Труды Кольского научного центра РАН. Серия: Технические науки. 2024. Т. 15, № 3. С. 50–60.
Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2024. Vol. 15, No. 3. P. 50–60.

Научная статья
УДК 004.912, 004.822
doi:10.37614/2949-1215.2024.15.3.004

ФОРМИРОВАНИЕ МНОГОСЛОЙНЫХ ГРАФОВ ЗНАНИЙ НА БАЗЕ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ТЕКСТОВ

Вадим Константинович Пимешков[1]✉, Марина Леонидовна Никонорова[2], Максим Геннадьевич Шишаев[3], Иван Геннадьевич Вишняков[4]
[1–4]Институт информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук, Апатиты, Россия
[1]v.pimeshkov@ksc.ru✉, https://orcid.org/0000-0001-7010-230X
[2]m.nikonorova@ksc.ru, https://orcid.org/0000-0002-4358-2683
[3]m.shishaev@ksc.ru, https://orcid.org/0000-0001-7070-7878
[4]i.vishnyakov@ksc.ru, https://orcid.org/0009-0003-4938-5693

Аннотация

В работе рассматривается проблематика формирования многослойных графов знаний для использования при интеллектуальном анализе данных социальных медиа. Технология основана на «мягкой» структуризации имеющегося пула документов в соответствии с построенной тематической моделью. Проведено экспериментальное опробование технологии и сравнение формальных свойств графовых моделей знаний, полученных с помощью тематического моделирования и традиционным способом. Результаты сравнения указывают на потенциальную эффективность использования многослойных графов знаний, построенных на базе тематической модели текстов, при реализации интеллектуальных процедур обработки данных в условиях множественности и динамичности предметных областей.

Ключевые слова:

граф знаний, тематическая модель, социальные медиа

Благодарности:

исследование выполнено в рамках государственного задания Института информатики и математического моделирования имени В. А. Путилова Кольского научного центра Российской академии наук от Министерства науки и высшего образования Российской Федерации, тема научно-исследовательской работы «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (регистрационный номер 122022800551-0).

Для цитирования:

Пимешков В. К., Никонорова М. Л., Шишаев М. Г., Вишняков И. Г. Формирование многослойных графов знаний на базе тематического моделирования текстов // Труды Кольского научного центра РАН. Серия: Технические науки. 2024. Т. 15, № 3. С. 50–60. doi:10.37614/2949.1215.2024.15.3.004.

Original article

FORMATION OF MULTILAYER KNOWLEDGE GRAPHS BASED ON THEMATIC TEXTS MODELING

Vadim K. Pimeshkov[1]✉, Marina L. Nikonorova[2], Maxim G. Shishaev[3], Ivan G. Vishnyakov[4]
[1–4]Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
[1]v.pimeshkov@ksc.ru✉, https://orcid.org/0000-0001-7010-230X
[2]m.nikonorova@ksc.ru, https://orcid.org/0000-0002-4358-2683
[3]m.shishaev@ksc.ru, https://orcid.org/0000-0001-7070-7878
[4]i.vishnyakov@ksc.ru, https://orcid.org/0009-0003-4938-5693

Abstract

The paper considers the problem of forming multilayer knowledge graphs for use in intelligent analysis of social media data. The technology is based on the «soft» structuring of the existing document pool in accordance with the constructed topic model. Experimental testing of the technology and comparison of the formal properties of knowledge graph models obtained using topic modeling and the traditional method were conducted. The results of the comparison indicate the potential effectiveness of using multilayer knowledge graphs built on the basis of a topic model of texts when implementing intelligent data processing procedures in the conditions of multiple and dynamic subject areas.

Keywords:

knowledge graph, topic model, social media

Acknowledgments:
the study was carried out within the framework of the Putilov Institute for Informatics and Mathematical Modeling of the Kola Science Centre of the Russian Academy of Sciences state assignment of the Ministry of Science and Higher Education of the Russian Federation, research topic “Methodology for creating information and analytical systems to support the management of regional development based on formative artificial intelligence and big data” (registration number of the research topic 122022800551-0).
For citation:
Pimeshkov V. K., Nikonorova M. L., Shishaev M. G., Vishnyakov I. G. Formation of multilayer knowledge graphs based on thematic texts modeling // Transactions of the Kola Science Centre of RAS. Series: Engineering Sciences. 2024. Vol. 15, No. 3. P. 50–60. doi:10.37614/2949.1215.2024.15.3.004.

Введение

Графы знаний (ГЗ) являются эффективным способом формализованного представления знаний, подразумевающих использование в автоматизированных интеллектуальных информационных системах различного назначения. Привлекательной стороной ГЗ является тот факт, что аккумулируемые и систематизируемые в их рамках знания извлекаются из источников данных, в противовес общепринятому подходу к построению онтологий, где система знаний формируется априорно экспертами и затем используется при формировании и обработке наборов данных. То есть в случае ГЗ данные играют роль источника эталонных структур понятий и отношений в системе знаний. Это обусловливает высокий уровень актуальности содержащихся в ГЗ знаний и их адаптацию к динамике пользовательских представлений о предметной области.

