УДК 004.89 *Научная статья* DOI: 10.18287/2223-9537-2024-14-1-55-70

Проектирование информационной системы комплексного тематического анализа больших данных социальных медиа

© 2024, А.М. Фёдоров, И.О. Датьев ✉, И.Г. Вишняков

Институт информатики и математического моделирования им. В.А. Путилова Кольского научного центра РАН (ИИММ КНЦ РАН), Апатиты, Россия

Аннотация

Открытые сообщества пользователей в социальных медиа являются источником данных, оперативно представляющим тематическую повестку актуальных для населения вопросов. Индикаторы активности пользователей - просмотры, комментарии и репосты - обладают динамической природой. В статье представлен новый взгляд на задачи тематического моделирования, результаты которого исследуются на наличие динамических свойств. Эти данные актуальны для решения задач информационной поддержки регионального и муниципального развития. Представлен опыт проектирования информационной системы комплексного тематического анализа больших открытых данных социальных медиа. Система основана на использовании трёх технологий: построения динамических тематических моделей для мониторинга социальных медиа; интеллектуального анализа результатов тематического моделирования объектов и процессов социальных медиа; когнитивной визуализации результатов динамического тематического моделирования. Для учёта проектной неопределённости использованы средства объектного моделирования, системного проектирования и модульный подход.

Ключевые слова: *управление региональным развитием, информационно-аналитические системы, сообщества социальных сетей, анализ данных, тематическое моделирование.*

Цитирование: *Федоров А.М., Датьев И.О., Вишняков И.Г.* Проектирование информационной системы комплексного тематического анализа больших данных социальных медиа // Онтология проектирования. 2024. Т.14, №1(51). С. 55-70. DOI:10.18287/2223-9537-2024-14-1-55-70.

Финансирование: Исследование выполнено в рамках государственного задания ИИММ КНЦ РАН Министерства науки и высшего образования РФ, тема НИР «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (регистрационный номер темы НИР: 122022800551-0).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Проникновение в повседневную жизнь и доступность социальных медиа обуславливает интерес властных структур, представителей бизнеса и общественных организаций к размещённой там информации. Пользователи социальных медиа оперативно реагируют на актуальные для населения вопросы. Основные компоненты содержания: тексты и индикаторы активности пользователей - просмотры, комментарии и репосты - обладают динамической природой. Следует подчеркнуть важность исследования этой динамичности, которая характеризует стремления пользователей, что необходимо для прогнозирования развития социума и соответствующих вариантов отклика со стороны органов управления. Эти данные отражают интересы и настроения людей в режиме реального времени и представляют большую ценность для задач информационной поддержки принятия решений при региональном и муниципальном управлении.

В работе представлен опыт проектирования информационной системы комплексного тематического анализа больших открытых данных социальных медиа. Рассматриваются концептуальные основы исследования динамических свойств результатов тематического моделирования (ТМ) и технологии, составляющие архитектурный каркас системы: построения динамических тематических моделей (ТМь) для мониторинга социальных медиа; интеллектуального анализа результатов ТМ объектов и процессов социальных медиа; когнитивной визуализации результатов динамического ТМ. Для учёта проектной неопределённости [1] и повышения доступности информации лицам, готовым к выполнению познавательно-деятельных функций, [2] использован модульный подход.

1 Методы и меры для исследования динамических свойств результатов ТМ

В [3] отмечено, что впервые в формальной постановке задача обнаружения и отслеживания тем упомянута в отчёте [4].

ТМь - модель коллекции текстовых документов, которая определяет: к каким темам относится каждый документ коллекции и какие слова (термины) образуют каждую тему [5]. Под темой понимается набор слов, а не названия, схожие с заголовками научной статьи или элементами классификаторов типа универсальной десятичной классификации. ТМ — построение ТМь. Динамическое ТМ представляет собой способ построения ТМь, позволяющий учитывать временную компоненту для выявления и отслеживания истории развития тем. В данной работе исследуются динамические свойства ТМь. Такое исследование подразумевает анализ динамики, в т.ч. нетекстовых атрибутов информационной среды социальных медиа.

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) употребляемых в них терминов. Терминами могут быть слова и словосочетания. Каждый документ $d \in D$ представляет собой последовательность n_d терминов $w_1, ..., w_{\text{nd}}$ из словаря W.

В основе ТМ лежит низкоранговое матричное разложение, которое позволяет представить исходную матрицу (матрицу документ-термин) в виде произведения двух матриц более низкого ранга. Такое представление опирается на интуитивно понятное предположение о том, что число тем $|T|$ меньше $|D|$ и $|W|$. Каждый документ состоит из нескольких тем, и каждая тема состоит из некоторых терминов. Для каждого документа определяется вероятность того, что он содержит каждую из тем, и для каждой темы - вероятность того, что она содержит каждый из терминов. Эти вероятности можно записать в матрицы Θ и Φ соответственно. Задача сводится к поиску приближённого представления матрицы частот терминов в документах $P=(\hat{p}(w|d))_{W\times D}$ в виде произведения $P=\Phi\text{x}\Theta$ двух неизвестных матриц меньшего размера — матрицы терминов тем $\Phi=(\phi wt)_{W\times T}$ и матрицы тем документов $\Theta=(\theta td)_{T\times D}$. Матрицы P, Φ, Θ являются стохастическими и имеют неотрицательные нормированные столбцы p_{d}, ϕ_{t}, θ_{d}, представляющие дискретные распределения [6].

ТМ позволяет автоматически выделять темы из текстовых документов и широко применяется в области анализа текстов, а также для информационного поиска. Базовыми методами ТМ являются латентно-семантический анализ (ЛСА) [7], вероятностный ЛСА [8], латентное размещение Дирихле (*Latent Dirichlet Allocation, LDA*)) [9], неотрицательное матричное разложение [10], иерархическая языковая модель Дирихле [11], иерархический процесс Дирихле (*Hierarchical Dirichlet Processes, HDP*) [12] и др.

Одной из особенностей текстов в социальных сетях является их малая длина. Для ТМ коротких текстов предложено несколько подходов [13]: прямой учёт встречаемости слов [14], рассмотрение каждого короткого документа как принадлежащего одной теме [9], учёт эвристических связей между документами для объединения их в «псевдо-документы» для получения документов большего размера [15].

Известны методы ТМ, позволяющие учитывать эволюцию тем во времени: динамическая ТМь [16], байесовская сеть с непрерывным временем [17], фреймворк для выявления тем в корпусе данных и отслеживания сложных структурных изменений во времени [18] и др.

Особенности оценивания методов ТМ в социальных медиа обсуждались в работе [19]. Среди автоматически вычисляемых наибольшее распространение получили метрики [20], основанные на встречаемости терминов. К интегральным показателям ядра темы относятся характеристики, вычисляемые на основе частотных значений входящих в ядро темы токенов [21]. Нахождение универсальных автоматически вычисляемых метрик качества разных ТМь является открытым вопросом.

В обзоре мер сходства текста [22] выделяются четыре типа мер, основанных на: символах, терминах, корпусе, знаниях; а также гибридные меры, представляющие собой комбинации перечисленных типов. При использовании символьных мер тексты рассматриваются как последовательности символов, которые могут быть преобразованы с помощью операций редактирования [23]. Чтобы применить эти меры, тексты (документы) представляются в виде списков частот или векторной модели, в которой каждому слову сопоставляется вес в соответствии с выбранной весовой функцией. Получив такое представление для документов, можно находить расстояние между документами в пространстве [3].

Для назначения весов словам используется метод *TF-IDF* (от англ. *TF — term frequency, IDF — inverse document frequency*) [24]. Для сравнения векторов документов в [4, 25] применялись косинусное сходство, манхэттенское расстояние, евклидово расстояние и др.

