1.6. ТЕМАТИЧЕСКИЕ МОДЕЛИ КАК ИНСТРУМЕНТ «ДАЛЬНЕГО ЧТЕНИЯ»

Милкова М.А., научный сотрудник,
Центральный экономико-математический институт РАН

Статья представляет собой обзор подходов к тематическому моделированию – современному направлению исследования больших текстовых коллекций. В настоящее время сверхвысокие темпы накопления информации приводят к тому, что при изучении той или иной темы пользователю становится все труднее разобраться в исследуемом предмете. Таким образом, актуальным вопросом является смысловая компрессия информации – своего рода «дальнее чтение» – необходимое условие получения знаний в условиях стремительного разрастания доступного объема информации. «Дальнее чтение» может быть реализовано с помощью тематического моделирования –направления, находящегося на стыке компьютерной лингвистики и машинного обучения и призванного определять структуру коллекции текстовых документов путем выявления скрытых тем в документах, а также термов (слов или словосочетаний), характеризующих каждую из тем.

«Мы умеем читать тексты,
теперь нужно научиться не читать их»
Ф. Моретти

1. Введение

Тематическое моделирование – современный инструмент, позволяющий автоматически выявлять тематическую структуру больших текстовых коллекций, что является актуальной задачей в эпоху больших интернет-данных. В настоящее время процесс накопления информации настолько стремителен, что простого информационного поиска уже недостаточно для оперативного и адекватного получения информации. Так, например, ответ на вопрос «где находится передний край науки по данной теме» по-прежнему требует времени, квалификации и личного общения с экспертами (Воронцов, 2016). Разработка новых методов и алгоритмов автоматической обработки естественного языка по-прежнему не решает задачи смысловой компрессии, не позволяет получить «дорожную карту» интересуемого направления. Несмотря на высокий уровень современных поисковых систем, сама концепция итерационного поиска уже кажется устаревший. Ставя перед собой задачу поиска знаний в новой области, исследователь вынужден долго карабкаться по лестнице, в которой становится все больше сломанных ступенек.

Интересным является факт, что вопрос о смысловой компрессии возникает и при анализе литературы в целом. Так, известный литературовед, социолог и историк литературы Ф. Моретти[1] предлагает новый принцип изучения литературы —«дальнее чтение» (distant reading)[2] , противопоставленный привычному «пристальному чтению» (close reading), и использует его для работы с большими корпусами текстов, обычно остающимися за пределами внимания (и возможностей) исследований, применяющих более традиционную оптику. В своих работах Моретти подчеркивает необходимость установления связи между анализом и синтезом литературы, однако он отмечает, что в таком случае история литературы будет историей «из вторых рук» —«мозаика, состоящая из исследований других людей, без какого-либо непосредственного прочтения текстов» (Моретти, 2016). Продолжая объяснять концепцию дальнего чтения, Моретти (2016) заключает: «Мы умеем читать тексты, теперь нужно научиться не читать их. Дальнее чтение, для которого расстояние, повторюсь, является условием получения знаний, дает возможность сосредоточиться на единицах, намного больших или намного меньших, чем текст: приемах, темах, тропах или же жанрах и системах. И если в промежутке между очень маленьким и очень большим сам текст исчезнет — что ж, это будет одним из случаев, когда позволительно сказать: «Меньше значит больше» (less is more). Если мы хотим понять, как устроена система в своей целостности, то нужно быть готовым потерять что-то. За теоретизирование всегда приходится расплачиваться: действительность неизмерима в своем разнообразии, концепции же абстрактны и скудны. Однако именно их «скудость» и позволяет овладеть ими и, следовательно, познать. Именно поэтому меньше действительно значит больше».

[1] Франко Моретти (р. 1950) – итальянский литературовед, профессор Стэнфордского университета, автор десятка книг по истории романа, социологии литературы, проблемам точных методов в литературоведении. Моретти является центральной фигурой в активно развивающихся цифровых гуманитарных науках (digital humanities), которые меняют взгляд на изучение целых пластов культурной продукции.

[2] Перевод «Distant reading» на русский язык как «дальнее чтение», возможно, несколько утрачивает терминологическое звучание понятия. «Distant» употребляется в значении «удаленный», подразумевая, что чем дальше мы находимся от объектов, тем большим обзором обладаем.

Мы не случайно приводим такую длинную выдержку из книги Моретти – сверхвысокие темпы накопления текстовой информации, а также стремительное развитие новых методов в области анализа текстов и обработки естественного языка позволяют по-новому взглянуть на смысловую компрессию текста как на своего рода «дальнее чтение».

«Дальнее чтение» невозможно без перехода на новый уровень поиска по сверхбольшим корпусам текста. Последнее десятилетие развивается новая парадигма так называемого разведочного поиска – Exploratory Search (White and Roth, 2009). И если современные поисковые системы отвечают на короткие четко сформулированные запросы, то разведочный поиск характеризует отсутствие точной формулировки запроса и отсутствие единого ответа (Янина и Воронцов, 2016). Если терминология области заранее не определена, а перед исследователем стоит задача представить структуру интересуемой области, получить дорожную карту направления, разведочный поиск должен быть инструментом, посредством которого возможно «дальнее чтение» – смысловая компрессия информации.

Итак, мы предполагаем, что «дальнее чтение», о котором говорил Моретти, может быть реализовано с помощью тематического моделирования – современного направления, находящегося на стыке компьютерной лингвистики и машинного обучения и призванного определять структуру коллекции текстовых документов путем выявления скрытых тем в документах, а также термов (слов или словосочетаний), характеризующих каждую из тем.

Построение тематической модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров— тем. Тема —результат би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. Обычно выполняется нечёткая кластеризация, то есть документ может принадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием (Daud et.al., 2010).

Тематическое моделирование интенсивно развивается с конца 90-х годов. Предложено множество моделей для решения самых разнообразных задач: тематическая сегментация текстов, классификация и категоризация документов, многоязычный информационный поиск, поиск тематической структуры в сообществах, анализ тональности, тематическая визуализация больших текстовых коллекций и др. Тематические модели могут учитывать различные особенности языка и текстовых коллекций. Существуют модели, выявляющие ключевые фразы, учитывающие морфологию слов и синтаксическую структуру предложений, а также различные характеристики (модальности) документов – авторство, тэги, ссылки и др., отслеживающие изменения тем во времени, строящие иерархические отношения между темами и др.