Проблемным местом ГЗ является то, что при извлечении знаний из данных формируется мультипредметная система знаний, охватывающая в общем случае множество предметных областей. Это порождает проблему множественности возможных (в том числе противоречивых) интерпретаций знаний, а также усложняет техническую задачу оперирования системой знаний в силу ее большого объема. Решением этой проблемы является построение тематизированных (предметно-ориентированных) ГЗ, однако это, в свою очередь, сужает область применимости формируемого ГЗ, что является неприемлемым в некоторых прикладных задачах. Компромиссом между специфичностью и универсальностью ГЗ является ГЗ, имеющий внутреннее разбиение на подграфы, ориентированные на ту или иную предметную область (в общем случае — на специфику интерпретации знаний, требуемую для эффективного решения прикладных задач в предполагаемой сфере применения ГЗ). Такие ГЗ именуются многослойными.

В настоящей работе рассматривается проблематика построения многослойных ГЗ в контексте прикладной задачи мониторинга социальных медиа с целью анализа контента информационных коммуникаций пользователей. Эта задача является характерным примером проблемной области, описываемой множественностью и динамичностью систем семантических понятий и отношений, используемых пользователями. Для автоматизации и интеллектуализации обработки данных в таких условиях требуются адекватные инструменты. Главной идеей работы является использование тематической модели корпуса документов в качестве основы для секционирования ГЗ в соответствии с наиболее адекватными коммуницирующим сообществам стереотипными интерпретациями понятий из различных предметных областей. Цель исследования — оценка эффекта от использования тематического моделирования при построении многослойного ГЗ в сравнении с ГЗ, построенным традиционным способом.

Графы знаний как инструмент интеллектуальной обработки данных предметной области

ГЗ описывает концепты или сущности реального мира и отношения между ними способом, близким к представлению (мышлению) человека, и в виде, пригодном для компьютерной обработки. Формально ГЗ можно представить в виде [1]:

$$G \coloneqq (V, E, r, \Sigma_V, \Sigma_E, l_V, l_E),$$

где V — множество вершин, выраженное уникальными концептами или терминами; E — множество ребер, заданное выделенными между концептами отношениями; $r: E \to \{\{x, y\}: x, y \in V\}$ — функция, присваивающая каждому ребру неупорядоченную пару вершин; Σ_V — множество меток вершин; Σ_E — множество меток ребер; $l_V: V \to \Sigma_V$ — функция, определяющая метки вершин; $l_E: E \to \Sigma_E$ — функция, определяющая метки ребер.

Иногда о ГЗ говорят как о семантических сетях или сетях знаний, извлеченных из данных реального мира. В настоящее время ГЗ используются в задачах поиска информации, вопросно-ответных системах, системах поддержки принятия решений, рекомендательных системах [2]. ГЗ, как и другие модели знаний, могут быть двух типов — общего назначения и специфичные для предметной области. Первый тип, как правило, представляет собой довольно обширные графы, содержащие знания из множества предметных областей (например, Wikidata, DBpedia и др.), в то время как второй тип ориентирован на более узкую область или отрасль или на решение конкретной задачи.

ГЗ неразрывно связаны с технологией их построения и использования, включающей этапы извлечения знаний (knowledge extraction), слияния знаний (knowledge fusion), обработки знаний (knowledge processing) [3].

Извлечение знаний — первостепенная и основная задача при построении ГЗ, в рамках которой из необработанных данных извлекаются различные сущности, атрибуты и отношения. Слияние знаний предполагает объединение одинаковых объектов из различных источников для получения более точной и согласованной информации, что способствует поддержанию ГЗ в актуальном состоянии. Данный этап включает в себя подзадачу согласования объектов (entity alignment) для обнаружения семантически одинаковых объектов в разных источниках и подзадачу устранения неоднозначности объектов (entity disambiguation) для сопоставления объектов из входных данных с соответствующими уникальными объектами в целевом ГЗ. Обработка знаний предполагает обработку простых фактов и формирование структурированных систем знаний с данными высокого качества. Обычно включает в себя построение онтологий, оценку качества и иногда логический вывод.