Следующим уровнем сравнения текстов является сравнение тем. На предварительном этапе для корпусов текстов строятся ТМь, которые сопоставляются между собой. Для количественной оценки различия коллекций в рамках сравнения ТМь предложено использовать сумму модулей отклонений от равномерного распределения тематик, делённую на количество тематик – коэффициент контентной аутентичности [26].

В области корпусной лингвистики задача подобия текста и корпуса, а также сравнения коллекций (корпусов) текстов относится к направлению сравнительного текстового анализа (СТА) [27, 28].

2 Концептуальные основы комплексного исследования динамических свойств результатов ТМ

Для исследования динамическими свойствами результатов ТМ предлагается разделение результатов по способам представления на:

- множества вероятностных элементов;
- связи ТМь и атрибутов исходных текстов, использованных для построения этих моделей;
- специализированные ТМь с мультимодальной архитектурой.

Способы представления определяют направления работы с динамическими свойствами результатов ТМ, по каждому из которых создано концептуальное описание технологии и соответствующих программных компонентов:

- проектирование и формирование архитектур ТМь путём определения необходимых компонент и выбора инструментов реализации;
- построение ТМь и их интеллектуальный анализ;
- когнитивная визуализация динамики в результатах ТМ.

За основу технологии принята созданная *система мониторинга* (С.М.) сообществ социальных медиа [29]. Получаемые с помощью этой системы данные регулярно обрабатываются с целью выявления динамических аспектов объектов и процессов, с которыми эти данные связаны. На рисунке 1 представлены в виде диаграммы использования *UML* (*Unified*

Modeling Language) возможные варианты исследования динамических свойств результатов ТМ на основе открытых больших данных социальных медиа.

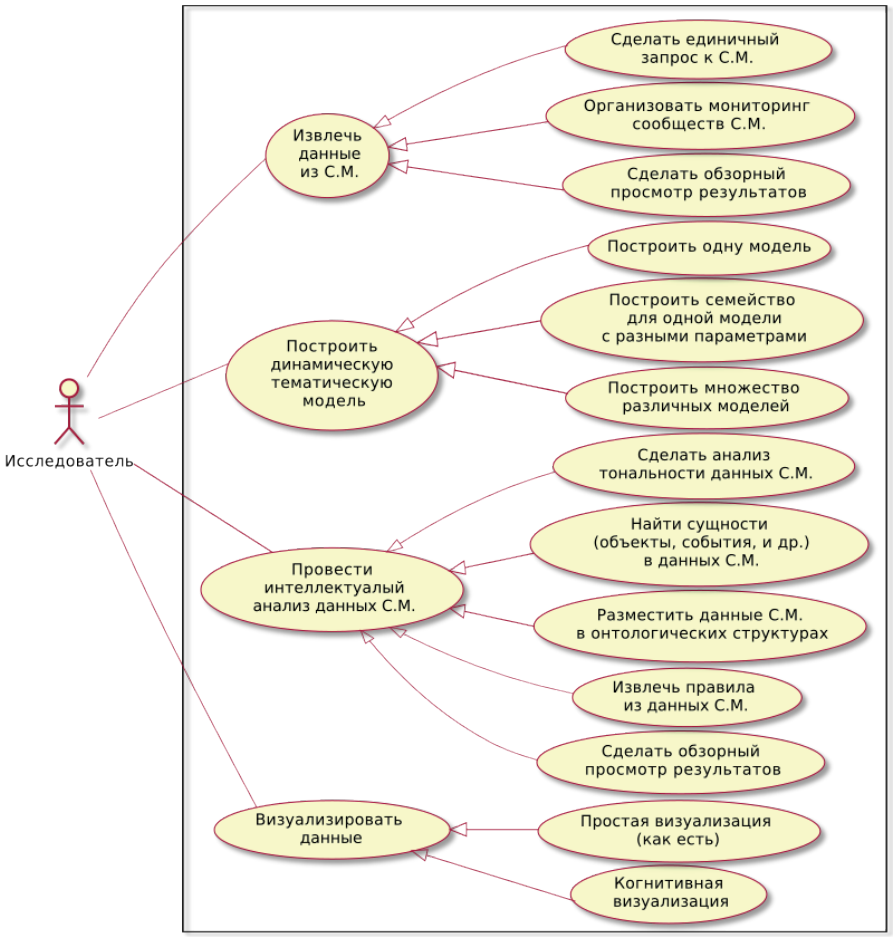

Рисунок 1 – Диаграмма использования. Комплексное исследование динамических свойств результатов тематического моделирования

Первый уровень прецедентов содержит действия, включающие предварительный этап извлечения данных посредством применения С.М. и соответствующие разрабатываемым технологиям: построение динамических ТМь, интеллектуальный анализ результатов, когнитивная визуализация. Второй уровень детально раскрывает варианты использования каждой из технологий.

На рисунке 2 в виде диаграммы последовательности *UML* представлен пример использования разработанных технологий для исследования динамических свойств ТМь. На предварительном этапе пользователи определяют конфигурацию мониторинга данных. В результате работы соответствующих инструментов извлечения данные социальных медиа сохраняются в базе данных (БД) мониторинга, а пользователь получает уведомление. Подобная схема применяется и на следующих основных этапах, ассоциированных с использованием разработанных технологий: построения ТМь, анализа и визуализации.

3 Построение динамических тематических моделей на основе данных мониторинга социальных медиа

Динамическая ТМь отражает распределённые во времени и/или пространстве тематические свойства определённого корпуса текстов. Корпуса текстов, сформированные на основе

открытых данных социальных медиа, наделены определённой спецификой. Помимо содержания и связанных с ним лексико-семантических и других языковых свойств, такие тексты характеризуются множеством дополнительной атрибутивной информации. Например, к такой информации относятся аккаунт автора текстов и ассоциированные с ним пользовательский профиль, дата, время, место публикации, а также связанные с этой публикацией другие публикации, их прямые и косвенные характеристики.