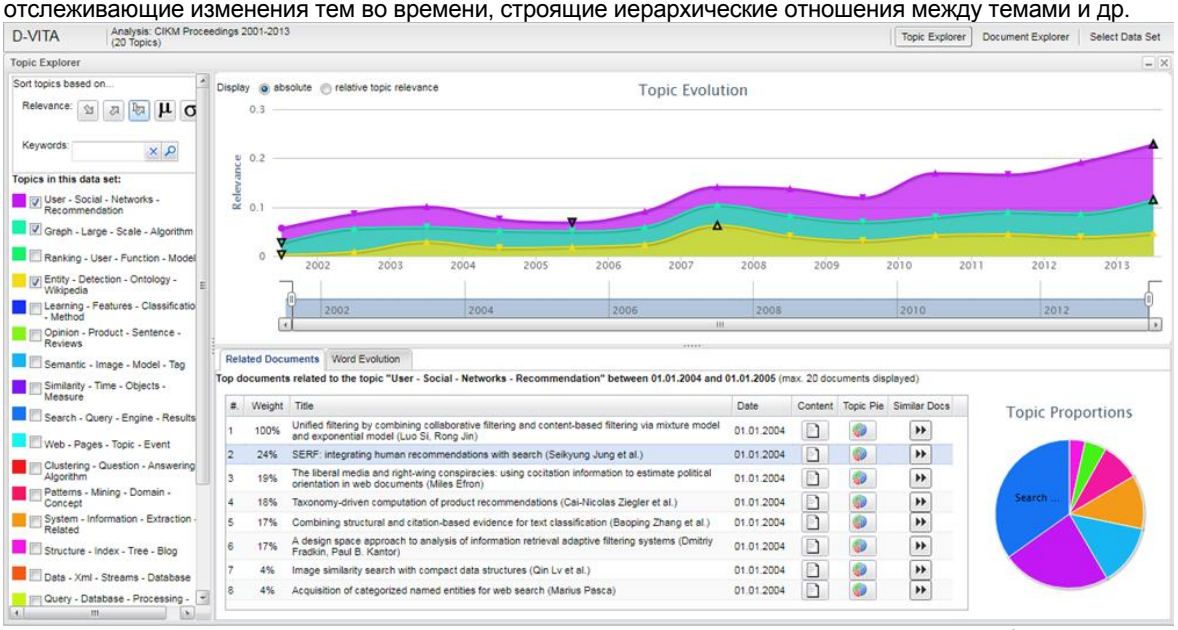

Рис. 1. Пример визуализации динамической тематической модели с помощью D-VITA[3] (Günnemann, et.al., 2013).

[3] На рисунке приведена визуализация модели, построенной в демонстрационной версии D-VITA, на основе аннотаций статей в Proceedings of ACM Conference on Information and Knowledge Management, 2001-2013 гг. Демонстрационная версия доступна по ссылке http://monet.informatik.rwth-aachen.de/DVita

В левой части окна представлены выделенные темы, при выборе которых отображается динамика их развития; для каждой «тематической реки» (themeriver) можно отобразить список документов, относящихся к данной теме, а выбрав конкретный документ – увидеть частоту встречаемости выделенных в нем тем с помощью круговой диаграммы. На основе распределения тем в выбранном документе строится список похожих документов.

Доминирующим подходом к тематическому моделированию в настоящее время является байесовское обучение. Большинство моделей разрабатываются на основе модели латентного размещения Дирихле (Blei, et.al. 2003). Также активно развивается и альтернативный, многокритериальный подход, получивший название Аддитивная регуляризация тематических моделей (Воронцов, 2014), в котором модель оптимизируется по взвешенной сумме критериев.

Данная статья построена следующим образом. Во втором разделе мы вводим понятие векторного представлении документов, а также приводим ранние подходы к исследованию текстов. В третьем разделе дается описание базовых тематических моделей, а в четвертом – обзор тематических моделей второго поколения, которые развивались на основе базовых. Пятый раздел содержит некоторые актуальные вопросы тематического моделирования, а также информацию о его современных программных реализациях. В заключении выводится предположение о том, что аппарат тематического моделирования мог бы быть эффективно применен в разрезе изучения цифровой трансформации экономики.

2. Текст как вектор

Формально, методы тематического моделирования можно разделить на две группы – дискриминативные и вероятностные, причем дискриминативные модели являются наиболее примитивным подходом, так как подразумевают, что темы в документах распределены равномерно. Однако мы начнем описание тематических моделей с дискриминативного подхода как основы вероятностных моделей.

Итак, большинство методов анализа текстов используют векторную модель представления информации (Vector Space Model, VSM) (Salton, 1975), представляющую каждый документ как вектор размерности N, где N – число выделенных термов во всей коллекции документов. I-й компонент вектора содержит вес i-го терма для данного вектора.

Наиболее распространенным методом назначения веса терму является вычисление метрики TF-IDF – статистической меры частоты встречаемости терма в конкретном документе (TF, term frequency – частота терма), определяемой в сравнении с частотой его использования в других документах коллекции (IDF, inverse document frequency – обратная частота документа).

$$TF(w_i, d_j) = \frac{fr_{ij}}{\sum_i fr_{ij}}, \qquad IDF(w_i, D) = log\frac{|D|}{|\{d_j \in D: w_i \in d_j\}|} \quad , \qquad (1)$$

Где fr_{ij} – частота встречаемости терма w_i в документе d_j; |D| – число документов коллекции; $|\{d_j \in D: w_i \in d_j\}|$ – число документов в коллекции, в которых встретился терм w_i.

В зависимости от решаемой задачи используются различные варианты метрики TF-IDF. В классическом случае это $TF \times IDF$, однако, используются и более сложные комбинации данных показателей. В любом случае, учёт IDF уменьшает вес широкоупотребительных слов, и общее значение TF-IDF для терма будет тем больше, чем выше частота встречаемости терма в конкретном документе и ниже в других документах коллекции.

Таким образом, располагая представлением всех документов в виде N-мерных векторов весов слов, мы можем находить расстояние между точками пространства и тем самым решать задачу подобия документов. Для сравнения векторов документов существует множество метрик, например, косинусное расстояние, и другие методы (в публикации Choi, et. al. (2010) отмечено более 70 способов вычисления мер схожести векторов).

Несмотря на то, что построение простейшей векторной модели для задач сравнения текстов между собой актуально и сейчас, модель неприменима для текстов больших размеров, к тому же она не позволяет разрешить проблему синонимии и полисемии слов. Развитием векторной модели стало представление наборов векторов термов из документа как общей терм-документной матрицы. Первым из методов, реализующих терм-документное представление коллекции документов, стал метод латентно-семантического индексирования.

Латентное-семантическое индексирование (Latent Semantic Indexing, LSI или, что то же – латентно-семантический анализ, Latent Semantic Analysis, LSA) – метод, предложенный Deerwester, et al. (1990), с целью повышения эффективности работы информационно-поисковых систем. В LSA задача состоит в том, чтобы спроецировать часто встречающиеся вместе термы в одно и то же измерение семантического пространства, которое имеет пониженную размерность по сравнению с оригинальной терм-документной матрицей. Элементы матрицы содержат веса термов в документах, назначенные с помощью выбранной весовой функции.

В основе LSA лежит метод сингулярного разложения терм-документной матрицы X: $X = T_0 S_0 D_0^T$, где X – прямоугольная матрица размерности $t \times d$, T_0, D_0^T – ортогональные матрицы размерности $t \times r$ и $r \times d$ соответственно, r – ранг матрицы X, S – диагональная матрица.

Как результат – мы имеем матрицу $\hat{X}$ пониженной размерности k (k – число наибольших сингулярных значений матрицы S) являющуюся наилучшим приближением матрицы X: $\hat{X} = TSD^T$, где матрицы T, D имеют размерности $t \times k$ и $k \times d$ соответственно.

Таким образом, каждый терм и документ представляются при помощи векторов в общем семантическом пространстве размерности k, в котором и определяется их близость.

К основным недостаткам LSA относят предположение модели о нормальном распределении термов в документах, а также сложность интерпретации результатов (Rosario, 2000). В работе Hofmann (1999) был предложен статистический взгляд на LSA, что положило начало развитию вероятностного тематического моделирования.

3. Первое поколение вероятностных тематических моделей

3.1. Вероятностный латентно-семантический анализ

Вероятностный латентно-семантический анализ (Probabilistic Latent Semantic Analysis, PLSA) был предложен Hofmann (1999) и базировался на принципе максимума правдоподобия как альтернатива классическим методам кластеризации текстов, основанным на вычислении функций расстояния.