Говоря о применении ГЗ для решения различных задач, так или иначе связанных с анализом данных социальными медиа, можно выделить следующие работы. Авторы работы [4] представили ГЗ, построенный на основе данных социальных сетей, для обнаружения фейковых новостей. В основе графа лежит онтология Fandet, предназначенная для представления сложных и часто неполных данных социальных сетей, а также облегчения их анализа. В работе [5] авторы используют ГЗ в задаче деанонимизации группы людей. Они моделируют начальные знания злоумышленника при помощи ГЗ, и в дальнейшем используют его для моделирования двух этапов атаки — деанонимизации и выведение конфиденциальной информации (privacy inference). Такая модель позволяет лучше описать процесс атаки и количественно оценить степень раскрытия конфиденциальной информации. Авторы исследования [6] строят ГЗ на основе новостных статей, посвященных COVID-19. Такой граф обеспечивает инфраструктуру для анализа данных социальных медиа, связанных с COVID-19, которая может быть полезна для исследователей, специалистов по обработке данных и организаций. В работе [7] предлагают архитектуру и прототип платформы, включающей ГЗ для поддержки работы журналистов некоторой исходной информацией в рамках вычислительной журналистики.

Тематические ГЗ, в отличии от обычных, как правило, отражают конкретную узкую тематику, например, аккумуляторы электромобиля, и содержат подробные знания предметной области, включая специализированные сущности и тройки. Авторами работы [8] предлагается фреймворк для автоматизированного построения таких «тематизированных» (theme-specific) ГЗ. Данный фреймворк принимает на вход необработанный тематический корпус и создает ГЗ, который включает в себя значимые сущности и отношения между ними в рамках заданной темы. Создание графа начинается с построения онтологии сущностей темы из «Википедии», на основе которой затем генерируется отношения-кандидаты с помощью больших языковых моделей (LLM) для построения онтологии отношений. Далее, чтобы проанализировать документы из тематического корпуса, авторы сначала сопоставляют извлеченные пары сущностей с онтологией и извлекают отношения-кандидаты. Наконец, авторы учитывают контекст употребления сущностей совместно с онтологией отношений для окончательного определения отношений между сущностями и, соответственно, построения ГЗ.

По мере появления более сложных данных и, соответственно, необходимости разработки более сложных систем знаний, свое развитие получила концепция многослойных ГЗ. Ключевой особенностью таких графов является представление знаний с помощью нескольких слоев, уровней или измерений. Каждый такой слой в графе может представлять различные типы отношений между

сущностями. Например, один слой может представлять социальные связи, другой — профессиональные отношения, а третий — взаимодействия в определенном контексте. Известным примером многослойного ГЗ является KnowWhereGraph [9]. На текущий момент данный граф содержит около 13 миллиардов триплетов и более 30 слоев, включающих пространственные данные (места, регионы), данные о населении, об экстремальных явлениях, инфраструктуре, сельскохозяйственные данные и др. Таким образом, он обеспечивает достаточную полноту и актуальность данных, необходимых для решения различных прикладных задач. В этой работе, в отличие от KnowWhereGraph, многослойность отражает разные контексты совместной встречаемости терминов. Предполагается, что такие контекстные слои зависят от характера исходных данных и решаемой задачи.

Технология формирования многослойного графа знаний на основе тематического моделирования

В рамках настоящей работы рассматривается разбиение на контексты в соответствии с тематическим распределением документов, полученным на основе тематической модели. Такое распределение позволяет посмотреть на данные с точки зрения некоторых скрытых тем, выделенных в наборе данных. Вершинами полученного графа будут являться термины, а отношения между ними будут выражаться через взвешенную направленную ассоциативную связь, полученную на основе статистики совместного употребления данных терминов, принимая во внимание результаты тематического моделирования. Предполагается, что разбиение графа на контексты, которые могут рассматриваться как слои графа или подграфы, на основе тематического распределения документов позволит выделять значимые связи, даже невзирая на малый вес темы в наборе данных. Справедливо заметить, что эффективность такого подхода будет в том числе зависеть от свойств и качества полученной тематической модели.

Формально предлагаемую технологию, использующую тематическую модель в качестве основы для разбиения на контексты (рис. 1), можно представить следующим образом.

Дано: D — набор документов различной тематики; $L \subseteq D \times D$ — асимметричное транзитивное отношение «является откликом», определяет на множестве документов структуру коммуникации. $d_1 L d_2$ означает, что документ (пост или комментарий) d_2 является откликом на документ d_1. Структура коммуникации состоит из веток обсуждения, представляющих собой последовательности Br документов из D ($Br \subseteq D$), удовлетворяющие условию: $Br = \{d_1, \dots, d_N\}: \forall i < j, d_i L d_j$.

Полная структура коммуникации (полное дерево) задается корневым документом d и всеми транзитивно связанными с ним документами $Br(d)$.

Найти:

$$G \coloneqq (V, E, K, W_K, w_E, s, t, r),$$

где V — множество вершин — уникальные термины из набора документов D; E — множество ребер — взвешенные направленные ассоциативные отношения между терминами; K — множество меток контекстов, заданное темами тематической модели; W_K — множество весов ребер, заданное для каждого контекста; $w_E: E \to W_K$ — функция, присваивающая вес каждому ребру; $s: E \to V$ — функция, присваивающая каждому ребру начальную вершину; $t: E \to V$ — функция, присваивающая каждому ребру конечную вершину; $r: E \to K$ — функция, присваивающая каждому ребру метку контекста.

