УДК 004.912
doi:10.15217/issn1684-8853.2016.6.43

# ТЕМАТИЧЕСКАЯ МОДЕЛЬ С БЕСКОНЕЧНЫМ СЛОВАРЕМ

***С. Н. Карпович***[a, 1], *руководитель направления поисковой оптимизации*
[a]*Rambler&Co интернет холдинг, Москва, РФ*

***Постановка проблемы:*** *в связи с постоянным ростом Интернета, увеличением количества новостей, сообщений в электронной почте, постов в блогах растет потребность в алгоритмах для автоматического анализа текстовых данных. Одним из перспективных направлений машинного обучения и анализа текстов на естественном языке являются алгоритмы тематического моделирования. Большинство методов тематического моделирования рассматривают данные в статичном виде, с конечным словарем, но на практике необходимы методы, позволяющие работать с пополняемым словарем. Каждый год появляются новые слова, какие-то слова выходят из обихода, поэтому вопрос пополнения словаря особенно актуален для онлайн тематических моделей.* ***Цель:*** *разработка подхода определения тематического вектора нового слова с использованием произведения Адамара тематических векторов документов, где это слово встретилось, который будет альтернативным подходу с использованием распределения Дирихле или процесса Дирихле.* ***Результаты:*** *исследования показали, что сумма векторов тем документов, где встретилось новое слово, дает неверное представление о тематической принадлежности нового слова. При этом для определения тематики нового слова по тематикам документов, где это слово встретилось, эффективнее использовать произведение Адамара. В результате перемножения векторов тем документов получаем тематический вектор нового слова с наибольшими значениями вероятностей у нескольких тематик, значение слабо выраженных тематик либо стремится к нулю, либо обнуляется.* ***Практическая значимость:*** *использование предложенного алгоритма позволяет бесконечно увеличивать словарь онлайн тематической модели, а следовательно, учитывать новые и старые слова.*

***Ключевые слова*** *— тематическое моделирование, обработка текста на естественном языке, машинное обучение.*

## Введение

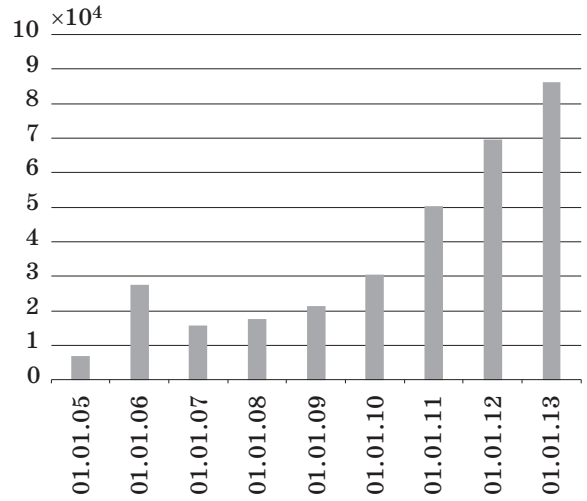
Тематическое моделирование — одно из современных направлений машинного обучения при анализе текстов. Вероятностная тематическая модель коллекции текстовых документов определяет, к каким темам и с какой вероятностью относится каждый документ, а также вероятности слов, составляющих каждую тему. Задача тематического моделирования рассматривается как задача одновременной кластеризации слов и документов по одному и тому же множеству кластеров, называемых темами. На выходе тематической модели (ТМ) каждому документу и слову определяется тематический вектор, состоящий из оценок степени принадлежности данного документа или слова каждой из тем. Размерность вектора равна числу тем. Алгоритмы тематического моделирования используются в исследованиях, применяются в задачах кластеризации, используются в информационном поиске.

В работах [1, 2] представлены базовые алгоритмы тематического моделирования — PLSI (Probabilistic Latent Semantic Indexing) и LDA (Latent Dirichlet Allocation), которые описывают построение ТМ с конечным набором документов, конечным размером словаря модели. В реальных задачах анализа текстов на естественном языке все чаще возникает потребность в онлайн анализе динамической ТМ, определении тематической принадлежности новых документов, определении тематики новых слов. Под «новым словом» в данной работе подразумевается слово, отсутствующее в словаре ТМ. Под «новым документом» подразумевается документ, который не участвовал в построении ТМ.

С каждым годом словарный состав, используемый в повседневной жизни любого человека, меняется. Изменяются частотные характеристики употребления слов и выражений, устаревающие выходят из обихода, появляются новые, прочно занимая свое место в нашей жизни. Действительно, исследуя данные корпуса текстов SCTM-ru [3], подсчитаем уникальные слова в документах корпуса за каждый год (рис. 1). Если снижение количества уникальных слов в 2007 г. могло быть связано с особенностью наполнения Викиновостей в это время, то, начиная с 2009 г., количество уникальных слов стремительно растет.

