравчук И.*С. Методика проведения поисковых исследований по выявлению возможностей импортозамещения высокотехнологичной продукции на основе мировых патентных и финансовых информационных ресурсов // Экономика и предпринимательство. 2019. № 4. С. 157–167. [Andrejchikov A.V., Teveleva O.V., Nevolin I.V., Milkova M.A., Kravchuk I.S. (2019). Methodology for conducting search research to identify opportunities for import substitution of high-tech products based on world patent and financial information resources. *Ekonomika i Predprinimatel'stvo*, no. 4, pp. 157–167 (in Russian).]

Гибсон Э., Дайм Т., Гарсес Э., Дабич М. Библиометрический анализ как инструмент выявления распространенных и возникающих методов технологического Форсайта // Форсайт. 2018. Т. 12. № 1. С. 6–24. [Gibson Je., Dajm T., Garses Je., Dabich M. (2018). Bibliometric analysis as a tool for identifying common and emerging methods of technological Foresight. *Forsajt*, vol. 12, no. 1, pp. 6–24 (in Russian).]

Девяткин Д.А., Смирнов И.В., Соченков И.В., Тихомиров И.А. Современные методы компьютерной лингвистики для патентного поиска и анализа // Интеллектуальная собственность. Промышленная собственность, Специальный выпуск. 2016. № 1. С. 71–77. [Devjatkin D.A., Smirnov I.V., Sochenkov I.V., Tihomirov I.A. (2016). Modern methods of computer linguistics for patent search and analysis. *Intellektual'naja Sobstvennost'. Promyshlennaja Sobstvennost'. Special'nyj Vypusk*, no. 1, pp. 71–77 (in Russian).]

Милкова М.А. Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. 2019. № 1 (5). С. 57–69. [Milkova M.A. (2019). Topic models as a tool for distance reading. *Cifrovaja Ekonomika*, no. 1 (5), pp. 57–69 (in Russian).]

Миловидов В. Услышать шум волны: что мешает предвидеть инновации? // Форсайт. 2019. Т. 12. № 1. С. 88–97. [Milovidov V. (2019). Hearing the sound of the wave: What makes it difficult to anticipate innovation? *Forsajt*, vol. 12, no. 1, pp. 88–97 (in Russian).]

Стрельцова Е.А., Фурсов К.С., Чулок А.А. Анализ патентной информации как инструмент выявления и оценки технологического профиля страны // Интеллектуальная собственность. Промышленная собственность. Специальный выпуск. 2016. № 1. С. 63–70. [Strel'cova E.A., Fursov K.S., Chulok A.A. (2016). Analysis of patent information as a tool for identifying and evaluating the technological profile of a country. *Intellektual'naja Sobstvennost'. Promyshlennaja Sobstvennost'. Special'nyj vypusk*, no. 1, pp. 63–70 (in Russian).]

Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука. 1986. 287 с. [Tihonov A.N., Arsenin V.Ya. (1986). Metody resheniya nekorrektnyh zadach. Moscow, Nauka, 287 p. (in Russian).]

Шваб К. Четвертая промышленная революция. М.: Эксмо, 2016. С. 208. [Shvab K. (2016). The fourth Industrial Revolution. Moscow, Jeksmo, p. 208 (in Russian).]

Эриванцева Т.Н. Применение патентного анализа для оценки перспектив импортозамещения на примере отечественных ранорасширителей и сшивающих изделий // Экономика науки. 2016. № 4. С. 261–275. [Jerivanceva T.N. (2016). The use of patent analysis to assess the prospects of import substitution on the example of domestic retractors and crosslinking products. *Ekonomika Nauki*, no. 4, pp. 261–275 (in Russian).]

Эриванцева Т.Н. Оценка конкурентоспособности российских научно-технологических заделов в области создания медицинских инструментов // Экономика науки. 2017. № 1. С. 53–69. [Jerivanceva T.N. (2017). Assessment of the competitiveness of Russian scientific and technological backlogs in the field of creating medical instruments. *Ekonomika Nauki*, no. 1, pp. 53–69 (in Russian).]

*Янина А.О., Воронцов К.*В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных. 2016. Т. 2. № 2. С. 173–186. [Janina A.O., Voroncov K.V. (2016). Multimodal topic models for exploratory search in a collective blog. *Mashinnoe Obuchenie i Analiz Dannyh*, vol. 2, no. 2, pp. 173–186 (in Russian).]

Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K. Mining ethnic content online with additively regularized topic models // Computación y Sistemas. 2016. Vol. 20. No. 3, pp. 387–403.

Blei D., Ng A., Jordan M. Latent dirichlet allocation // Journal of Machine Learning Research. 2003. No. 3.

Boyd-Graber J., Hu Y., Mimmo D. Applications of topic models // Foundations and Trends in Information Retrieval. 2017. P. 1–154.

Chen L., Shang W., Yang G., Zhang J., Lei X. A topic model integrating patent classification information for patent analysis // Geomatics and Information Science of Wuhan University. 2016. Vol. 41. P. 123–126.

Choi D., Song B. Exploring technological trends in logistics: Topic modeling-based patent analysis // Sustainability. 2018. No. 10 (8). P. 2810.