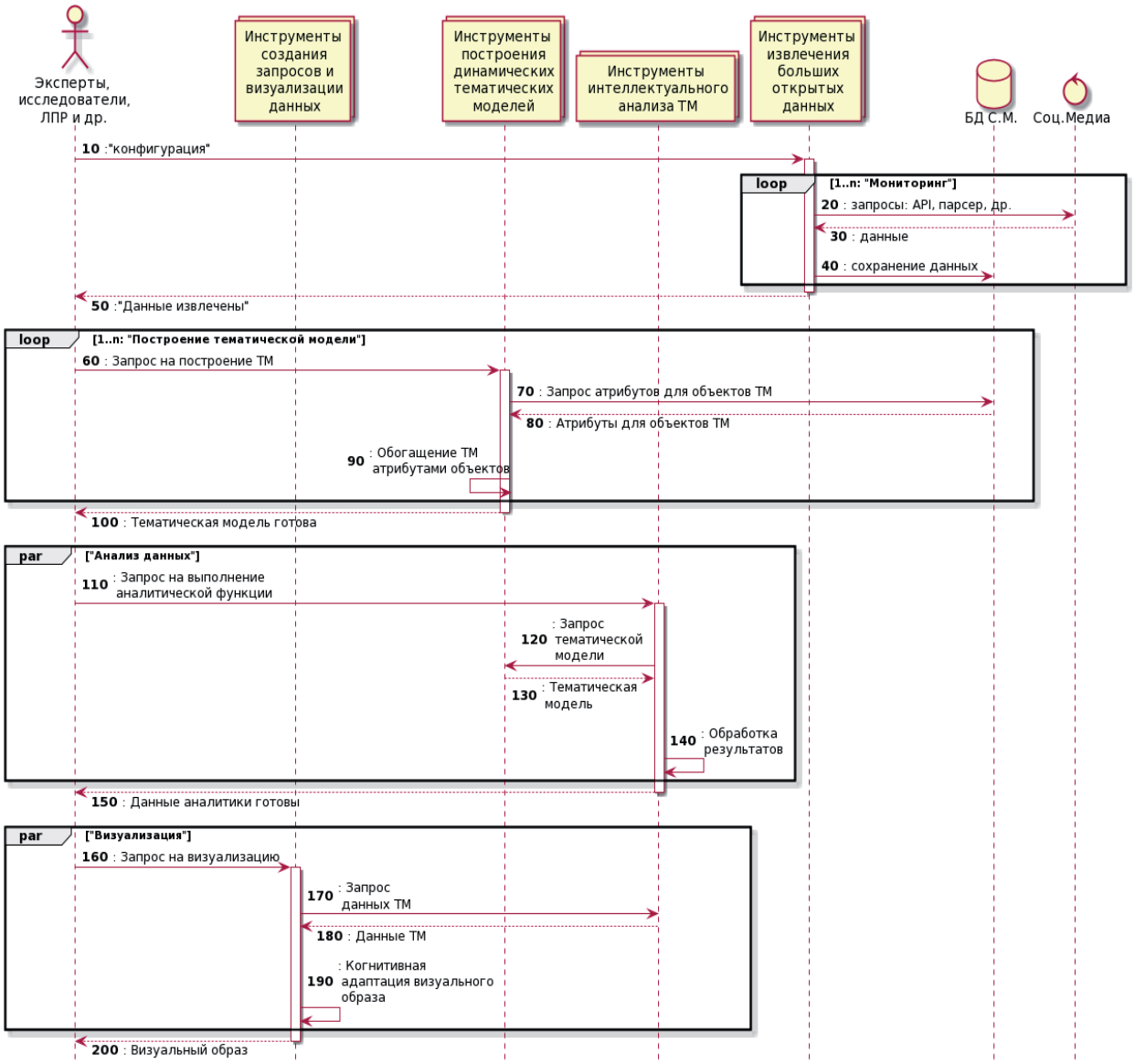

Рисунок 2 – Диаграмма последовательности. Применение разработанных технологий для исследования динамических свойств результатов тематического моделирования

Исследование динамических свойств объектов и процессов, связанных с ТМ, позволило выявить способы представления результатов ТМ, в которых проявляется динамика (см. рисунок 3). Все варианты разбиваются на два подмножества: аспект динамики и инструмент выявления динамики. Динамика ТМь проявляется в виде следующих вариантов на основе: последовательностей тем; содержания текстов; атрибутов текста (контекста). Динамические особенности этих вариантов проявляются в результате исследования неразрывных связей между характеристиками исходных корпусов текстов и получаемыми на их основе ТМь. Каждый из вариантов может рассматриваться отдельно и в сочетаниях с остальными.

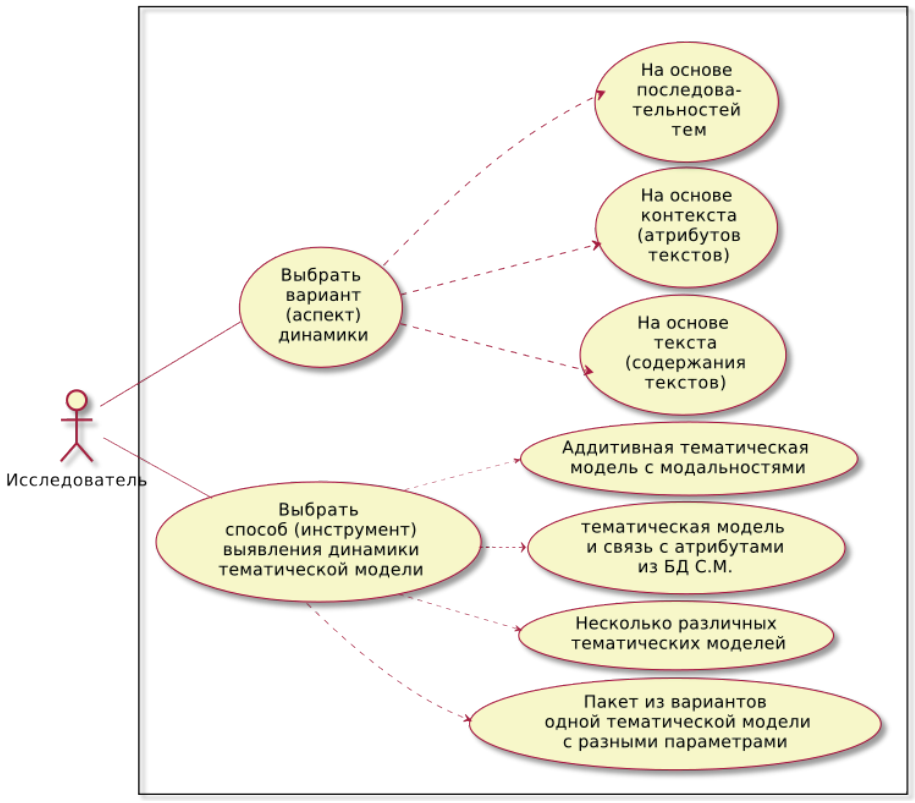

Рисунок 3 – Диаграмма использования. Технология создания динамических тематических моделей

В первом случае рассматриваются темы в ТМь. Интерес представляют изменение порядка тем, их возникновение, преобразование и исчезновение, а также временные характеристики этих изменений.

Содержание текстов документов, на основе которых строятся ТМь, задаёт второй динамический аспект. В данном случае интерес представляет то, как и какие изменения в исходных текстах приводят к изменению ТМь и каков характер этих изменений.

Третий путь работы с динамическими аспектами ТМь предусматривает рассмотрение их совместно с атрибутивной информацией, с которой связаны тексты исходных документов. Тексты социальных медиа неразрывно связаны с такими атрибутами как счётчики активности (комментарии, просмотры, авторы, дата и время публикации и др.). Интерес представляет то, как динамические свойства атрибутов связаны с динамическими свойствами ТМь.

В первом варианте подхода к оценке динамических характеристик ТМь результаты ТМ рассматриваются как множества вероятностных векторов. В общем случае на вход в ТМь передаётся исследуемый корпус текстов и словарь, а на выходе получаются две матрицы: Φ (слова на темы) и Θ (темы на документы). Матрица Φ состоит из вектор-столбцов, задающих распределения слов в словаре в соответствии с каждой темой. Матрица Θ формируется из вектор-столбцов с вероятностным распределением тем в каждом документе исследуемого корпуса текстов. Работа с динамическими характеристиками таких объектов включает использование мер и метрик, отмеченных в разделе 1. Вычисление таких метрик позволяет находить для компонентов ТМь степень их сходства и различия, определять расстояния между ними и т.п.

Второй вариант исследования динамики ТМь ориентирован на извлечение и анализ атрибутивных особенностей объектов. Исследуются переход от тематических свойств текстов к их динамичным атрибутивным особенностям и обратный переход - от атрибутов с заданными свойствами к связанным с ними текстам и их тематическим характеристикам.

Для комплексного анализа динамики и тематического наполнения необходимо на этапе подготовки к сбору данных определить то, какие тексты и их атрибуты из какой социальной сети должны быть получены и положены в основу ТМь.

При извлечении открытых данных из социальных медиа руководствуются утверждением о том, что лишних данных не бывает. При построении ТМь используются только текстовые данные, но на следующем этапе анализа подключаются все атрибутивные метаданные.

Мультимодальный, комплексный вариант построения или представления динамических ТМь основывается на внутренних свойствах и возможностях определённого класса таких моделей. ТМь с аддитивной регуляризацией позволяют включать непосредственно в модель дополнительные, в т.ч. нетекстовые, данные. Это реализовано с помощью т.н. модальностей - маркированных единой меткой непересекающихся групп данных, на базе которых строится ТМь. Для её построения достаточно определения основной текстовой модальности, в которую включаются исходные тексты. Дополнительные модальности представляют собой сопутствующие основной модальности группы текстовых атрибутов, которые совместно, но не пересекаясь, обрабатываются алгоритмами ТМ. Гибкость в управлении модальностями добавляют коэффициенты, которыми регулируется степень значимости каждой модальности. Эти коэффициенты влияют на ТМь, определяя в ней вклад каждой модальности пропорционально значениям коэффициентов. Несмотря на то, что модальности — это текстовые группы, они могут быть сформированы из нетекстовых элементов, т.к. алгоритмы ТМ работают с текстами, разбитыми на токены (обособленные части текста, имеющие своё символьное представление).