Пусть есть коллекция документов $\mathcal{D} = \{d_1, \dots, d_N\}$, а также термы, составляющие словарь $\mathcal{W} = \{w_1, \dots, w_M\}$. Игнорируя последовательность, с которой термы встречаются в документе, данные обобщаются в терм-документной матрице $N \times M$, каждый элемент которой соответствует частоте встречаемости терма в документе.

PLSA связывает с каждой наблюдаемой переменной (термом и документом) латентную (скрытую) тему $t \in T = \{t_1, \dots, t_k\}$. Совместная вероятностная модель над $\mathcal{D} \times \mathcal{W}$ определяется вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w|t)$ и тем в документах $\theta_{td} = p(t|d)$:

$$p(d,w) = p(d)p(w|d), \quad p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td} \qquad (2)$$

В модели вводится предположение об условной независимости d и w – термы в документе определяются латентной темой (t), а не документом. Также предполагается, что число тем много меньше, чем документов и термов.

Итак, вероятностная модель (2) описывает процесс порождения коллекции по известным распределениям $p(w|t)$, $p(t|d)$. Задача тематического моделирования – это обратная задача: по заданной коллекции $\mathcal{D}$ требуется найти параметры распределения термов в темах и тем в документах, при которых тематическая модель (2) хорошо приближает частотные оценки условных вероятностей $\hat{p}(w|d) = n_{dw}/n_d$ (частота встречаемости терма в документе).

Для определения оптимальных значений скрытых параметров модели в PLSA используется стандартная процедура оценки максимального правдоподобия – EM-алгоритм, в котором каждая итерация состоит из двух шагов – E (expectation), на котором вычисляются апостериорные вероятности для скрытых параметров, и M(maximization), на котором параметры обновляются.

На E-шаге алгоритма оценивается вероятность:

$$p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} \qquad (3)$$

Формулы для параметров на M-шаге:

$$\varphi_{wt} = \frac{n_{wt}}{\sum_w n_{wt}}; \ \theta_{td} = \frac{n_{td}}{\sum_t n_{td}} \qquad (4)$$

Модель PLSA, описанная Hofmann (1999), является важной вехой в развитии вероятностного моделирования текстов, однако она имеет существенные ограничения. Так, в PLSA каждый документ представляется числовым вектором, каждая компонента которого равна доле соответствующей темы в документе. Однако вероятностная модель не описывает закон распределения этих долей, а также вероятности самих документов. В результате число параметров модели линейно растёт с ростом размера текстовой коллекции, что может приводить к переобучению. Кроме того, непонятно, как оценивать вероятности новых документов, не входивших в состав обучающей выборки. Другими словами, модель задаёт закон порождения слов, но не закон порождения документов (Daud, et.al, 2010). Данные недостатки были устранены в модели скрытого размещения Дирихле.

3.2. Модель скрытого размещения Дирихле

Модель скрытого размещения Дирихле (Latent Dirichlet Allocation, LDA) предложена в работе Blei et. al. (2003). LDA – порождающая вероятностная модель, в которой документы представляются как вероятностная смесь скрытых тем (каждое слово в документе порождено некоторой латентной темой), при этом в явном виде моделируется распределение слов в каждой теме, а также априорное распределение тем в документе. Темы всех слов в документе предполагаются независимыми.

На первом шаге для каждого документа d выбирается случайный вектор распределения тем θ_d из распределения Дирихле с параметром α. На втором шаге выбирается тема t_{di} (в классической модели

LDA количество тем фиксировано изначально) из мультиномиального распределения с параметром θ_d. Наконец согласно выбранной теме t_{di} выбирается слово w_{di} из распределения φ_t, которое является распределением Дирихле с параметром β.

Таким образом, порождающая модель слова w из документа d представляется в виде:

$$p(w|d,\theta,\varphi) = \sum_t p(w|t,\varphi_t)\, p(t|d,\theta_d), \quad (5)$$

$\theta \sim Dir(\alpha)$, $\varphi \sim Dir(\beta)$, где α и β — задаваемые так называемые гиперпараметры распределения Дирихле.

Как правило, все компоненты параметров α и β распределения Дирихле берутся равными, поскольку отсутствует априорная информация о распределении слов в темах и тем в документах (Коршунов, Гомзин, 2012). Предложены подходы, позволяющие восстановить оптимальные значения гиперпараметров модели по обучающей выборке (например, Heinrich, 2005). Так, например, в Griffiths and Steyvers (2004) α принимается равным 50/T, β =0.1.

Для оптимизации параметров φ, θ чаще всего используется сэмплирование Гиббса (Collapsed Gibbs Sampling), но также используются и другие подходы, такие как Максимизация апостериорной вероятности (Maximum a posteriori probability), Вариационный байесовский вывод (Variational Bayes). Подробнее о методах оптимизации параметров LDA см. Heinrich (2005).

Большинство сравнительных экспериментов демонстрировали превосходство качества модели LDA над PLSA (Blei et al., 2003, Boyd-Graber et al., 2009). Однако более поздние эксперименты показали, что переобучение модели PLSA (одна из основных «претензий» к данной модели) не наблюдается на больших текстовых коллекциях, правдоподобие моделей PLSA, LSA отличаются незначительно (Воронцов, Потапенко, 2012). Различия проявляются только на низкочастотных термах, которые не важны для образования тем. Второй «недостаток» PLSA, относящийся к неадекватному описанию новых текстовых документов, может быть устранен путем реорганизации итерационного процесса обучения модели (Воронцов, 2014).

3.3. Критерии качества тематических моделей

Отметим отдельно наиболее распространенные критерии качества тематических моделей. Одним из основных является перплексия (perplexity) – мера, используемая для оценивания языковых моделей в компьютерной лингвистике (Azzopardi et.al., 2003). Перплексия является мерой несоответствия или «удивленности» модели $p(w|d)$ терминам w, наблюдаемым в документах коллекции $\mathcal{D}$.

$$perplexity(\mathcal{D};p) = exp\left(\frac{\sum_d \log p(w_d)}{\sum_d N_d}\right), \quad (6)$$

где w_d – вектор слов документа d, N_d – число слов в документе d.

Чем меньше перплексия, тем лучше модель предсказывает появление термов w в документе d. Важно, что с помощью перплексии некорректно сравнивать тематические модели, построенные на разных словарях.

Качество модели также зависит от того, насколько выявленные темы являются интерпретируемыми. Общепринятой численной оценкой интерпретируемости, не требующей привлечения экспертов, является когерентность (Newman, et.al., 2010; Mimno, et.al., 2011). Согласно Mimno, et.al. (2011) когерентность темы t определяется как:

$$coherence(t,V^t) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D\left(v_m^{(t)},v_l^{(t)}\right)+1}{D(v_l^{(t)})}, \quad (7)$$

Где $V^t = (v_1^{(t)},\dots,v_M^{(t)})$ – список M наиболее вероятных слов темы t; $D(v,v')$ – число документов, содержащих термы v и v'; $D(v)$ – число документов, содержащих только терм v.

Существуют и другие внутренние критерии качества тематических моделей, такие как межтематическое расстояние Кульбака-Лейблера, энтропия (Daud, et.al., 2010) и др., а также различные внешние критерии качества (например, точность и полнота информационного поиска; сопоставление найденных тем с заранее известными концептами и др.).

4. Второе поколение вероятностных тематических моделей

Подчеркнем, что стандартные модели PLSA и LDA подразумевают следующие допущения:

1. гипотеза «мешка слов» (bag of words model) – предположение о том, что для выявления тематики текстов важна только частота слов в документах, но не их порядок;
2. порядок следования документов в коллекции может быть любым;
3. количество тем определяется заранее и не меняется.