Daud A., Li J., Zhu L., Muhammad F. A generalized topic modeling approach for maven search. In: Li Q., Feng L., Pei J., Wang S.X., Zhou X., Zhu QM. (eds.) Advances in data and web management. APWeb 2009. WAIM 2009. Lecture Notes in Computer Science. 2009. Vol 5446. Berlin, Heidelberg: Springer.

*Eisenstein J., Chau D.H., Kittur A., Xing E.*P. TopicViz: Interactive topic exploration in document collections. Proceeding of CHI EA '12. Extended Abstracts on Human Factors in Computing Systems. 2012. P. 2177–2182.

Frei O., Apishev M. Parallel non-blocking deterministic algorithm for online topic modeling. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. 2016. Vol. 661. Springer, Cham.

Grant C.E., Clint P.G., Virupaksha K., Nirkhiwale S., Wilson J.N., Wang D.Z. A topic-based search, visualization, and exploration system. Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference. 2015. P. 43–48.

Halibas A.S., Shaffi A.S., Mohamed M.A. Application of text classification and clustering of Twitter data for business analytics // Majan International Conference (MIC). Muscat. 2018. P. 1–7.

Helbing D. Towards digital enlightenment: Essays on the dark and light sides of the digital revolution. Springer, Cham, 2019.

Hofmann T. Probabilistic latent semantic analysis. Uncertainty in Artificial Intelligence. Stockholm, UAI'99, 1999.

Kahneman D., Frederick S. Representativeness revisited: Attribute substitution in intuitive judgment. In: T. Gilovich, D. Griffin, D. Kahneman (eds.). Heuristics and biases. New York, Cambridge University Press, 2002. P. 49–81.

Kahneman D. A perspective on judgment and choice: Mapping bounded rationality // American Psychologist. 2003. No. 58 (9). P. 697–720.

Krishna A., Aich A., Akhilesh V., Hegde C. Analysis of customer opinion using machine learning and NLP techniques // International Journal of Advanced Studies of Scientific Research. 2018. Vol. 3(9).

Sulea O.-M., Zampieri M., Malmasi S., Vela M., Dinu L.P., Genabith J. Exploring the use of text classification in the legal domain // Proceedings of the 2nd

Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL), 2017.

Suominen A., Toivanen H., Seppänen M. Firms' knowledge profiles: Mapping patent data with unsupervised learning // Technological Forecasting and Social Change. 2017. Vol. 115. P. 131–142.

Tang J., Wang B., Yang Y., Hu P., Zhao Y., Yan X., Gao B., Huang M., Xu P., Li W., Usadi A.K. PatentMiner: Topic-driven patent analysis and mining // KDD'12. August 12–16. 2012. Beijing, 2012. P. 1366–1374.

Tseng Y.-H., Lin C.-J. Text mining techniques for patent analysis // Information Processing & Management. 2007. No. 43. P. 1216–1247.

Vorontsov K.V., Potapenko A.A. Additive regularization of topic models // Machine Learning Journal, Special Issue «Data Analysis and Intelligent Optimization». Springer. 2014. P. 1–21.

Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M. Bigartm: Open source library for regularized multimodal topic modeling of large collections // AIST'2015, Analysis of Images, Social networks and Texts. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015. P. 370–384.

Рукопись поступила в редакцию 16.10.2019 г.

INNOVATIVE APPROACH TO INFORMATION SEARCH BY EXAMPLE OF A PATENT ANALYSIS OF AN IMPORTANT SUBSTITUTION PLAN

M.A. Milkova

DOI: 10.33293/1609-1442-2020-1(88)-143-157

Maria A. Milkova, Central Economics and Mathematics Institute of the Russian Academy of Sciences, Moscow, Russia; ORCID 0000-0002-9393-1044; m.a.milkova@gmail.com

This article was prepared with the financial support of the Russian Foundation of Basic Research (project No. 19-010-00293 «Development of methodology, economic and mathematical models, methods and decision support systems for search research to identify opportunities for import substitution of high-tech products based on world patent and financial information resources»).

Nowadays the process of information accumulation is so rapid that the concept of the usual iterative search requires revision. Being in the world of oversaturated information in order to comprehensively cover and analyze the problem under study, it is necessary to make high demands on the search methods. An innovative approach to search should flexibly take into account the large amount of already accumulated knowledge and a priori requirements for results. The results, in turn, should immediately provide a roadmap of the direction being studied with the possibility of as much detail as possible. The approach to search based on topic modeling, the so-called topic search, allows you to take into account all these requirements and thereby streamline the nature of working with information, increase the efficiency of knowledge production, avoid cognitive biases in the perception of information, which is important both on micro and macro level. In order to demonstrate an example of applying topic search, the article considers the task of analyzing an import substitution program based on patent data. The program includes plans for 22 industries and contains more than 1,500 products and technologies for the proposed import substitution. The use of patent search based on topic modeling allows to search immediately by the blocks of a priori information – terms of industrial plans for import substitution and at the output get a selection of relevant documents for each of the industries. This approach allows not only to provide a comprehensive picture of the effectiveness of the program as a whole, but also to visually obtain more detailed information about which groups of products and technologies have been patented.

Keywords: innovative search, topic search, topic modeling; import substitution, patent search, patent analysis, additive regularization of topic models.

JEL: D83.

Manuscript received 16.10.2019