Первым этапом извлекаются термины из D следующим образом: $TermExtrMethod(D, L, \dots) \to Br^T$ — структура коммуникации Br с выделенными в ней терминами T.

Вторым этапом строится тематическая модель на основе Br^T и заданного количества тем K:

$$TopicModelConst(Br^T, K) \to \{\Theta, \Phi\},$$

где Θ — матрица, задающая распределение документов по темам; Φ — матрица, задающая распределение терминов по темам.

Третьим этапом производится построение и заполнение тематизированных матриц совместного употребления терминов: $f(Br^P, \Theta) \to CM$ — набор матриц совместного употребления, где элемент матрицы совместного употребления по k-й теме вычисляется по формуле:

$$cm_{ij}^{k} = \sum_{Br^P=1}^{M} \theta_{Br^P}^{k} \cdot Br_{ij}^{P} \cdot \left(\frac{\phi_i^k}{\max(\phi_i)} + \frac{\phi_j^k}{\max(\phi_j)}\right), \quad (1)$$

где $\theta_{Br^P}^{k}$ — значение принадлежности структуры Br^P теме k; Br_{ij}^{p} — значение совместного употребления терминов i и j в рамках Br, получаемое функцией подсчета совместного употребления терминов в структуре коммуникации $fp(Br^T) \rightarrow Br^P$ — вида $<i, j, c_{ij}>$, где i, j — термины, c_{ij} — значение совместного употребления данных терминов в конкретной структуре Br; ϕ_i^k — значение принадлежности термина i теме k; $\max(\phi_i)$ — максимальное значение принадлежности термина i теме.

В результате получается набор матриц со значениями совместной встречаемости соответствующих терминов. Набор уникальных терминов интерпретируется как вершины графа, а ненулевые значения матриц совместной встречаемости — как взвешенные ассоциативные отношения между ними.

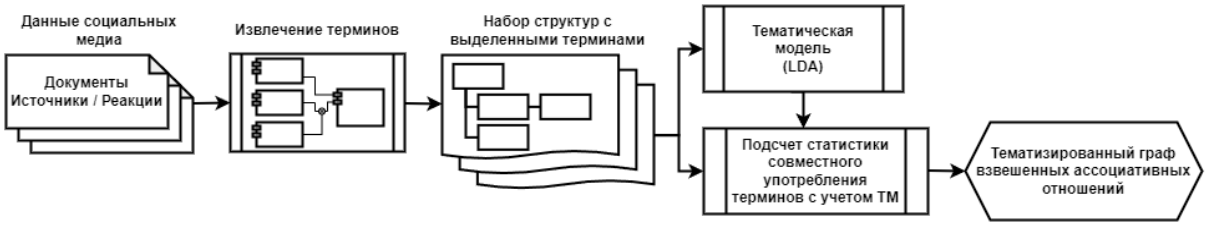


Рис. 1. Технология формирования многослойного ГЗ на основе тематического моделирования

Оценка эффективности применения тематического моделирования для формирования ГЗ

Для проверки предлагаемой технологии формирования многослойного ГЗ был использован датасет публикаций с комментариями социальной сети «ВКонтакте» из 8 групп за приблизительно 2 года. Исходный датасет содержал 52 375 публикации и 235 365 комментариев, из которых содержат текст 50 067 и 221 834 соответственно.

Очистка текста в датасете от нежелательных символов таких, как эмодзи, повторения различных знаков и конструкций, например, телефонных номеров, ссылок или идентификаторов пользователей, проводилась в два этапа. На первом этапе для очистки от нежелательных конструкций использовался набор регулярных выражений. На втором — белый список символов, состоящий из кириллицы, латиницы, цифр и ряда специальных символов.

Для дальнейшей обработки датасет был преобразован в набор структур коммуникации, где каждая структура является представлением публикации и ее откликов (комментариев). Вместе с этим преобразованием проводилось выделение терминов в этих структурах с помощью авторского комбинированного метода извлечения терминов [10]. В результате работы метода в документах были выделены термины, входящие в заранее заданный словарь, именованные сущности, осмысленные биграммы и триграммы, а также контекстно важные униграммы-существительные. Полученные структуры дополнительно были очищены от редко встречающихся терминов. В результате был получен набор из 45 363 структур, содержащих хотя бы два уникальных термина, а также список уникальных терминов датасета. Для обучения тематической модели и дальнейшего ее использования при построении матриц совместного употребления был выполнен препроцессинг терминов: приведение в нижний регистр, стеммминг (для приведения в единую словоформу), замена пробелов на нижние подчеркивания (в многословных терминах).