Очевидно, что на практике данные допущения в значительной степени не соответствуют реальности. Гипотеза мешка слов не позволяет учитывать связь слов в контексте; терминология, характерная для темы, может меняться в случае рассмотрения большого временного промежутка; а задача априорного определения числа тем вообще является нетривиальной (подробнее о проблеме определения оптимального числа тем см. Краснов, 2019).

Перечисленные ограничения служили мотивом для появления второго поколения вероятностных моделей, которые, расширяя базовые алгоритмы, позволили расширить границы применения тематического моделирования.

Основные расширения моделей представлены в таблице 1 в конце раздела. Отдельно стоит отметить Робастную вероятностную тематическую модель (Special Words with Background, SWB), предложенную Chemudugunta et al. (2006). В работе выдвигается предположение, что появление отдельных термов может объясняться не только тематикой документа, но и наличием общеупотребительных слов – фона и специфичных для конкретного документа редких термов, которые характерны для документа, но слабо представлены во всей коллекции – шума. SWB расширяет вероятностную модель добавлением шумовой и фоновой компонент.

Анализ литературы показывает, что LDA лидирует среди вероятностных тематических моделей благодаря многочисленным обобщениям, расширениям и приложениям к анализу коллекций текстовых документов (см. табл.1). Однако в работах Воронцов и Потапенко (2012), Potapenko and Vorontsov (2013), в которых критически пересмотрен взгляд на PLSA и LDA, отмечено, что широкое распространение LDA объясняется скорее его чисто математическим удобством для байесовского обучения, подчеркивается, что априорные распределения Дирихле и их обобщения не имеют убедительных лингвистических обоснований. Более того, переход от порождающей модели к алгоритму настройки её параметров требует весьма громоздких выкладок, которые резко усложняются при введении более сложных априорных распределений или совместном моделировании нескольких языковых явлений.

По этим причинам мощный импульс получило развитие так называемого подхода Аддитивной регуляризации тематических моделей (Additive Regularization of Topic Models, ARTM), разработанного Воронцовым (2014). ARTM– многокритериальный подход, в основе которого лежит представление задачи тематического моделирования как некорректно поставленной оптимизационной задачи, требующей введения регуляризатора – дополнительного критерия, учитывающего специфические особенности прикладной задачи или знания предметной области (Воронцов, 2014; Vorontsov and Potapenko, 2014).

Вероятностная модель порождения коллекции $\mathcal{D}$ понимается как задача приближенного представления известной терм-документной матрицы в виде произведения двух неизвестных матриц меньшего размера – матрицы термов в темах Φ и матрицы тем в документах Θ:

$$F \approx \Phi\Theta \quad (8)$$

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

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \qquad (9)$$

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

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

Искомое стохастическое матричное разложение $\Phi\Theta$ определено не единственным образом, а с точностью до невырожденного преобразования $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, то есть задача является некорректно поставленной. Согласно теории регуляризации (Тихонов и Арсенин, 1986), решение такой задачи возможно доопределить и сделать устойчивым. Для этого к основному критерию добавляется регуляризатор. Таким образом, наряду с правдоподобием (9) требуется максимизировать *n* критериев $R_i(\Phi,\Theta)$ – регуляризаторов.

Для решения задачи многокритериальной оптимизации максимизируется линейная комбинация критериев L и R_i с неотрицательными коэффициентами регуляризации τ_i , при условии неотрицательности и нормировки столбцов матриц Φ, Θ:

$$L(\Phi,\Theta) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \qquad (10)$$

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

Решение этой задачи строится на основе так называемого регуляризованного EM-алгоритма (см. Воронцов и Потапенко, 2014).

Таким образом, в настоящее время наметились два направления развития тематических моделей –на основе Байесовского обучения (модель LDA) и на основе Аддитивной регуляризации. В работе Vorontsov and Potapenko (2014) пересматриваются тематические модели, ранее разработанные в рамках байесовского подхода, для каждой из которых находится соответствующий регуляризатор, который приводит к тому же самому или очень похожему алгоритму обучения модели. По сравнению с байесовским подходом, ARTM радикально упрощает вывод алгоритма и позволяет комбинировать регуляризаторы в

произвольных сочетаниях. Также в недавних исследованиях показано превосходство ARTM над LDA по качеству выделенных тем (см., например, работу Apishev et.al., 2017, где ARTM и LDA сравниваются на примере мониторинга этнически обусловленного дискурса в социальных сетях).

Таблица 1.

Описание	Модель	Класс моделей
Расширения общего характера		
Добавление в тематическую модель фоновой (соответствует общеупотребительным словам) и шумовой (соответствует редким, специфичным словам) компонент	Робастная тематическая модель (Special Words with Background, SWB) – Chemudugunta et.al. (2006).	LDA
Непараметрические модели. Отказ от необходимости точного определения числа тем, тематическое моделирование с потенциально бесконечным числом тем	Иерархический процесс Дирихле (Hierarchical Dirichlet Process, HDP) – Teh, et.al.(2006)	LDA
Онлайн-модели. Алгоритмы, работающие не с фиксированным набором данных, а с данными, обновляющимися в режиме реального времени	Вариационное онлайн оценивание параметров LDA – Hoffman, et.al.(2010); онлайн-модификации сэмплирования по Гиббсу – Canini, et.al.(2009)	LDA
Мультиязычные тематические модели	Polylingual Topic Model – Mimmo, et.al.(2009)	LDA
Тематические модели с учителем	Supervised LDA (sLDA) – Blei and McAuliffe (2007)	LDA
Модели, учитывающие внешние отношения		
Учет взаимосвязей посредством авторства	Автор-тематическая модель (Author-Topic Model, ATM) – Rosen-Zvi, et.al. (2004).	LDA
Учет цитирования документов	Совместная вероятностная модель (Joint Probabilistic Model, JPM) – Cohn and Hofmann (2001).	PLSA
Учет как внешних, так и внутренних ссылок документа	Скрытая тематическая модель гипертекста (Latent Topic Hypertext Model, LTHM) -Gruber, et.al.(2008).	LDA
Учет взаимосвязей терминов и авторов по корпусу текстов докладов на научных конференциях	Тематическая модель автор-конференция (Author-Conference Topic Model, ACT) – Tang, et.al. (2008). Обобщение ACT: Решения задачи поиска экспертов на основе информации о семантике и времени (Semantic and Temporal Information Based Maven Search, STMS) – Daud, et.al. (2009)	LDA
Учет связей между участниками социальных сетей	Модели автор-получатель (Author-Recipient Topic Model) - McCallum, et.al. (2004)	LDA
Учет произвольных сетевых структур документов	NetPLSA – Mei et.al.(2008)	PLSA
Учет пользовательских меток документов (тегов), в том числе множественных меток	Labeled LDA – Ramage, et.al.(2009); Flat-LDA – Rubin, et.al.(2012)	LDA
Модели, учитывающие внутренние отношения		
Модели, учитывающие взаимосвязи между темами	Корреляционные тематические модели (Correlated Topic Models, CTM) – Blei and Lafferty (2006). Модель на основе ориентированного ациклического графа (Pachinko Allocation Model, PAM) – Li and McCallum (2006).	LDA
Моделирование иерархии тем – от более общих до узких	Модель иерархического скрытого размещения Дирихле (Hierarchical Latent Dirichlet Allocation, hLDA) – Blei et.al.(2003)	LDA
Модели, учитывающие зависимости между словами документа	Скрытая марковская модель и LDA (HMM-LDA) – Griffiths et.al.(2005). Биграммная тематическая модель (Bigram Topic Model) – Wallach (2006).	LDA
	Модель контекстных смесей (Contextual Mixture, CPLSA) – Mei and Zhai (2006).	PLSA
Модель на основе предположения, что последовательность тем в документе является марковской цепью	Скрытая тематическая марковская модель (hidden topic Markov model, HTMM) – Gruber et.al.(2007). N-граммная тематическая модель (Topical N-gram model, TNG) – Wang et.al.(2007).	Марковская модель
Динамические тематические модели		
Отображение динамики изменения тем – дискретное время	Dynamic Topic Model, DTM – Blei and Lafferty (2006)	LDA
Отображение динамики изменения тем – несколько различных масштабов времени	Многомасштабная томографическая модель (Multiscale-topic Tomography Model, MTTM) -Nallapati et.al. (2007)	LDA
Отображение динамики изменения тем – непрерывное время	Continuous Time Dynamic Topic Model, cDTM – Wang et.al.(2008).	LDA
Модель тематики во времени, где тема порождает и термы, и отметку времени	Topics Over Time, TOT – Wang and McCallum (2006); Continuous Time Model, CTM – Wang et.al.(2006)	LDA