С помощью таких токенов можно записать идентификационные номера аккаунтов, даты, значения счётчиков активности, закодировать динамические аспекты текстов соцсетей.

Полученная в результате модель содержит данные о динамике. В описанном варианте все интересующие динамические аспекты определяются на этапе построения модели. В предыдущем варианте предусматривалась возможность связывания тематических и динамических характеристик различными способами после получения модели.

4 Интеллектуальный анализ результатов ТМ объектов и процессов социальных медиа

Для совместного исследования результатов ТМ и расширенного атрибутами динамично изменяющегося исходного корпуса текстов разработана технология интеллектуального анализа данных социальных медиа. Реализованы базовые процедуры получения:

- тематических характеристик по заданным исходным текстам (объектам) и их атрибутам;
- атрибутивной информации по заданным тематическим свойствам.

Разработанная технология и реализующие её средства расширяют возможности анализа корпуса текстов с помощью ТМ. Здесь используется свойство текстов социальных медиа, которые, по сути, являются метатекстами. Метатекстовая структура исследуемых объектов расширяет возможности ТМ пропорционально объёму и структуре метатекстовых атрибутов. Такими атрибутами являются счётчики активности, а также мультимедийные приложения (графические и видео изображения, аудио файлы и др.). Важным элементом является динамический характер исследуемых объектов. Тексты социальных медиа изменяются во времени и в пространствах, задаваемых своими атрибутами. Для работы с ними применяются различные метрики и меры.

Особенность интеллектуального анализа заключается в совместном использовании полученных результатов ТМ и исходных данных, имеющих объёмную атрибутивную структуру. Использование такого расширения структур данных позволяет проводить гибкий много-

уровневый тематический анализ. Таким образом, интеллектуальный анализ результатов ТМ реализуется поэтапно (см. рисунок 4):

- построение ТМь одним из способов, описанных в разделе 3;
- установление связей между ТМь и БД с атрибутивными данными;
- формирование запросов к тематической модели, позволяющих получать:
  - тематические характеристики на основе заданных атрибутивных данных;
  - атрибутивные данные на основе заданных тематических характеристик;
- построение последовательности запросов к ТМь, позволяющих исследовать её динамические свойства на основе заданной последовательности (множества) атрибутивных данных и тематических атрибутов;
- применение к результатам запросов метрик и мер, соответствующих типам получаемых данных;
- интерпретация полученных результатов.

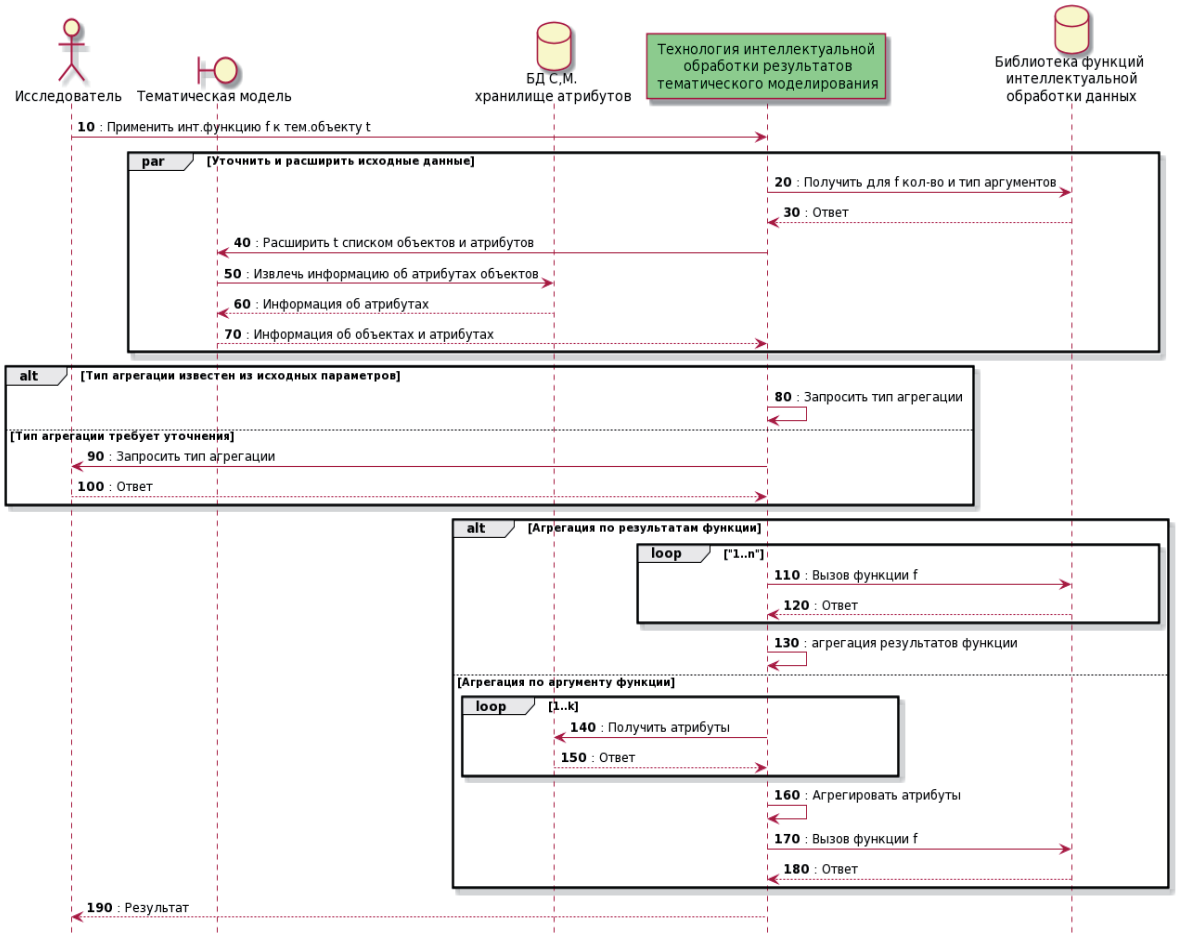

Рисунок 4 – Диаграмма последовательности. Технология интеллектуальной обработки результатов тематического моделирования

5 Когнитивная визуализация результатов динамического ТМ для поддержки решения задач регионального развития

В данной работе ТМ и анализ больших открытых данных социальных медиа используются для разработки концептуальных и прикладных средств поддержки решения задач регионального развития. Конечные пользователи разработок - управленцы и эксперты в различных предметных областях, которым необходимая в работе информация о социальных про-

цессах будет представлена в виде результатов ТМ и производных от них. Такие результаты нуждаются в дополнительной подготовке для представления экспертам.

Когнитивная визуализация предполагает представление результатов ТМ в удобном и понятном интерактивном виде. Такой подход позволяет эксперту иметь доступ к полученным аналитическим результатам и к связанным с ними первичным данным, иметь возможность либо сразу принимать необходимые решения, либо скорректировать модельные параметры и построить следующий вариант ТМь. Для когнитивной визуализации предложены базовые принципы (см. рисунок 5), на основе которых производится проектирование и программная реализация когнитивной визуализации.