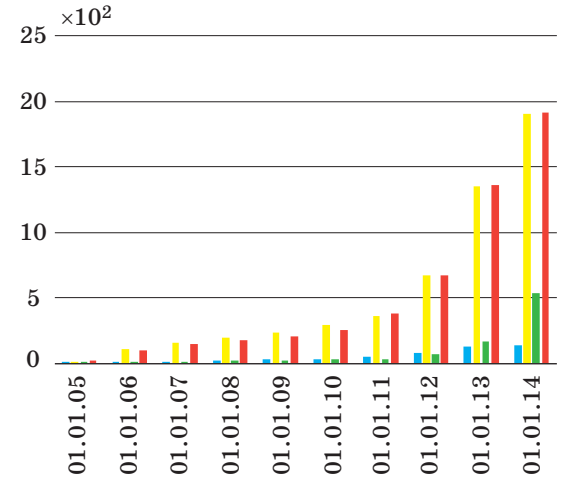
Тематическая модель, построенная на текстах до 2010 г., не будет содержать информацию о половине слов ТМ 2013 г., так как словарь новостей в 2013 г. в два раза больше словаря новостей 2010 г. Можно заметить, что рост количества уникальных слов произошел вследствие увеличения количества документов в Викиновостях. На рис. 2, *а* представлено количество документов в тематиках «Некрологи», «Россия», «Украина», «Политика». Хорошо заметен рост

[1] Научный руководитель — профессор, доктор технических наук, заведующий лабораторией интегрированных систем автоматизации СПИИРАН *А. В. Смирнов*.

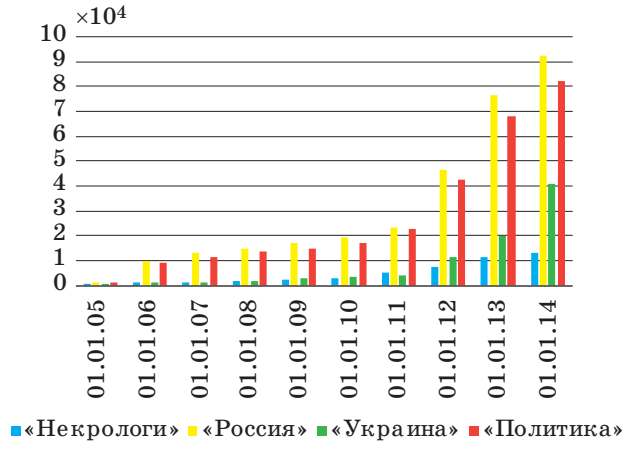



■ ***Рис. 1.*** Количество уникальных слов в корпусе SCTM-ru, 2005–2013 гг.

*а*)




*б*)




■ ***Рис. 2.*** Количество документов (*а*) и уникальных слов (*б*) по тематикам в корпусе SCTM-ru, 2005–2014 гг.

количества новостей по тематикам «Россия» и «Политика». На рис. 2, *б* представлено количество уникальных слов в этих же тематиках.

Из графиков видно, что количество новых документов в тематиках «Россия» и «Политика» от года к году росло равномерно, при этом словарный состав тематики «Россия» рос быстрее, и по количеству уникальных слов тематика «Россия» опережает тематику «Политика». Тематика «Украина», в четыре раза уступающая лидерам по количеству документов, по своему словарному составу лишь в два раза меньше этих же тематик лидеров.

Цель данной работы — выбрать подход к определению тематики нового слова в ТМ по тематическому вектору документов, где это слово встретилось. Разработать метод расчета тематического вектора для новых слов. Предложить алгоритм расширения словаря ТМ.

## Обзор существующих методов расширения словаря тематической модели

В работах [4–7] описаны подходы к построению динамических ТМ с фиксированным словарем. Эти ТМ позволяют проследить изменение тематики во времени, но не позволяют оценить изменение словарного состава модели.

В работе [8] описан алгоритм создания онлайн ТМ с бесконечным словарем. Основное отличие модели заключается в том, что мультиномиальное распределение берется из бесконечного процесса Дирихле по всем возможным словам, вместо конечного распределения Дирихле. В словарь добавляются новые слова с каждой итерацией добавления новых документов в ТМ, но количество слов в словаре ограничено заданным значением, поэтому словарь является усеченным упорядоченным множеством вероятностного распределения слов. Таким образом, словарный состав ТМ не меняется в размере, но изменяется по своему составу по мере добавления новых документов.