На основе полученных документов с извлеченными терминами был обучен набор из 11 тематических моделей (ТМ) (количество тем варьировалось от 5 до 15) на основе латентного размещения Дирихле (Latent Dirichlet allocation, LDA), реализованной в библиотеке Gensim [11]. При обучении всех моделей использовались следующие параметры: update_every = 1, chunksize = 1000, passes = 6. Для полученных моделей была вычислена метрика согласованности (coherence), которая показывает, насколько выявленные темы значимы и интерпретируемы. Результаты подсчета метрики приведены на рис. 2. Таким образом, в дальнейшей работе было принято решение использовать модель с 7 темами, т. к. она имеет наибольший показатель согласованности тем.

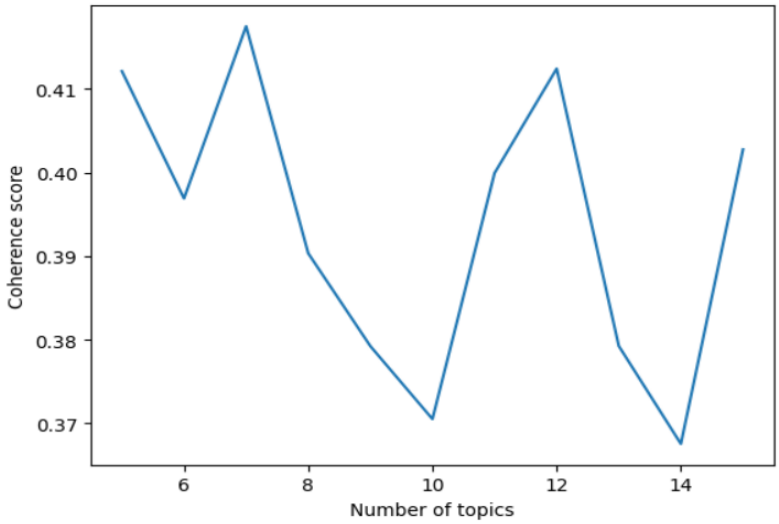
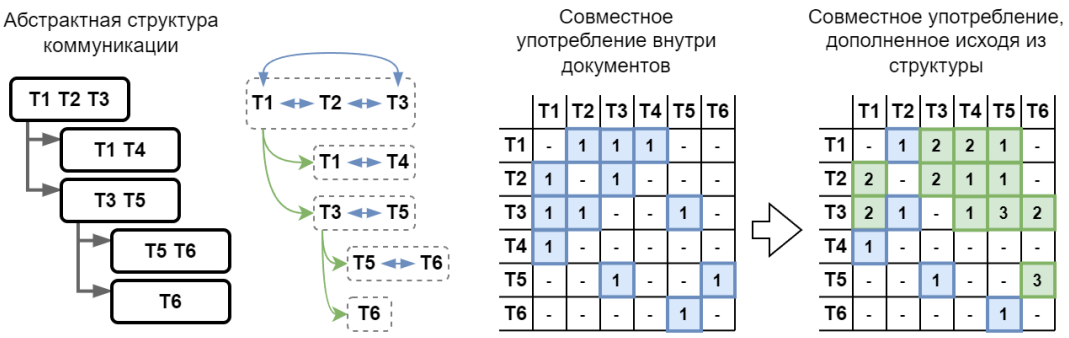


Рис. 2. Изменение метрики согласованности в зависимости от количества тем

Далее на основе набора структур и выбранной тематической модели был произведен подсчет статистики совместного употребления терминов. Значения совместного употребления рассчитывались отдельно для каждой структуры методом, в рамках которого предполагалось, что, во-первых, термины связаны между собой в рамках каждого документа, а, во-вторых, термины источника приводят к появлению термина в реакции и, соответственно, термины источника связаны с каждым термином в реакциях. При этом пары идентичных терминов игнорировались для исключения петель в графе. Пример расчета совместного употребления таким методом приведен на рис. 3. Далее значения совместного употребления (Br_{ij}^{P}), в соответствии с формулой (1), умножались на соответствующий компонент вектора распределения документа по темам ($\theta_{Br^P}^{k}$) и на сумму коэффициентов влияния терминов. Полученные значения фиксировались в наборе матриц размером $N \times N$, где N — это число уникальных терминов в датасете (13 171), в ячейке, соответствующей текущей теме и терминам. Таким образом, мы получили набор матриц, в котором каждая матрица отражает определенный слой графа, соответствующий одной из тем, выделенных тематической моделью.