5. Некоторые актуальные вопросы тематического моделирования

Несмотря на то, что тематические модели эффективно используются уже более 10 лет, ряд вопросов до сих пор остаётся открытым.

Так, существует проблема определения оптимального количества тем, связанная с фактом того, что в реальных текстовых коллекциях истинного числа тем просто не существует. Поэтому, несмотря на ряд предложенных подходов (см., например, Краснов, 2019), выбор числа тем является своего рода эвристикой, зависящей как от объема и структуры коллекции, так и от субъективного взгляда исследователя. Кроме того, в случае построения онлайн-модели, подразумевающей добавление новых данных, возникает вопрос об обнаружении новых тем и добавлении их в модель.

После определения «оптимального» числа тем возникает вопрос их интерпретируемости, который также является дискуссионным – предполагается, что каждая тема характеризуется небольшим числом термов, каждый документ относится к небольшому числу тем. В хорошей модели темы являются хорошо интерпретируемыми – считается, что эксперт может понять, о чем данная тема, посмотрев на список наиболее вероятных слов. Однако на практике вопрос интерпретации тем человеком является открытым (более подробно о данном вопросе см., например, Chang, et.al.,2009; Mavrin, et.al., 2018; Alekseev, et.al. 2018). Для облегчения интерпретируемости тем можно воспользоваться, например, одним из способов визуализации матрицы Ф (см. рис. 2) – системой Termite (Chuang et.al., 2013).

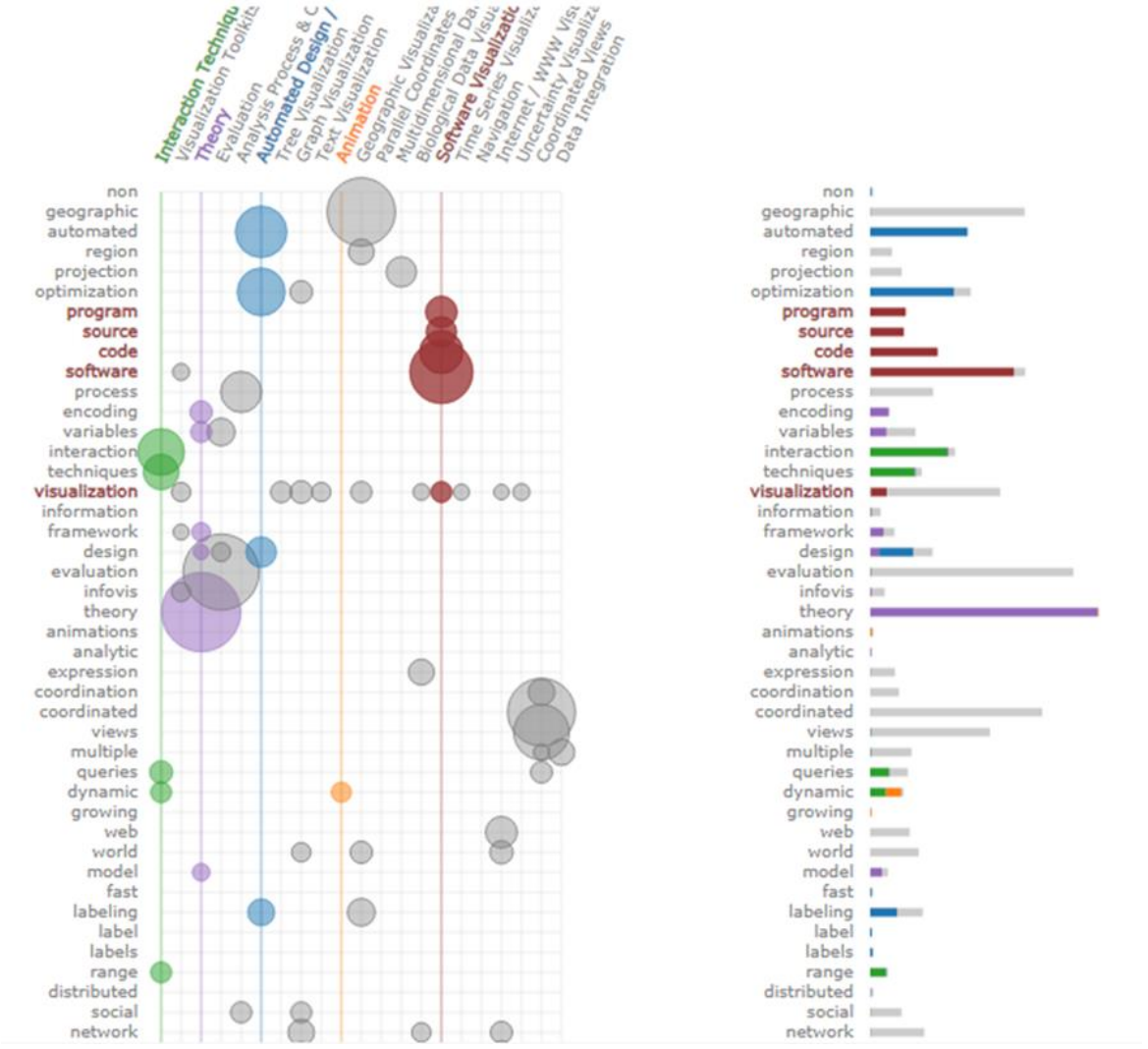

Рис. 2. Интерактивная система Termite для визуализации матрицы Ф.
Выделенный столбец позволяет наглядно увидеть, какие термы наиболее характерны для данной темы, и дать ей интерпретируемое название. Termite позволяет также увидеть пересечение тем и, наоборот, по выделенному терму отобразить темы, для которых он вероятен.

В настоящее время существует большое число программных реализаций методов тематического моделирования. Наиболее распространенными и активно поддерживаемыми реализациями являются библиотеки Python: Gensim, Scikit-learn, BigARTM (Vorontsov et.al., 2015). Также стоит отметить библиотеки на других языках: Vowpal Wabbit (C++), Mallet (Java) (McCallum, 2002), Matlab Topic Modeling Toolbox.

Важным вопросом является визуализация результатов тематического моделирования. Разработано большое число средств визуализации тематических моделей: Termite System (Chuang et.al. 2013),TIARA (Wei et.al.2010), HierarchicalTopics (Dou et.al.2013) и др. Для подробного обзора средств визуализации мы адресуем читателя к Айсина (2015), где систематизированы основные актуальные инструменты. Один из примеров визуализации тематической модели приведен на Рисунке 3.