Алгоритм построения онлайн ТМ с изменяемым словарем рассмотрен в работе [9]. Время в модели дискретно, документы, поступающие в ТМ, разбиты на временные отрезки, временной срез может быть равен часу, дню, месяцу или году. Вводится понятие «окна», которому соответствует набор документов из нескольких временных срезов. В модели документы обрабатываются итеративно в рамках одного «окна». Одно из ключевых отличий этого подхода от онлайн LDA [7] является передача параметров от ранее построенной модели в новую. Второе отличие — это динамически изменяющийся словарь ТМ, где тематические векторы слов, которые при-

сутствовали в предыдущем «окне», рассчитываются по данным старой модели, а тематические векторы новых слов берутся из равномерного распределения Дирихле. Новые слова, встретившиеся более 10 раз, добавляются в модель, а старые слова, встретившиеся менее 10 раз в «окне», удаляются из ТМ. Отсутствует ограничение на размер словаря ТМ, поэтому он может изменяться по размеру и составу в ходе добавления новых документов.

Таким образом, существуют алгоритмы тематического моделирования с изменяемым словарем, в которых тематический вектор нового слова берется из равномерного распределения Дирихле либо из процесса Дирихле, но не предложено подходов в определении тематики нового слова по тематическим векторам документов, где это слово встретилось.

## Исследование темы нового слова

В качестве коллекции документов для исследования возьмем уже упоминавшийся корпус SCTM-ru, созданный специально для тестирования алгоритмов тематического моделирования. Источником данных корпуса является сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7000 документов. События, описанные в документах, происходят с ноября 2005 г. по июнь 2014 г. Словарный состав корпуса — 150 600 уникальных словоформ, 59 000 уникальных лемм. Для проведения исследования разделим корпус на две части, над первой частью построим ТМ, вторую будем использовать в качестве объекта исследования. Новости с 2005-го до конца 2012 г. возьмем в качестве данных для обучения. Применим алгоритм multi-label PLSI (ml-PLSI) [10] для обучения ТМ, который позволяет предсказывать тематическую принадлежность новых документов. Тестовый набор данных формируется из новостей за 2013 г.

По определению тематического моделирования, задача построения ТМ рассматривается как задача одновременной кластеризации слов и документов по одному и тому же множеству кластеров, называемых темами, поэтому размерность тематических векторов документов и слов одинакова. Алгоритм ml-PLSI определяет вектор тем нового документа по словарному составу этого документа: новые слова, которые впервые встретились в ТМ, никак не влияют на определение тематики нового документа. Резонно предположить, что тематика нового слова, впервые появившегося в ТМ, каким-то образом связана с тематикой документа, где это слово встретилось. Для того чтобы определить эту связь, возьмем набор новых документов и определим их тематическую принадлежность. Для определения тематики нового слова нужны тематические векторы документов, где эти слова встретились.

Слова, редко встречающиеся в коллекции документов, не значимы для ТМ. Чем больше документов содержат новое слово, тем точнее определяется тематическая принадлежность этого слова по векторам тем документов. Положим важными для ТМ те слова, которые встретились более чем в пяти документах. Таких новых слов оказалось 352. Например, слово «Кипру» встретилось в 14 документах, слово «олигарха» — в 11, «Царнаева» — в 10, «Тамерлан» — в девяти, «Анастасиадис» — в семи, «метеорита» — в шести, «вальц» — в пяти.

Для определения тематики нового слова «вальц» подсчитаем сумму тематических векторов документов, где это слово встретилось:

$$\mathbf{P}_{new}(w|t) = \sum_{i=1}^{n} \mathbf{P}_i(d|t),$$

где $\mathbf{P}_{new}(w|t)$ — тематический вектор нового слова; $\mathbf{P}_i(d|t)$ — тематический вектор документа, где встретилось новое слово; $n$ — количество документов, где встретилось новое слово.

Для визуального представления отнесения нового слова ко всем тематикам модели отобразим нормированный вектор на графике. Результат представлен на рис. 3, где по оси абсцисс — темы, по оси ординат — вероятность отнесения нового слова к теме. На рис. 4 показаны пять наиболее значимых тем, к которым относится слово «вальц».