	T1	T2	T3	T4	T5	T6
T1	-	1	1	1	-	-
T2	1	-	1	-	-	-
T3	1	1	-	-	1	-
T4	1	-	-	-	-	-
T5	-	-	1	-	-	1
T6	-	-	-	-	1	-

	T1	T2	T3	T4	T5	T6
T1	-	1	2	2	1	-
T2	2	-	2	1	1	-
T3	2	1	-	1	3	2
T4	1	-	-	-	-	-
T5	-	-	1	-	-	3
T6	-	-	-	-	1	-

Рис. 3. Пример работы метода подсчета совместного употребления терминов

Для того, чтобы оценить эффект от разбиения графа на темы, был построен аналогичный граф, но без применения тематической модели. Он строился без учета распределения документов по темам, соответственно, в ячейках матрицы фиксировались просто значения совместного употребления терминов без дополнительных коэффициентов.

В обоих случаях, с применением ТМ и без, было замечено, что в матрицах преобладают значения совместного употребления близкие к 0. Соответствующие отношения, со значениями близкими к нулю, говорят о том, что совместное употребление значительного количества терминов наблюдалось всего лишь один или пару раз во всем датасете, следовательно, эта связь не является значимой в рамках рассматриваемой модели знаний. Поэтому было принято решение исключить из матриц такие незначимые связи путем выделения значимых связей.

Для выделения значимых связей из тематизированной матрицы на первом этапе для каждой темы был выбран процент самых сильных связей, значение которого соответствует доле терминов в конкретной теме (значения получены на этапе обучения тематической модели). На втором этапе была проведена оценка значимости в соответствии с эвристически выведенным порогом значимости:

$$Br_k^P(i,j) > \bar{x}\left(\bar{x}\left(Br_k^P(i,)\right), \bar{x}\left(Br_k^P(,j)\right), \bar{x}\left(Br_k^P\right)\right),$$

где $\bar{x}\left(Br_k^P(i,)\right)$ — среднее значение совместного употребления для термина i; $\bar{x}\left(Br_k^P(,j)\right)$ — среднее значение совместного употребления для термина j; $\bar{x}\left(Br_k^P\right)$ — среднее значение совместного употребления для всех терминов в соответствующей матрице (теме) k.

При этом среднее значение совместного употребления для термина рассчитывалось как среднее суммы векторов входящих и исходящих отношений. Для матрицы без ТМ применялась только фильтрация по порогу значимости. В результате такой фильтрации отношений их количество значительно сократилось — с 754 993 до значений, указанных в табл. 1. Там же отражено изменение количества уникальных терминов с хотя бы одной ненулевой связью для каждой матрицы соответственно. До фильтрации в каждой матрице было 11 889 таких терминов.

Таблица 1

Количество отношений и терминов в полученных матрицах после фильтрации

Матрица	Количество отношений	Количество терминов
Без ТМ	117716	6592
Тема № 1	7475	520
Тема № 2	25215	2536
Тема № 3	14555	1875
Тема № 4	16356	1912
Тема № 5	27987	2832
Тема № 6	12004	1390
Тема № 7	9827	968

Стоит заметить, что без фильтрации вовсе мы могли получить практически полносвязный граф, что не позволит эффективно отслеживать какие-либо связи в такой структуре. С другой стороны, при слишком строгой фильтрации мы однозначно рискуем удалить нужные связи и, соответственно, термины. В данном случае способ фильтрации полагался на специфику анализируемых данных, поэтому оценка его эффективности остается открытой задачей.

По факту наличия отношений между терминами полученные тематические матрицы в значительной степени пересекаются с матрицей без ТМ (рис. 4). Это говорит о том, что после удаления незначимых отношений предложенным порогом фильтрации матрицы все еще близки по структуре, и большая часть связей, присутствующих в тематизированном наборе матриц, также присутствуют и в матрице без ТМ. При этом в некоторых из них все же есть отношения, которые были удалены при фильтрации из матрицы без ТМ. Этот факт, возможно, подтверждает гипотезу о том, что тематизированные ГЗ способствуют выделению некоторых связей на общем фоне.

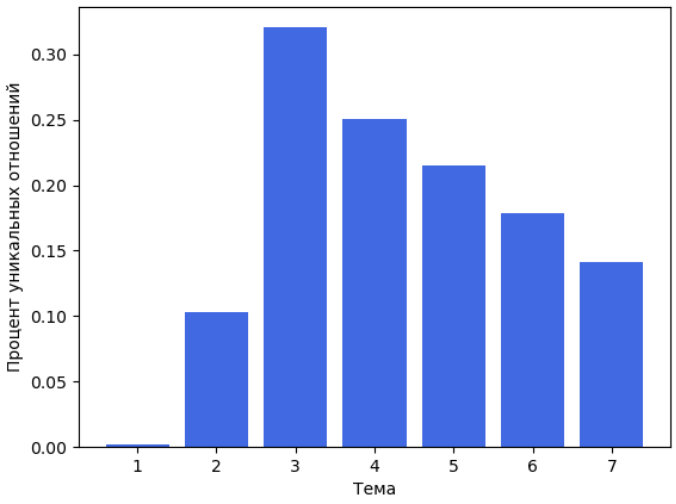


Рис 4. Процент уникальных связей относительно матрицы без тематической модели

Полученные тематизированные матрицы в среднем по факту наличия связи пересекаются не более чем на 5–6 %, а, значит, вероятнее всего, отражают разные ассоциации пользователей. Построенная матрица корреляций (рис. 5) показывает процент общих связей между тематическими слоями матрицы.