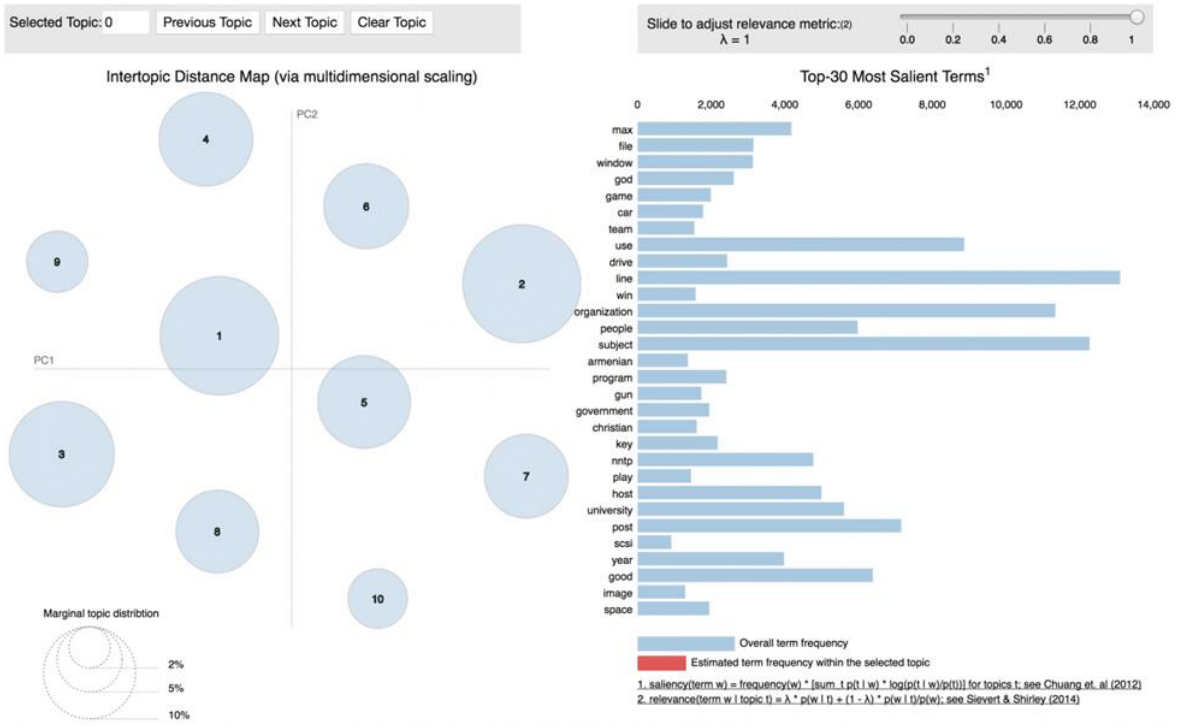

Рис. 3. Визуализация тематической модели с помощью библиотеки pyLDAvis в Python, отображающая распределение термов в темах.

Интерактивный график позволяет для каждой выбранной темы (круги слева) отображать распределение термов в ней (столбцы справа); и, наоборот, для каждого выбранного терма отображать темы, в которых он наиболее вероятен[4].

Отдельно стоит отметить системы, реализующие на основе тематической модели навигацию по коллекции. Каноническим примером такого тематического навигатора может служить система Topic Model Visualization Engine (Chaney and Blei, 2012). На Рисунке 4 приведен представленный авторами для демонстрации пример на основе англоязычной Wikipedia. Также так называемый тематический браузер – интерактивный инструмент для просмотра тематических моделей – реализован в системе Topic Browser (Gardner, et.al., 2010), в работе Carlson (2016) и др.

[4] Для построения графика использовались данные о 11000 новостных сообщениях. Источник данных: https://raw.githubusercontent.com/selva86/datasets/master/newsgroups.json

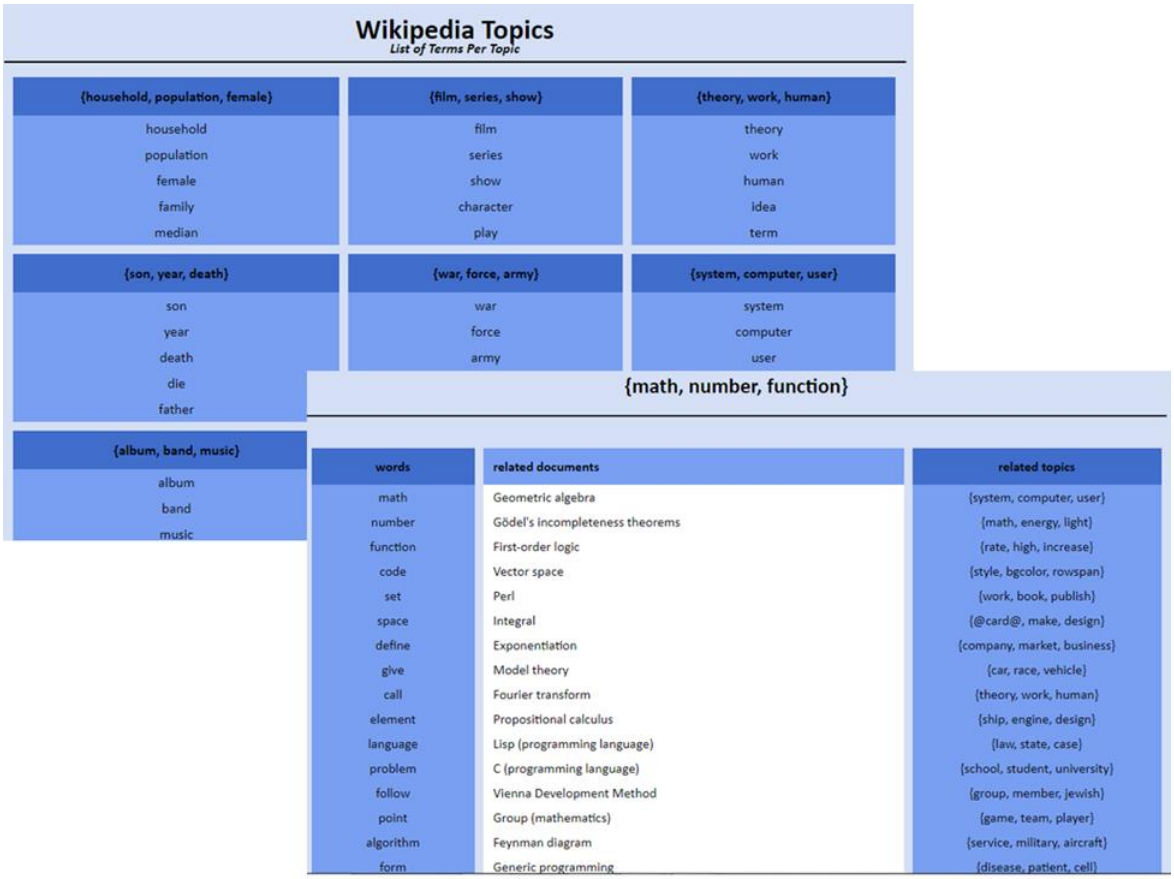

Рис. 4. Topic Model Visualization Engine[5] – пример навигации по коллекции Wikipedia на основе тематической модели.