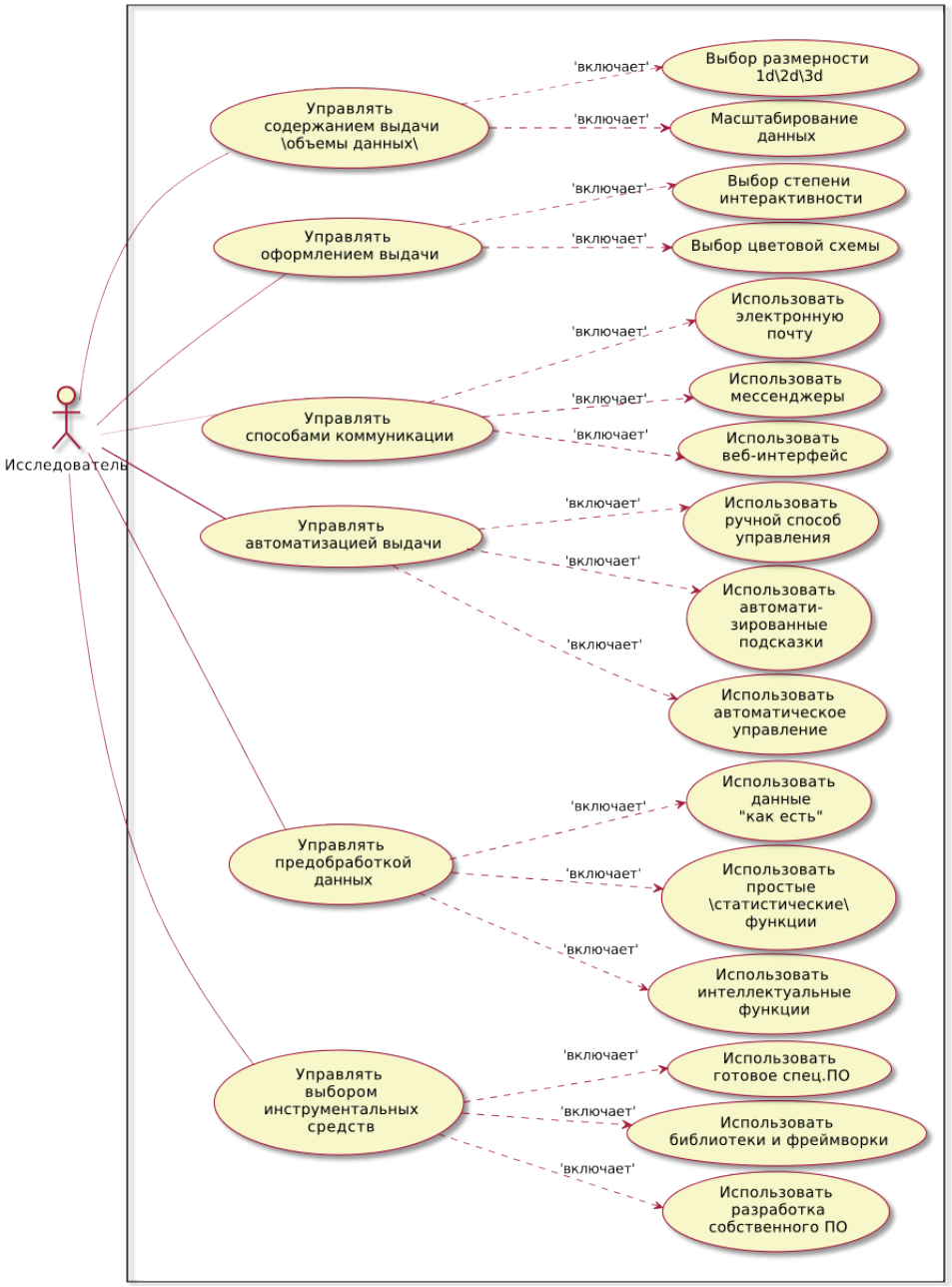

Рисунок 5 – Диаграмма использования. Базовые принципы технологии когнитивной визуализации

Для пользователя-эксперта визуальный образ формируется с помощью управления содержанием и оформлением выдачи, выбором инструментальных средств и способов комму-

никации, а также предобработкой данных. Эти виды управления могут быть использованы по отдельности и в различных сочетаниях. При этом управление может осуществляться в автоматизированном режиме или быть полностью переданным пользователю.

Например, если данных для отображения слишком много, то можно выдавать их пользователю порциями, предоставляя ему интерактивные элементы управления для выбора нужных данных. Для взаимодействия с пользователем можно использовать различные способы коммуникации (мессенджеры, электронную почту и др.). Для повышения наглядности данные могут быть дополнительно обработаны, например, с помощью ТМ, средства кластеризации и классификации, построения онтологических конструкций и тезаурусов и др.

Технологии ТМ постоянно развиваются. Существуют средства когнитивной визуализации результатов ТМ (например, *LDAvis* [30]). Однако данное средство позволяет видеть лишь часть результатов ТМ, которая представляет собой визуализацию содержимого матрицы Φ, т.е. тематическую разбивку словаря анализируемого корпуса текстов. На координатной плоскости, задаваемой номинальными координатами [31], отображается взаимное расположение выявленных тем. В данной работе развитие функции *pyLDAvis* расширено возможностью работы с содержимым матрицы Θ, т.е. распределением документов корпуса текстов по темам. На основе результатов ТМ пользователю предоставлена возможность выбора темы для анализа и интерпретации результатов моделирования. Производится автоматическая выборка и отображение строк матрицы Θ, связанных с исходными текстами, обладающими максимальной вероятностью принадлежности к выбранной теме. Посредством автоматического размещения гиперссылок рядом со строками матрицы Θ обеспечивается доступ к исходным текстам (см. рисунок 6).

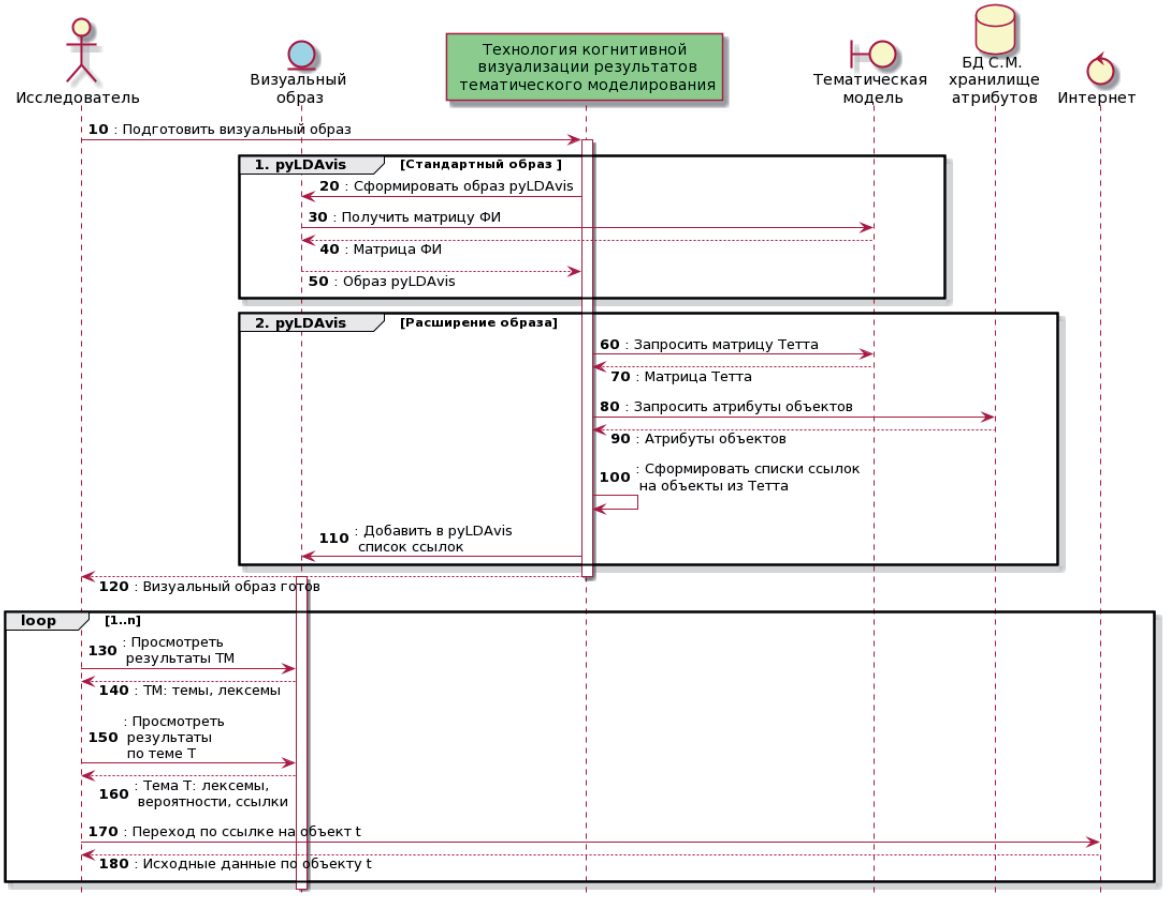

Рисунок 6 – Диаграмма последовательности. Технология когнитивной визуализации результатов динамического тематического моделирования