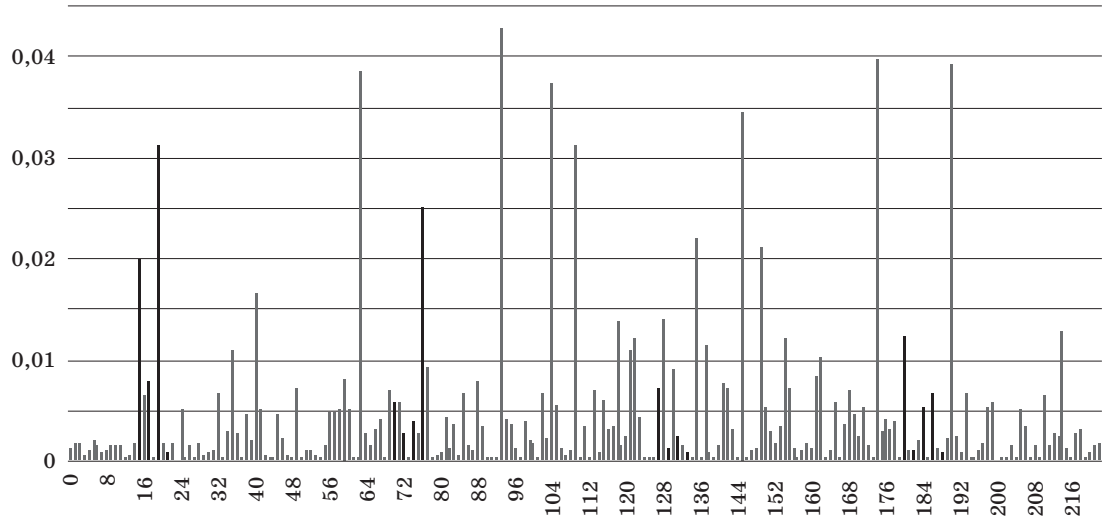
После нормирования слово «вальц» с вероятностью более 0,001 относится к 149 темам (см. рис. 3). Сумма тематических векторов может определять тематический вектор нового слова, но в этом случае новое слово относится к большому количеству тем, что зачастую не так.

Используем произведение тематических векторов документов, где встретилось новое слово, а именно покомпонентное произведение Адамара [11]:

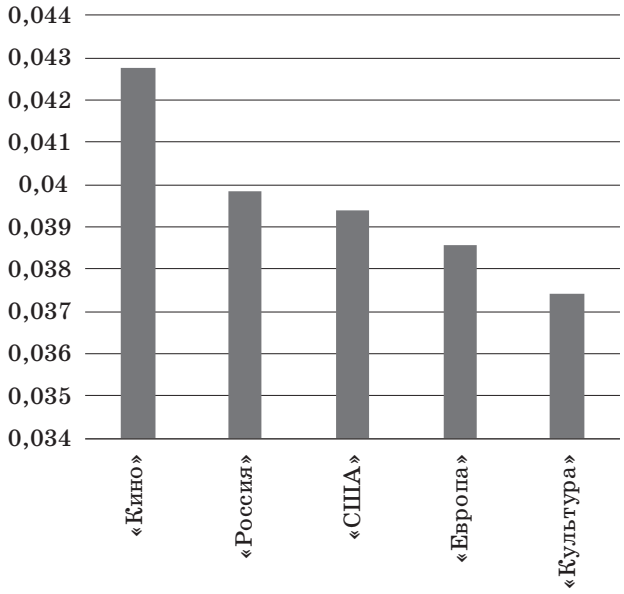
$$\mathbf{P}_{new}(w|t) = \mathbf{P}_{i=1}(d|t) \cdot \mathbf{P}_{i+1}(d|t) \cdot \ldots \cdot \mathbf{P}_n(d|t).$$

На рис. 5 представлен график нормированного распределения вероятностей тем для слова «вальц», полученный произведением Адамара тематических векторов документов, где это слово встретилось, а на рис. 6 — наиболее вероятные темы для этого слова.

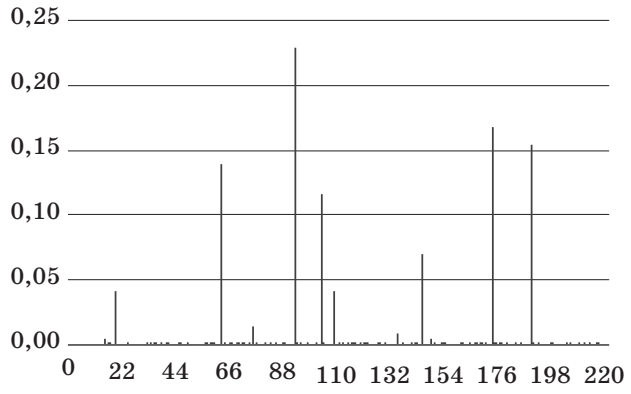
Рассмотрим подробнее результаты суммы и произведения Адамара векторов на примере. Предположим, что в ТМ — три темы и три документа с новым словом. Векторы тем новых