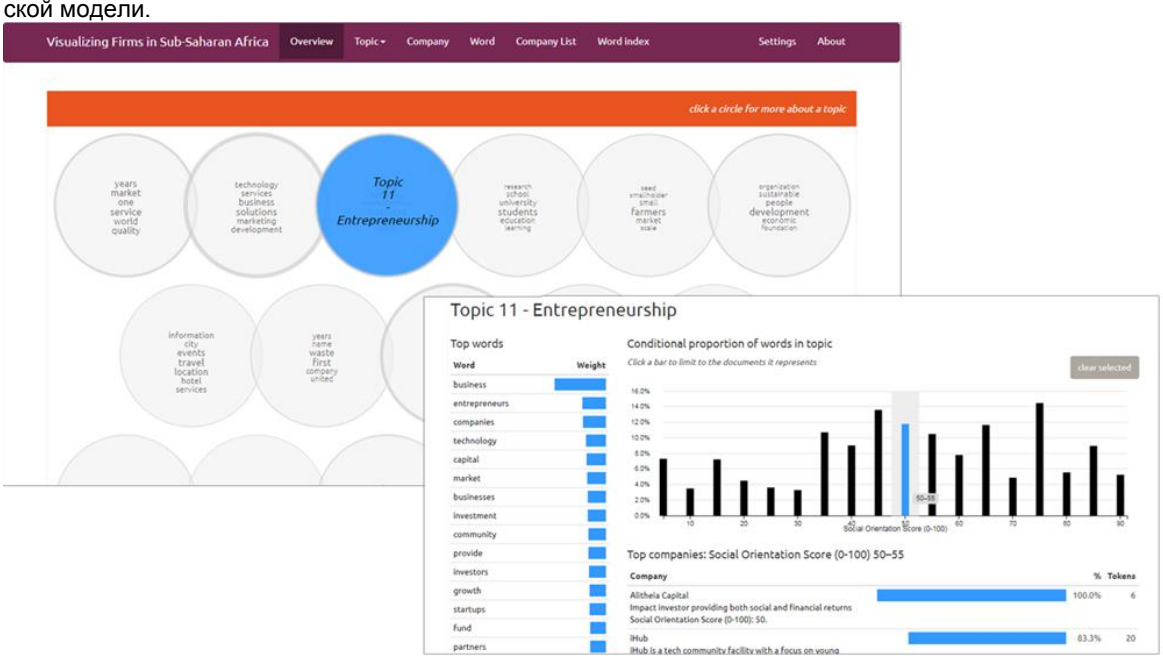

Рис.5. Тематический навигатор по компаниям-стартапам[6] . Визуализация модели к работе Carlson (2016)

[5]Демонстрационная версия Topic Model Visualization Engine доступна по ссылке http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html

[6] Реализация тематического навигатора представлена по ссылке http://www.natalieannecarlson.com/SSA_browser_demo/

6. Вместо заключения: «Дальнее чтение» цифровой экономики

Несмотря на большое число обзоров, посвященных тематическому моделированию (Daud, et.al., 2010; Коршунов и Гомзин, 2012; Boyd-Graber et.al.,2017 и др.), нам показалось актуальным рассмотреть это направление с целью дальнейшего использования в разрезе исследования цифровой трансформации экономики. Термин «цифровая экономика» в последнее время является одним из самых упоминаемых как в прессе, интернет-площадках, так и на многочисленных экономических форумах (Устюжанина и др., 2017), причем в данном случае речь идет не об очередном модном лозунге, а об объективно обусловленном процессе. Однако до сих пор у большинства исследователей нет ясного понимания того, что такое цифровая экономика как общественная система (Устюжанина и др., 2017). Даже на уровне экспресс-анализа терминологии цифровой экономики ясно, что она (терминология) носит характер несложившейся, тем самым иллюстрируя нерешенность проблем государственного и законодательного плана (Милкова, 2018).

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, нам необходимо именно «дальнее чтение» цифровой экономики, которое и стало возможным благодаря развитию цифровых технологий.

Литература

26. Alekseev, V. A., Bulatov, V. G., Vorontsov, K. V. (2018). Intra-text coherence as a measure of topic models ‘ interpretability. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”
27. Apishev,M., Koltsov S., Koltsova, O., Nikolenko, S., and Vorontsov, K. (2017). Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. Conference Paper in Lecture Notes in Computer Science.
28. Azzopardi, L., Girolami, M., Risjbergen, K. V. (2003). Investigating the relationship between language model perplexity and IR precision-recall measures. In: Proceedings of the 26th ACM SIGIR, Toronto, Canada
29. Blei, D. M., Moreno, P. J. (2001). Topic segmentation with an aspect hidden Markov model. In: Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans. LA USA, 343–348
30. Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3.
31. Blei, B., Griffiths, T., Jordan, M., and Tenenbaum, J. (2003). Hierarchical topic models and the nested Chinese restaurant process. Neural Information Processing Systems, 16
32. Blei, D. M. and Lafferty, J. (2006). Correlated topic models. In: Advances in Neural Information Processing Systems (NIPS), Cambridge, MA, MIT Press, 147–154
33. Blei, D. M. and Lafferty, J. (2006). Dynamic topic models. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA.
34. Blei, D. M., and McAuliffe, J. (2007). Supervised topic models. In: Advances in Neural Information Processing Systems (NIPS), Cambridge, MA, MIT Press
35. Boyd-Graber, J., Hu, Y., and Mimmo, D. (2017). Applications of Topic Models. Foundations and Trends in Information Retrieval, 1–154
36. Canini, K.R., Shi, L., and Griffiths, T.L. (2009). Online Inference of Topics with Latent Dirichlet Allocation. Journal of Machine Learning Research – Proceedings Track 5, 65-72
37. Carlson, N. (2016). Social Entrepreneurship, Language, and Funding: Evidence from Tech Startups in Sub-Saharan Africa. Columbia Business School.
38. Chaney, A., Blei, D. (2012). Visualizing topic models. Frontiers of Computer Science in China, 55(4), 77-84.
39. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D.M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In: Advances in Neural Information Processing Systems, 288–296
40. Chemudugunta C., Smyth P., and Steyvers M. (2006). Modeling general and specific aspects of documents with a probabilistic topic model . In: Advances in Neural Information Processing Systems. – MIT Press, Vol. 19, 241–248
41. Choi, S., Cha, S., and Tappert, C. (2010). A Survey of Binary Similarity and Distance Measures. Journal of Systemics, Cybernetics and Informatics, 8(1), 43-48
42. Chuang, J., Manning, C., and Heer J. (2013). Termite: Visualization techniques for assessing textual topic models. Working Conference (International) on Advanced Visual Interfaces Proceedings. -ACM, 74-77.
43. Cohn, D., Hofmann, T. (2001). The missing link- a probabilistic model of document content and hypertext connectivity. In: Advances in Neural Information Processing Systems (NIPS) , Cambridge, MA, MIT Press.