Пользовательский веб-интерфейс, в котором реализована указанная последовательность действий, представлен на рисунке 7. Интерфейс *pyLDAvis* расширен размещёнными в верхней части изображения блоками выбора темы и представления строк матрицы Θ, обогащённых гиперссылками на исходные тексты. Данное расширение является демонстрацией совместной работы базовых принципов когнитивной визуализации, представленных на рисунке 5 в виде управления содержанием, оформлением и автоматизацией выдачи, а также выбором инструментальных средств.

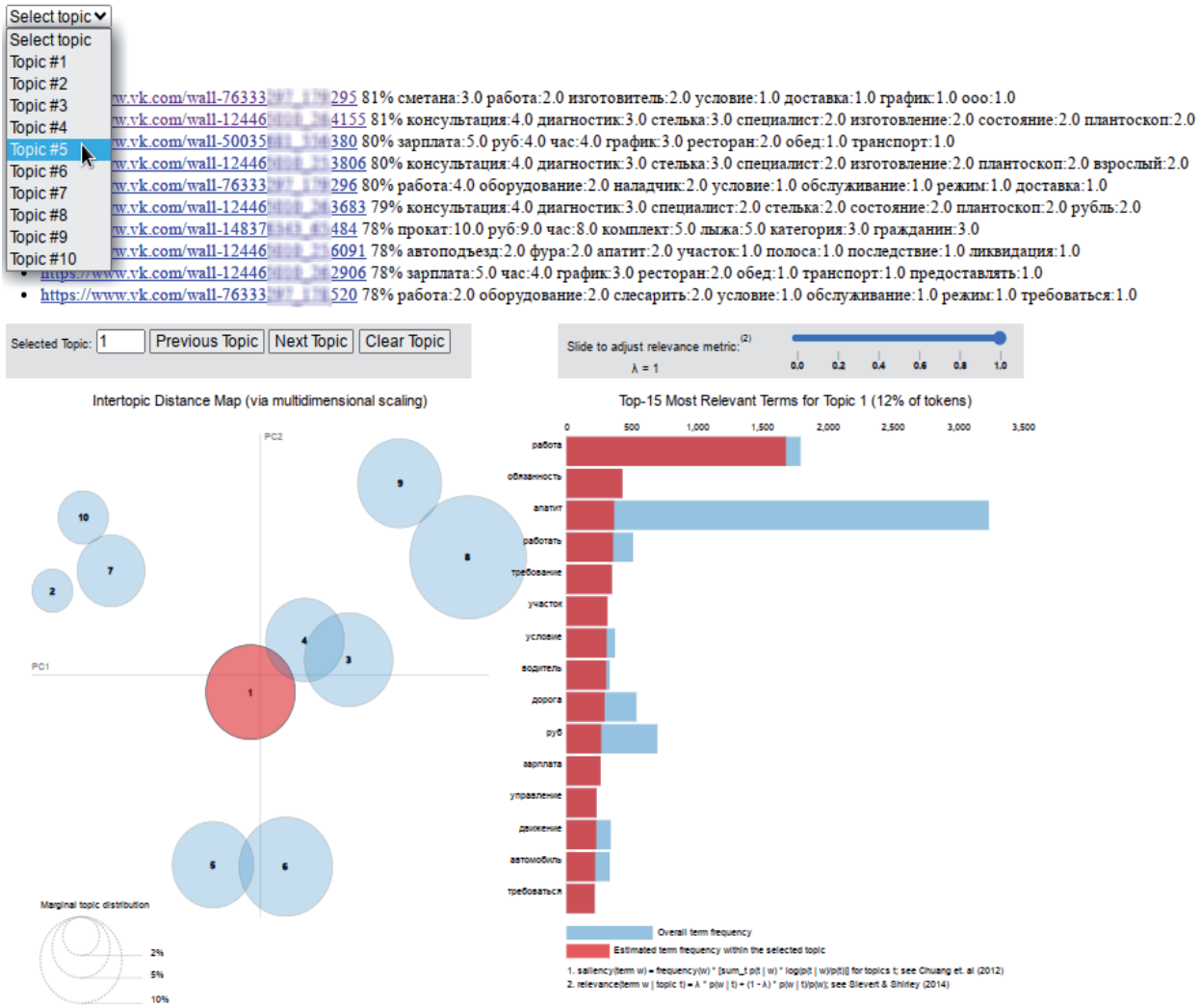

Рисунок 7 – *pyLDAvisPLUS*. Расширение веб-интерфейса инструментария *pyLDAvis* при анализе тематических моделей и атрибутивной информации и связанных с ними исходных текстов

Заключение

Представлен опыт создания информационной системы комплексного тематического анализа больших данных социальных медиа на основе разработанных информационных технологий.

Эти технологии описаны на принципиальном уровне в нотации диаграмм *UML*. Базовые компоненты технологий реализованы на языке программирования *Python* с использованием архитектуры *web*.

Изложенные подходы использования ТМ тесно связаны с развитием технологий и инструментов поддержки управления региональным развитием [32].

Список источников

[1] ***Боргест Н.М.*** Научный базис онтологии проектирования // Онтология проектирования. 2013. №1 (7). С.7-25.

[2] ***Смирнов С.В.*** Онтологическое моделирование в ситуационном управлении // Онтология проектирования. 2012. №2. С.16-24.

[3] ***Коршунов А.В., Гомзин А.Г.*** Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. №23. С.215-244.

[4] ***Allan J., Carbonell J., Doddington G., Yamron J., Yang Y.*** Topic Detection and Tracking Pilot Study. Final Report // Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), Feb. 1998.

[5] ***Воронцов К.В.*** Вероятностное тематическое моделирование. 2013. http://www.machinelearning.ru.

[6] ***Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.*** Автоматическая обработка текстов на естественном языке и анализ данных. М.: Изд-во НИУ ВШЭ, 2017. 269 с. https://www.hse.ru/data/2017/07/22/1173852775/NLPandDA_4print.pdf.

[7] ***Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R.*** Indexing by Latent Semantic Analysis // J. Am. Soc. Inf. Sci. Vol.41(6). 1990. P.391-407.

[8] ***Hofmann T.*** Probabilistic latent semantic indexing // In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 1999. P.50–57. DOI:10.1145/312624.312649.

[9] ***Blei D.M., Ng A.Y., Jordan M.I.*** Latent Dirichlet allocation // J. Mach. Learn. Res. Vol. 3. 2003. P.993-1022.

[10] ***Kuang D., Choo J., Park H.*** Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering // In: Celebi M. (eds) Partitional Clustering Algorithms. Springer, Cham. 2015. DOI:10.1007/978-3-319-09259-1_7.