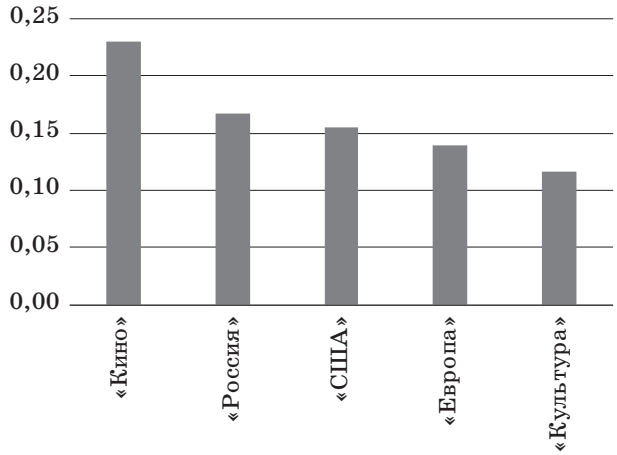

■ ***Рис. 3.*** Распределение вероятностей тем для слова «вальц»




■ ***Рис. 4.*** Значимые темы от суммы вероятностей тем для слова «вальц»




■ ***Рис. 5.*** Распределение вероятностей тем для слова «вальц»




■ ***Рис. 6.*** Значимые темы от произведения Адамара вероятностей тем для слова «вальц»

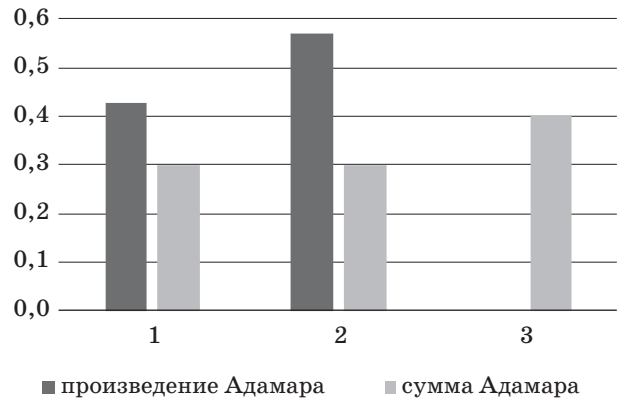
слов равны {0,5 0,5 0,0}, {0,3 0,2 0,5}, {0,1 0,2 0,7}. Сумма векторов

$$\begin{pmatrix}0{,}5\\0{,}5\\0{,}0\end{pmatrix}+\begin{pmatrix}0{,}3\\0{,}2\\0{,}5\end{pmatrix}+\begin{pmatrix}0{,}1\\0{,}2\\0{,}7\end{pmatrix}=\begin{pmatrix}0{,}9\\0{,}9\\1{,}2\end{pmatrix}.$$

Результат произведения Адамара этих векторов

$$\begin{pmatrix}0{,}5\\0{,}5\\0{,}0\end{pmatrix}\cdot\begin{pmatrix}0{,}3\\0{,}2\\0{,}5\end{pmatrix}\cdot\begin{pmatrix}0{,}1\\0{,}2\\0{,}7\end{pmatrix}=\begin{pmatrix}0{,}015\\0{,}02\\0{,}0\end{pmatrix}.$$

Нормированные векторы вероятностей тем нового слова представлены на рис. 7 для суммы и произведения Адамара векторов тем документов, где это слово встретилось.




■ ***Рис. 7.*** Пример векторов произведения Адамара и суммы для трех тем

Суммы вероятностей для первой и второй тем равны, хотя значения вероятностных оценок в документах для этих тем сильно отличаются. Вывод: при суммировании векторов тем теряется значимая для определения тематики нового слова информация. По сумме векторов тем новое слово, с наибольшей вероятностью, относится к третьей теме, но третий документ не относится к третьей теме, следовательно, и слово, которое в нем встретилось, не относится к этой теме. Сумма вероятностей ошибочно связывает новое слово с одной из тем, а произведение Адамара обнуляет вероятность для этой темы. Таким образом, применение произведения Адамара для определения тематики нового слова точнее отражает тематическую принадлежность этого слова, так как обнуляет значение вероятности для непересекающихся векторов тем.