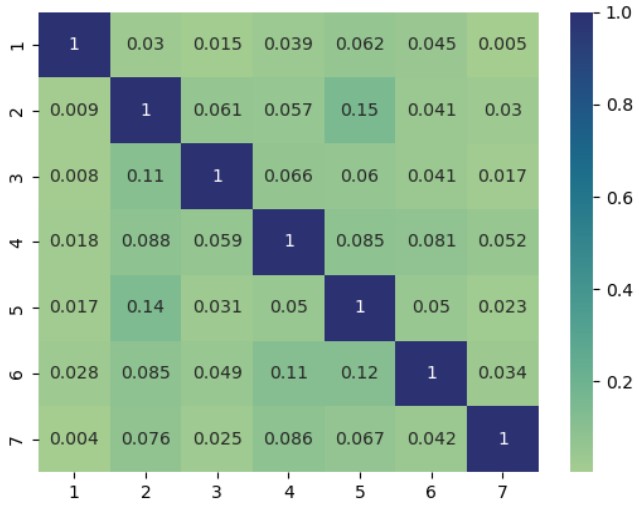


Рис. 5. Пересечение отношений между темами (относительно строки)

Полученные матрицы далее были преобразованы в направленные мультиграфы с помощью библиотеки NetworkX [12] следующим образом: вершинами стали уникальные термины, у которых имеется хотя бы одна связь, а ребрами — ненулевые значения матриц по слоям соответственно. Следовательно, данный набор мультиграфов представляет собой тематизированный или многослойный ГЗ.

Для оценки полученных мультиграфов были рассчитаны взвешенные степени их вершин и плотности с помощью методов, реализованных в упомянутой ранее библиотеке. Степень вершины соответствует количеству ребер, инцидентных данной вершине, а взвешенная степень вершины — это сумма весов ребер, инцидентных данной вершине. В табл. 2 по каждому мультиграфу, с применением ТМ и без, представлены топ-5 вершин по их взвешенной степени, на основе которых можно предположить об основных терминах, вокруг которых ведется дискурс в социальной сети в рамках определенной темы.

Таблица 2

Топ-5 вершин по их взвешенной степени

Без ТМ	Апатиты	Мурманск	Кировск	Город	Мурманская_область
	259 387	133 795	131 710	97 951	96 640
Слой 1	Средств	Действ	Антибиотик	Препарат	Фильм
	36 901,47	34 609,02	22 508,20	21 022,29	19 672,85
Слой 2	Апатиты	Город	Кировск	Кола	Улиц
	216 725,28	85 487,75	85 300,10	51 679,57	35 160,40
Слой 3	Квартир	Ремонт	Дом	Комнатн_квартир	Апатиты
	29 639,01	16 936,36	11 908,89	11 181,37	10 144,00
Слой 4	Работ	Светл_памя	Кировск	Апатиты	Вечн_памя
	16 446,07	3 964,45	3 423,50	3 339,41	3 090,55
Слой 5	Мурманск	Мурманская_область	Област	Росс	Мест
	82 211,37	63 065,60	54 506,15	22 862,24	18 227,96
Слой 6	Покупател	Препарат	Работник	Прав	Организац
	3 083,60	2 959,94	2 699,43	2 543,58	2 202,78
Слой 7	Имандра	Озер	Магазин	Водоем	Приток
	25 387,04	18 612,45	10 106,22	8 140,36	6 797,95

Плотность графа — это отношение числа ребер к максимально возможному. На основе вычисленных плотностей (рис. 6), где максимальное значение соответствует всего лишь 0,013, можно предположить, что полученные графы представляют собой леса.