[11] ***MacKay D.J.C., Peto L.C.B.*** A hierarchical Dirichlet language model // Nat. Lang. Eng. Vol. 1(3). 1995. DOI:10.1017/S1351324900000218.

[12] ***Teh Y.W., Jordan M.I., Beal M.J., Blei D.M.*** Sharing clusters among related groups: Hierarchical Dirichlet processes // In: NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, MA, United States, 2004. P.1385–1392.

[13] ***Vayansky I., Kumar S.*** A review of topic modeling methods // Information Systems. 2020. Vol.94. 101582. DOI:10.1016/j.is.2020.101582.

[14] ***Yan X., Guo J., Lan Y., Cheng X.*** A biterm topic model for short texts // In: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil. 2013. P.1445–1455. DOI:10.1145/2488388.2488514.

[15] ***Zuo Y. et al.*** Topic Modeling of Short Texts: A Pseudo-Document View // In: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, 2016. P.2105–2114. DOI:10.1145/2939672.2939880.

[16] ***Blei D.M., Lafferty J.D.*** Dynamic topic models // In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA, ACM, 2006. P.113–120. DOI:10.1145/1143844.1143859.

[17] ***Nodelman U., Shelton C.R., Koller D.*** Continuous time bayesian networks // In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. Alberta, Canada, 2002. P.378–387.

[18] ***Beykikhoshk A., Arandjelović O., Phung D., Venkatesh S.*** Discovering topic structures of a temporally evolving document corpus // Knowl Inf Syst. 2018. Vol. 55. P.599–632. DOI:10.1007/s10115-017-1095-4.

[19] ***Датьев И.О., Федоров А.М.*** Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайновых социальных сетей // Онтология проектирования. 2022. Том 12, №2(44). С.186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

[20] ***Mimno D. Wallach H., Talley Ed., Leenders M., McCallum A.*** Optimizing semantic coherence in topic models // In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. Association of Computational Linguistics, 2011. P.262-272.

[21] ***Vorontsov K., Potapenko A.*** Additive regularization of topic models // Mach Learn. 2015. Vol. 101. P. 303-323. https://doi.org/10.1007/s10994-014-5476-6.

[22] ***Gomaa W. H., Fahmy A. A.*** A Survey of Text Similarity Approaches // International Journal of Computer Applications. 2013. Vol. 68(13). P.13–18.

[23] ***Левенштейн В.И.*** Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. Том 163.4. С.845-848.

[24] ***Jones K.S.*** A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. MCB University: MCB University Press, 2004. Vol. 60, no. 5. P. 493-502.
[25] ***Allan J., Lavrenko V., Malin D., Swan R.*** Detections, bounds, and timelines: UMass and TDT-3 // In Proceedings of Topic Detection and Tracking Workshop. Vienna, VA, 2000. P.167–174.
[26] ***Краснов Ф.В., Дементов А.В., Шварцман М.Е.*** Использование тематических моделей для парного сравнения коллекций научных статей // Информатика и её применения. 2020. Том 14, выпуск 3. С.129–135.
[27] ***Kilgarriff A., Rose T.*** Measures for corpus similarity and homogeneity. 1998. http://aclweb.org/anthology/W98-1506.
[28] ***Fothergill R., Cook P., Baldwin T.*** Evaluating a topic modelling approach to measuring corpus similarity, In Proceedings of the Tenth International Conference on Language Re-sources and Evaluation (LREC 2016), Portorož, Slovenia. 2016. P.273-279.
[29] ***Федоров А.М., Датьев И.О., Щур А.Л.*** «ИС МСВ» //Роспатент: Свидетельство о государственной регистрации программы для ЭВМ №2020619469 от 17 августа 2020 г.
[30] ***Sievert C., Shirley K.*** LDAvis: A method for visualizing and interpreting topics // In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA. Association for Computational Linguistics, 2014. P.63-70.
[31] ***Jolliffe IT, Cadima J.*** Principal component analysis: a review and recent developments // Philos Trans A Math Phys Eng Sci. 2016 Apr 13; 374(2065):20150202. DOI:10.1098/rsta.2015.0202. PMID: 26953178; PMCID: PMC4792409.
[32] Информационно-аналитическая система поддержки управления региональным развитием на основе открытых больших данных социальных медиа: концепция разработки и практика реализации / А. М. Федоров и др. // Труды Кольского научного центра РАН. Серия: Технические науки. 2022. Т.13, № 2. С.5–22. DOI:10.37614/2949-1215.2022.13.2.001

Сведения об авторах

Федоров Андрей Михайлович 1978 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2000). К.т.н. (2005). В ИИММ КНЦ РАН ведущий научный сотрудник, заместитель директора по научной работе (с 2017 г.). Доцент кафедры информатики и вычислительной техники в филиале Мурманского арктического государственного университета (МАГУ) в г. Апатиты. Область научных интересов сосредоточена на разработке моделей и технологий информационной поддержки для регионального управления. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *fedorov@iimm.ru.*

Датьев Игорь Олегович 1981 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2004). К.т.н. (2011). В ИИММ КНЦ РАН старший научный сотрудник, ученый секретарь. Автор более 100 научных работ в области разработки моделей и технологий для региональных информационно-коммуникационных систем. Author ID (РИНЦ): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *datyev@iimm.ru.* ✉

Вишняков Иван Геннадьевич 1997 г. рождения. Окончил филиал МАГУ в г. Апатиты (2022). Магистрант второго курса ФИЦ КНЦ РАН по специальности 09.04.02 Информационные системы и технологии. В ИИММ КНЦ РАН системный администратор (с 2022 г.) Область научных интересов - разработка информационных систем анализа больших открытых данных социальных медиа. Author ID (RSCI): 1168901. *vishnyakov@iimm.ru.*

Поступила в редакцию 01.12.2023,после рецензирования 10.01.2024. Принята к публикации 02.02.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-1-55-70

Designing an information system for integrated topic analysis of social media big data

© 2024, A.M. Fedorov, I.O. Datyev ✉, I.G. Vishnyakov

Institute for Informatics and Mathematical Modeling named after V.A. Putilov of the Kola Science Center RAS, Apatity, Russia

Abstract

Open communities of users in social media are a source of data that quickly presents the thematic agenda of issues relevant to the population. The indicators of user activity are views, likes, comments and reposts, and they are of a dynamic nature. The article presents a new vision at the topic modeling problems, the results of which are examined for dynamic properties. These data are relevant to solve problems of information support for regional and municipal development. The authors reveal their experience in designing an information system for integrated topic analysis of large open social media data. The system is based on three technologies: building dynamic topic models for monitoring social media, intelligent analysis of topic modeling results; and cognitive visualization of dynamic topic modeling results. To take into account design uncertainty, object modeling tools, system design and a modular approach were used.

Keywords: *regional development management, information and analytical systems, social network communities, data analysis, topic modeling.*

For citation: *Fedorov AM, Datyev IO, Vishnyakov IG.* Designing an information system for integrated topic analysis of social media big data [In Russian]. *Ontology of designing*. 2024; 14(1): 55-70. DOI:10.18287/2223-9537-2024-14-1-55-70.

Financial Support: The work is supported by the Ministry of Science and Higher Education of the Russian Federation. Topic title: Methodology for creating information and analytical systems to support the management of regional develop-ment based on formative artificial intelligence and big data (reg.n. 122022800551-0).

Conflict of interest: The authors declare no conflict of interest.