В подходе с использованием произведения Адамара существует вероятность появления нулевых векторов из-за наличия ошибочно использованных слов в не соответствующих этим словам документах. Пример обнуления значений тематического вектора:

$$\begin{pmatrix}0{,}1\\0{,}9\\0{,}0\end{pmatrix}\cdot\begin{pmatrix}0{,}0\\0{,}0\\1{,}0\end{pmatrix}=\begin{pmatrix}0{,}0\\0{,}0\\0{,}0\end{pmatrix}.$$

На тестовых данных можно подсчитать количество нулевых векторов, полученных от произведения Адамара для случая, когда учитываем все новые слова, встретившиеся в двух и более документах. Таких слов 2193. Нулевых векторов нет. Следовательно, в корпусе SCTM-ru нет ошибок, связанных с некорректным использованием слова в документе, и произведение Адамара может быть выбрано как лучший способ определения тематического вектора для нового слова.

Прототип метода определения тематики нового слова реализован с помощью дистрибутива Anaconda, язык разработки Python, программные библиотеки для обработки и анализа данных — pandas, numpy, scikit-learn. Сначала, используя алгоритм ml-PLSI, создаем ТМ на данных для обучения. Затем последовательно обрабатываем все новости из тестовых данных. Запоминаем каждое новое слово и тематический вектор документа, где это слово встретилось. Для новых слов рассчитываем тематический вектор через сумму и через произведение Адамара тематических векторов документов, где это слово встретилось.

## Метод определения тематики нового слова. Алгоритм расширения словаря тематической модели

Как показал эксперимент на корпусе SCTM-ru, наиболее удобный способ определения тематики нового слова — это произведение Адамара тематических векторов документов, где это новое слово встретилось.

Для расширения словаря ТМ предлагается использовать следующий алгоритм:

Вход: ТМ, набор новых документов $d_{new}$.

Выход: ТМ с бесконечным словарем.

1. Выбираем группу новых документов, например, документы за один год или иное значимое количество новых документов.

2. Для всех новых документов $d_{new}$ группы

$$\mathbf{P}(d\,|\,t)=\sum_{w\in d}\mathbf{P}(w\,|\,t).$$

3. Для всех новых слов $w_{new}$ в группе новых документов

$$\mathbf{P}_{new}(w\,|\,t)=\mathbf{P}_{i=1}(d\,|\,t)\cdot\mathbf{P}_{i+1}(d\,|\,t)\cdot\ldots\cdot\mathbf{P}_{n}(d\,|\,t).$$

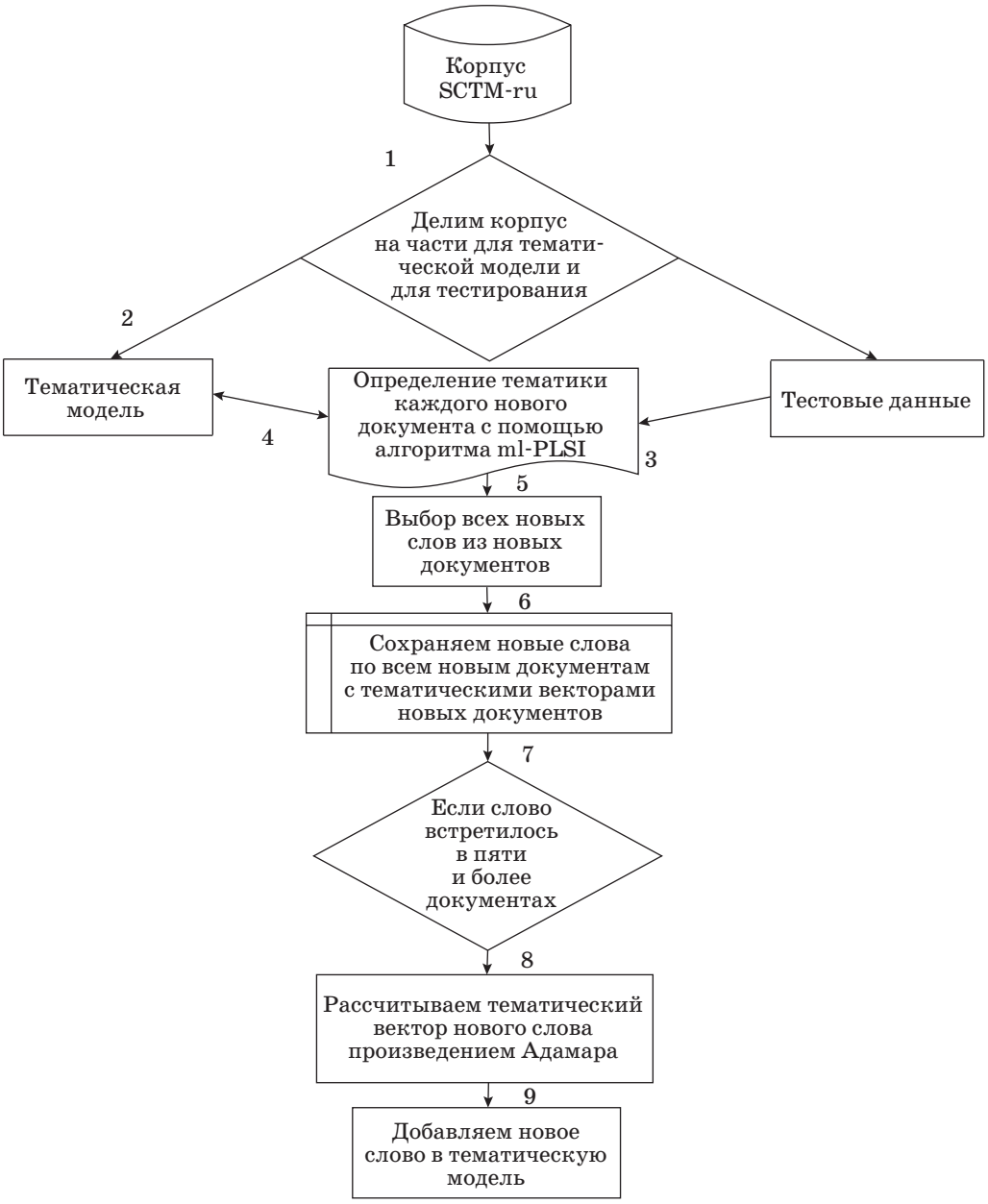
4. Обновляем матрицы «слово-документ», «документ-тема», «слово-тема» для всех документов группы:

а) добавляем в матрице «документ-тема» тематический вектор нового документа;

б) добавляем в матрицу «слово-документ» вектор с количеством слов в документе;

в) добавляем в матрицу «слово-тема» тематический вектор нового слова.

Схема метода представлена на рис. 8. Шаги *1–6* проходим для всех документов в тестовом множестве. Учитывая, что слова, редко встречающиеся в исследуемой коллекции документов, не значимы для ТМ, необходимо определить минимальное количество документов, где встретилось новое слово. В текущей работе зададим пороговое значение в пять документов. Для слов, которые равны или превышают пороговое значение,




■ ***Рис. 8***. Схема метода определения тематики нового слова и добавления нового слова в ТМ

рассчитываем тематический вектор по произведению Адамара. Затем добавим новое слово и новые документы в ранее построенную ТМ.

## Заключение

В результате проведенного исследования представлены методы определения тематической принадлежности нового слова в ТМ с использованием суммы и произведения Адамара тематических векторов документов, где это слово встретилось. Разработан прототип системы, использующий оба метода для определения тематики нового слова. Проведен анализ изменения размера словаря корпуса SCTM-ru в зависимости от времени описанных событий. Предложен алгоритм построения ТМ с бесконечным словарем, который позволяет дополнять словарь ТМ по мере добавления новых документов.

Проведенный анализ демонстрирует перспективность методов тематического моделирования. Предложено решение по важному вопросу тематического моделирования — определение тематики нового слова.

Статистическое исследование корпуса и программная часть метода определения тематики нового слова доступны на <https://github.com/cimsweb/TM-New-Word/>.