44. 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.
45. Daud, A., Li, J., Zhu, L., and Muhammad, F. (2009). A generalized topic modeling approach for maven search. In: Proceedings of International Asia- Pacific Web Conference and Web-Age Information Management (APWEB-WAIM), Suzhou, China.
46. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391—407.
47. Dou, W., Yu, L., Wang, X., Ma, Z., and Ribarsky, W. (2013). HierarchicalTopics: Visually exploring large text collections using topic hierarchies. IEEE Trans. Vis. Comput. Gr., 19 (12), 2002-2011.
48. Gardner, M. P., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E.K., and Seppi, K.D. (2010). The Topic Browser An Interactive Tool for Browsing Topic Models.
49. Griffiths, T.L, Steyvers, M. (2004) Finding scientific topics. In: Proceedings of the National Academy of Sciences. USA, 101: 5228–5235
50. Griffiths, T. L., Steyvers, M., Ble,i D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. In: Advances in Neural Information Processing Systems (NIPS) 17. Cambridge, MA, MIT Press, 537–544
51. Gruber, A., Rosen-Zvi, M., and Weiss, Y. (2007). Hidden topic Markov models. In: Proceedings of Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico, USA.
52. Gruber, A., Rosen-Zvi, M., and Weiss, Y. (2008). Latent topic models for hypertext. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland.
53. Günnemann, N., Derntl, M., Klamma, R., and Jarke, M. (2013). An Interactive System for Visual Analytics of Dynamic Topic Models. Datenbank-Spektrum, 13(3), 213-223, Springer Verlag
54. Heinrich, G. (2005). Parameter estimation for text analysis. Technical Note, University of Leipzig
55. Heinrich, G. (2005). Parameter estimation for text analysis. Technical Note, University of Leipzig, Germany
56. Hoffman, M., Blei, D., and Bach, F. (2010). Online learning for latent Dirichlet allocation. Neural Information Processing Systems.
57. Hofmann,T. (1999) Probabilistic Latent Semantic Analysis. Uncertainty in Artificial Intelligence, UAI'99, Stockholm.
58. Meho, L. (2007). The Rise and Rise of citation analysis. Physics World, 20(1)
59. Li, W., McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, 577–584
60. Mavrin, A., Filchenkov, A., and Koltsov, S. (2018). Four Keys to Topic Interpretability in Topic Modeling. In: Ustalov D., Filchenkov A., Pivovarova L., Žižka J. (eds) Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol 930. Springer, Cham
61. McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002
62. McCallum, A., Corrada-Emmanuel, A., Wang, X. (2004). The author-recipient- topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical Report UM-CS-2004-096.
63. Mei, Q. and Zhai, C. X. (2006). A mixture model for contextual text mining. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 649–655
64. Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008). Topic modeling with network regularization. Proceedings of the 17th International Conference on World Wide Web -WWW'08, New York, NY, USA,101-110.
65. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A. (2009). Polylingual Topic Models. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 880–889
66. Mimno, D.,Wallach, H. M., Talley, E., Leenders, M., McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the Conference on Empirical Methods in Natural Language Processing.– EMNLP ’11.– Stroudsburg, PA, USA: Association for Computational Linguistics, 262-272.
67. Nallapati, R., Cohen, W., Ditmore, S., Lafferty J, Ung, K. (2007). Multiscale topic tomography. In: Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA
68. Newman, D., Lau, J. H., Grieser, K., Baldwin, T. (2010). Automatic evaluation of topic coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.– HLT’10.– Stroudsburg, PA, USA: Association for Computational Linguistics, 100-108.

69. 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.
70. Ramage, D. Hall D., Nallapati R., and Manning, C.D. (2009) Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256.
71. Rosario, B. (2000). Latent Semantic Indexing: An overview. Technical report INFOSYS 240 Spring Paper, University of California, Berkeley
72. Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. (2004). The author-topic model for authors and documents. Proceedings of the 20th Conference on Uncertainty in artificial intelligence. UAI '04 – Arlington, Virginia, United States: AUAI Press, 487-494.
73. Rubin T. N., Chambers A., Smyth P., Steyvers M. (2012). Statistical topic models for multi-label document classification. Machine Learning, 88, 1-2, 157–208.
74. Salton, G.M. , Wong, A., and Yang C.S. (1975). A vector space model for automatic indexing. Communications of the ACM 18(11): 613--620 https://dl.acm.org/citation.cfm?id=361220
75. Tang, J., Zhang, J., Yao, L., L.i, J., Zhang, L., and Su, Z.. (2008). ArnetMiner: extraction and mining of academic social networks. In: Proceedings of ACM SIGKDD
76. Teh Y. W., Jordan, M.I.,Beal, M.J., and Blei, D.M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101 (476)
77. Vorontsov K. V., Potapenko A. A. (2014). Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization
78. Vorontsov, K., Frei,O., Apishev, M., Romov, P., Suvorova, M. (2015). 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), 370-384.
79. Wallach, J. M. (2006). Topic modeling: Beyond bag-of-words. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA.
80. Wang, C., Blei, M. D. and Heckerman, D. (2008). Continuous time dynamic topic models. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland
81. Wang, X., Li, W., and McCallum, A. (2006). A continuous-time model of topic co-occurrence trends. In: AAAI Workshop on Event Detection. Boston, Massachusetts, USA
82. Wang, X., McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA
83. Wang, X., McCallum, A., and Wei, X.(2007). Topical N-grams: phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), Omaha NE, USA
84. Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M., Qian, W., Shi, L., Tan, L., and Zhang, Q. (2010). TIARA: A visual exploratory text analytic system. 16th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining. – ACM, 153-162.
85. White R.W., Roth R.A. (2009). Exploratory search: Beyond the query-response paradigm. Morgan and Claypool Publs., 98 p.
86. Айсина, Р.М. (2015). Обзор средств визуализации тематических моделей коллекций текстовых документов. Машинное обучение и анализ данных, 1(11)
87. Воронцов К. В. (2014). Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, Т. 456 №3, 268-271
88. Воронцов, К. В., Потапенко, А. А. (2012). Регуляризация, робастность и разреженность вероятностных тематических моделей. Компьютерные исследования и моделирование, 4(4), 693–706.
89. Воронцов, К. (2016). Тематическое моделирование на пути к разведочному информационному поиску. Конференция Data Fest-3, Яндекс, Москва.
90. Корушнов, А., Гомзин, А. (2012). Тематическое моделирование текстов на естественном языке. Труды Института системного программирования РАН (электронный журнал), том 23
91. Краснов, Ф. (2019). Оценка оптимального количества тематик в тематической модели: подход на основе качества кластеров. International Journal of Open Information Technologies, 7(2).
92. Милкова М.А. (2018) Извлечение ключевых терминов направления «Цифровая экономика»: графоориентированный подход. Цифровая экономика, 4(4)
93. Моретти, Ф. (2016). Дальнее чтение. Перевод с английского А. Вдовин, О. Собчук, А. Шели, под научн. ред. И. Кушнаревой. Издательство института Гайдара, Москва.
94. Тихонов, А. Н., Арсенин, В. Я. (1986). Методы решения некорректных задач. М.: Наука.
95. Устюжанина, Е.В., Сигарев, А.В., Шеин, Р.А. (2017). Цифровая экономика как новая парадигма экономического развития. Экономический анализ: теория и практика, 16 (12), 2238 – 2253.
96. Шваб, К. (2016). Четвертая промышленная революция. М.: Эксмо, 208 с.

Милкова Мария Александровна m.a.milkova@gmail.com

Ключевые слова
тематические модели, дальнее чтение, модель скрытого размещения Дирихле, вероятностный латентно-семантический анализ, аддитивная регуляризация

Milkova Maria, Topic models as a tool for "long distance reading"

Keywords
Digital economy, Russian Digital Economy Program, graph-based approach, TextRank, semantic links, text mining

Abstract
The paper presents key terms extraction from the government documents issued in the period of 2013-2018 and linked to the Digital economy direction. One of the key interests of the analysis of government documents is to study them as primary source of digital economy terminology. The paper provides a brief review of the main approaches to key terms extraction and gives detailed description of one of the graph-based methods – a TextRank algorithm. The TextRank algorithm was tested on 13 government documents. The results of documents analysis are presented as weighted graphs of semantic links between keywords. Based on these words the lists of key terms are created for each document.

DOI: 10.34706/DE-2019-01-06