List of figures

Figure 1 - Usage diagram. Comprehensive study of the dynamic properties of topic modeling results
Figure 2 - Sequence diagram. Application of developed technologies to study the dynamic properties of topic modeling results
Figure 3 - Usage diagram. Technology for creating dynamic topic models
Figure 4 - Sequence diagram. Technology for intelligent processing of topic modeling results
Figure 5 - Usage diagram. Basic principles of cognitive visualization technology
Figure 6 - Sequence diagram. Technology of cognitive visualization of dynamic topic modeling results
Figure 7 - pyLDAvisPLUS. Extension of the pyLDAvis web interface when analyzing topic models and attribute information and associated source texts

References

[1] ***Borgest NM.*** Scientific basis of ontology of designing [In Russian]. *Ontology of Designing*. 2013; №1 (7): 7-25.
[2] ***Smirnov SV.*** Ontological modeling in situational management [In Russian]. *Ontology of Designing*. 2012; №2: 16-24.
[3] ***Korshunov AV, Gomzin AG.*** Topic modeling of natural language texts [In Russian]. Proceedings of the Institute of System Programming of the Russian Academy of Sciences. 2012; №23: 215-244.
[4] ***Allan J, Carbonell J, Doddington G, Yamron J, Yang Y.*** Topic Detection and Tracking Pilot Study. Final Report. Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), Feb. 1998
[5] ***Vorontsov KV.*** Probabilistic topic modeling [In Russian]. 2013. http://www.machinelearning.ru.

[6] ***Bolshakova EI, Vorontsov KV, Efremova NE, Klyshinsky ES, Lukashevich NV, Sapin AS.*** Automatic processing of texts in natural language and data analysis [In Russian]. Textbook allowance. Moscow: Publishing house of the National Research University Higher School of Economics. 2017. 269 p. https://www.hse.ru/data/2017/07/22/1173852775/NLPandDA_4print.pdf.

[7] ***Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R.*** Indexing by Latent Semantic Analysis. J. Am. Soc. Inf. Sci. 1990; 41(6): 391-407.

[8] ***Hofmann T.*** Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 1999. P.50–57. DOI:10.1145/312624.312649.

[9] ***Blei DM, Ng AY, Jordan MI.*** Latent Dirichlet allocation. J. Mach. Learn. Res. Vol. 3. 2003. P.993-1022.

[10] ***Kuang D, Choo J, Park H.*** Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In: Celebi M. (eds) Partitional Clustering Algorithms. Springer, Cham. 2015. DOI:10.1007/978-3-319-09259-1_7.

[11] ***MacKay DJC, Peto LCB.*** A hierarchical Dirichlet language model. Nat. Lang. Eng. Vol. 1(3). 1995. DOI:10.1017/S1351324900000218.

[12] ***Teh YW, Jordan MI, Beal MJ, Blei DM.*** Sharing clusters among related groups: Hierarchical Dirichlet processes. In: NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, MA, United States, 2004. P.1385–1392.

[13] ***Vayansky I, Kumar S.*** A review of topic modeling methods. Information Systems. 2020. Vol.94. 101582. DOI:10.1016/j.is.2020.101582.

[14] ***Yan X, Guo J, Lan Y, Cheng X.*** A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil. 2013. P.1445–1455. DOI:10.1145/2488388.2488514.

[15] ***Zuo Y et al.*** Topic Modeling of Short Texts: A Pseudo-Document View. In: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, 2016. P.2105–2114. DOI:10.1145/2939672.2939880.

[16] ***Blei DM, Lafferty JD.*** Dynamic topic models. In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA, ACM. 2006; P.113–120. DOI:10.1145/1143844.1143859.

[17] ***Nodelman U, Shelton CR, Koller D.*** Continuous time bayesian networks. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. Alberta, Canada. 2002; P.378–387.

[18] ***Beykikhoshk A, Arandjelović O, Phung D, Venkatesh S.*** Discovering topic structures of a temporally evolving document corpus. Knowl Inf Syst. 2018; 55: 599–632. DOI:10.1007/s10115-017-1095-4.

[19] ***Datyev IO, Fedorov AM.*** Additive regularization for topic modeling of social media communities [In Russian]. Ontology of design. 2022; 12, 2(44): 186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

[20] ***Mimno D, Wallach H, Talley Ed, Leenders M, McCallum A.*** Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK). Association of Computational Linguistics. 2011: 262–272.

[21] ***Vorontsov K, Potapenko A.*** Additive regularization of topic models. Mach Learn. 2015; Vol. 101: 303-323. https://doi.org/10.1007/s10994-014-5476-6.

[22] ***Gomaa WH, Fahmy AA.*** A Survey of Text Similarity Approaches. International Journal of Computer Applications. 2013; 68(13): 13–18.

[23] ***Levenshtein VI***. Binary codes with correction of deletions, insertions and substitutions of symbols [In Russian]. Reports of the USSR Academies of Sciences. 1965; 163.4: 845-848.

[24] ***Jones KS.*** A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. MCB University: MCB University Press, 2004; Vol. 60, no. 5: 493-502.

[25] ***Allan J, Lavrenko V, Malin D, Swan R.*** Detections, bounds, and timelines: UMass and TDT-3. In Proceedings of Topic Detection and Tracking Workshop. Vienna, VA, 2000; P.167–174.

[26] ***Krasnov FV, Dimentov AV, Shvartsman ME.*** Using topic models for pairwise comparison of collections of scientific articles. Informatics and its applications. 2020; 14(3): 129–135.

[27] ***Kilgarriff A, Rose T.*** Measures for corpus similarity and homogeneity. 1998. http://aclweb.org/anthology/W98-1506.

[28] ***Fothergill R, Cook P, Baldwin T.*** Evaluating a topic modelling approach to measuring corpus similarity. In Proceedings of the Tenth International Conference on Language Re-sources and Evaluation (LREC 2016), Portorož, Slovenia. 2016; P.273-279.

[29] ***Fedorov AM, Datyev IO, Shchur AL.*** "IS MSV" [In Russian]. Rospatent: Certificate of state registration of the computer program No. 2020619469 dated August 17, 2020.

[30] ***Sievert C, Shirley K.*** LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA. Association for Computational Linguistics, 2014; P.63-70.

[31] ***Jolliffe IT, Cadima J.*** Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016 Apr 13; 374(2065):20150202. DOI:10.1098/rsta.2015.0202. PMID: 26953178; PMCID: PMC4792409.

[32] ***Fedorov AM et al.*** Information and analytical system for supporting the management of regional development based on open big data of social media: concept of development and practice of implementation. Proceedings of the Kola Scientific Center of the Russian Academy of Sciences. Series: Technical Sciences. 2022; 13(2): 5–22. DOI:10.37614/2949-1215.2022.13.2.001.

About the authors

Andrei Mikhailovich Fedorov (b. 1978) graduated from the Kola branch of the Petrozavodsk State University (2000). Cand. Sci. Eng. (2005). A leading researcher and a deputy director for research (since 2017) at the Putilov Institute for Informatics and Mathematical Modeling of the KSC RAS. He is a senior lecturer at the Murmansk Arctic State University branch in the Apatity (until 2017, the Kola branch of Petrozavodsk State University), an associate professor at the Department of Information Technology (2005) and the Department of Informatics and Computer Engineering (since 2018). The area of scientific interests is currently focused on the development of models and technologies for information support for regional management. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *fedorov@iimm.ru.*

Igor Olegovich Datyev (b. 1981) graduated from the Kola branch of the Petrozavodsk State University in 2004. Cand. Sci. Eng. (2011). He is a research laboratory assistant, programmer, senior researcher, and scientific secretary at the Putilov Institute for Informatics and Mathematical Modeling of the KSC RAS. He is the author of more than 100 scientific papers in the field of development of models and technologies for regional information and communication systems. Author ID (RSCI): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *datyev@iimm.ru* ✉.

Ivan Gennadyevich Vishnyakov (b. 1997) graduated from the MASU branch in Apatity (2022). He is a second-year master's student of the Information systems and technologies specialty at the Federal Research Center KSC RAS. He is a system administrator at the Putilov Institute for Informatics and Mathematical Modeling of the KSC RAS. The area of scientific interests is focused on the development of information systems for analyzing big open data of social media. Author ID (RSCI): 1168901. *vishnyakov@iimm.ru.*

Received December 1, 2023. Revised January 10, 2024. Accepted February 2, 2024.