## Литература


1. **Hoffman T.** Probabilistic Latent Semantic Indexing// Proc. of the Twenty-Second Annual Intern. SIGIR Conf. on Research and Development in Information Retrieval. ACM. 1999. P. 50–57. doi:10.1145/312624.312649
2. **Blei D. M., Ng A. Y., Jordan M. I.** Latent Dirichlet Allocation// Journal of Machine Learning Research. MIT Press. Jan. 2003. Vol. 3. P. 993–1022.
3. **Карпович С. Н.** Русскоязычный корпус текстов SCTM-RU для построения тематических моделей // Тр. СПИИРАН. 2015. Т. 2. № 39. С. 123–142. doi:10.15622/sp.39.8
4. **Blei D. M., Lafferty J. D.** Dynamic Topic Models// Proc. of the 23rd Intern. Conf. on Machine Learning. ACM. 2006. P. 113–120. doi:10.1145/1143844.1143859
5. **Nallapati R. M.** et al. Multiscale Topic Tomography / R. M. Nallapati, S. Ditmore, J. D. Lafferty, K. Ung // Proc. of the 13th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. ACM. 2007. P. 520–529. doi:10.1145/1281192.1281249
6. **Wang C., Blei D., Heckerman D.** Continuous Time Dynamic Topic Models: preprint arXiv:1206.3298.2012.
7. **Hoffman M., Bach F. R., Blei D. M.** Online Learning for Latent Dirichlet Allocation// Advances in Neural Information Processing Systems. 2010. P. 856–864.
8. **Zhai K., Boyd-Graber J. L.** Online Latent Dirichlet Allocation with Infinite Vocabulary// ICML (1). 2013. Vol. 28. P. 561–569.
9. **Lau J. H., Collier N., Baldwin T.** On-line Trend Analysis with Topic Models: \# twitter Trends Detection Topic Model Online// COLING. 2012. P. 1519–1534.
10. **Карпович С. Н.** Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Тр. СПИИРАН. 2016. Т. 4. № 47. С. 92–104. doi:10.15622/sp.47.5
11. **Horn R. A.** The Hadamard Product// Proc. Symp. Appl. Math. 1990. Vol. 40. P. 87–169.


UDC 004.912
doi:10.15217/issn1684-8853.2016.6.43

**Topic Model with an Infinite Vocabulary**

Karpovich S. N.[a], Head of Search Engine Optimization, cims@yandex.ru
[a]Rambler Internet Holding LLC, 9, Varshvskoe Sh., 117105, Moscow, Russian Federation

**Introduction:** Due to the continuous growth of the internet, increasing amount of news, email messages, posts in blogs, etc., Natural Language Processing systems are in high demand. A popular and promising direction in machine learning and natural language processing is developing topic model algorithms. Most topic model methods deal with static information and a limited vocabulary. In practice, however, we need tools to work with a refillable vocabulary. New words come out every year, some words become obsolete, so refillable vocabularies are especially important for Online Topic Models. **Purpose:** We develop an approach to determine the topical vector for a new word using the Hadamard product of the topical vectors of the documents where this word was found. This approach will be an alternative to the use of Dirichlet distribution or Dirichlet process. **Results:** The research has shown that a sum of topical vectors in the documents with a new word gives a wrong idea about the topic of this new word. At the same time, it is better to use Hadamard product to specify the topic of a new word by the topics of the documents. Multiplying entrywise the topical vectors of the documents with a new word cancels the topics which do not overlap, separating out common topics with similar meanings. Multiplying the topical vectors of the documents provides a topical vector for the new word with the highest probability values for several most important topics. The values of weakly expressed topics either approach zero or are reset to zero. **Practical relevance:** The use of the proposed algorithm can infinitely expand the online vocabulary of a topic model and, hence, consider both new and old words.

**Keywords** — Topic Model, Natural Language Processing, Machine Learning.

**References**


1. Hoffman T. Probabilistic Latent Semantic Indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1999, pp. 50–57. doi:10.1145/312624.312649
2. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, MIT Press, Jan. 2003, vol. 3, pp. 993–1022.
3. Karpovich S. N. The Russian Language Text Corpus for Testing Algorithms of Topic Model. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2015, pp. 123–142 (In Russian). doi:10.15622/ sp.39.8
4. Blei D. M., Lafferty J. D. Dynamic Topic Models. *Proc. of the 23rd Intern. Conf. on Machine Learning*, ACM, 2006, pp. 113–120. doi:10.1145/1143844.1143859
5. Nallapati R. M., Ditmore S., Lafferty J. D., Ung K. Multiscale Topic Tomography. *Proc. of the 13th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, ACM, 2007, pp. 520–529. doi:10.1145/ 1281192.1281249
6. Wang C., Blei D. M., Heckerman D. *Continuous Time Dynamic Topic Models*. Preprint arXiv:1206.3298, 2012.
7. Hoffman M., Bach F. R., Blei D. M. Online Learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, 2010, pp. 856–864.
8. Zhai K., Boyd-Graber J. L. Online Latent Dirichlet Allocation with Infinite Vocabulary. *ICML* (1), 2013, vol. 28, pp. 561–569.
9. Lau J. H., Collier N., Baldwin T. On-line Trend Analysis with Topic Models: \# twitter Trends Detection Topic Model Online. *COLING*, 2012, pp. 1519–1534.
10. Karpovich S. N. Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2016, no. 47, pp. 92–104 (In Russian). doi:10.15622/sp.47.5
11. Horn R. A. The Hadamard Product. *Proc. Symp. Appl. Math.*, 1990, vol. 40, pp. 87–169.