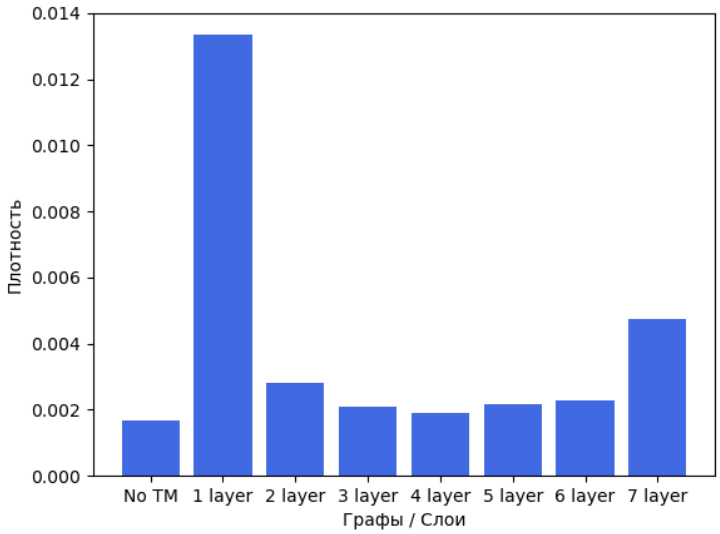


Рис. 6. Показатели плотности полученных графов

Заключение

В результате проделанной работы была разработана и опробована на практике технология формирования многослойных ГЗ на основе данных социальных медиа. В качестве базы для экспериментов использовался датасет публикаций с комментариями социальной сети «ВКонтакте», включающий 52 375 публикации и 235 365 комментариев.

Сравнение ГЗ, полученных по предложенной методике с ГЗ, сформированными традиционным способом, показывает, что в первом случае знания имеют лучшую структуризацию. Это создает предпосылки для реализации более эффективных процедур обработки данных социальных медиа с точки зрения объемов вычислений и корректности результатов.

Список источников

1. Baclawski K. et al. Ontology summit 2020 communiqué: Knowledge graphs // Applied Ontology. 2021. Vol. 16, № 2. P. 229–247.
2. Zou X. A Survey on Application of Knowledge Graph // J. Phys.: Conf. Ser. IOP Publishing, 2020. Vol. 1487, № 1. P. 012016.
3. Tian L. et al. Knowledge graph and knowledge reasoning: A systematic review // Journal of Electronic Science and Technology. 2022. Vol. 20, № 2. P. 100159.
4. Hani A.B. et al. Fane-KG: A Semantic Knowledge Graph for Context-Based Fake News Detection on Social Media // 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS). 2020. P. 1–6.
5. Qian J. et al. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model // IEEE Transactions on Dependable and Secure Computing. 2019. Vol. 16, № 4. P. 679–692.
6. Al-Obeidat F. et al. Cone-KG: A Semantic Knowledge Graph with News Content and Social Context for Studying Covid-19 News Articles on Social Media // 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS). 2020. P. 1–7.
7. Berven A. et al. A knowledge-graph platform for newsrooms // Computers in Industry. 2020. Vol. 123. P. 103321.
8. Ding L. et al. Automated Construction of Theme-specific Knowledge Graphs: arXiv:2404.19146. arXiv, 2024.
9. Janowicz K. et al. Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence // AI Magazine. 2022. Vol. 43, № 1. P. 30–39.
10. Пимешков В. К., Никонорова М. Л., Шишаев М. Г. Комбинированный метод извлечения терминов для задачи мониторинга тематических обсуждений в социальных медиа // Информатика и автоматизация. 2024. Т. 23, № 4. С. 1110–1138.
11. Gensim: Topic modelling for humans [Электронный ресурс]. URL: https://radimrehurek.com/gensim/ (дата обращения: 05.10.2024).
12. NetworkX [Электронный ресурс]. URL: https://github.com/networkx/networkx (дата обращения: 05.10.2024).

References

1. Baclawski K. et al. Ontology summit 2020 communiqué: Knowledge graphs. *Applied Ontology*, 2021, vol. 16, no. 2, pp. 229–247.
2. Zou X. A Survey on Application of Knowledge Graph. *Journal of Physics: Conference Series*, 2020, vol. 1487, no 1, p. 012016.
3. Tian L. et al. Knowledge graph and knowledge reasoning: A systematic review. *Journal of Electronic Science and Technology*, 2022, vol. 20, no. 2, p. 100159.
4. Hani A.B. et al. Fane-KG: A Semantic Knowledge Graph for Context-Based Fake News Detection on Social Media. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2020, pp. 1–6.
5. Qian J. et al. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. *IEEE Transactions on Dependable and Secure Computing*, 2019, vol. 16, no. 4, pp. 679–692.
6. Al-Obeidat F. et al. Cone-KG: A Semantic Knowledge Graph with News Content and Social Context for Studying Covid-19 News Articles on Social Media. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2020, pp. 1–7.
7. Berven A. et al. A knowledge-graph platform for newsrooms. *Computers in Industry*, 2020, vol. 123, pp. 103321.
8. Ding L. et al. Automated Construction of Theme-specific Knowledge Graphs. arXiv:2404.19146, 2024.
9. Janowicz K. et al. Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence. *AI Magazine*, 2022, vol. 43, no 1, pp. 30–39.

10. Pimeshkov V.K., Nikonorova M.L., Shishaev M.G. Kombinirovannyj metod izvlecheniya terminov dlya zadachi monitoringa tematicheskih obsuzhdenij v social'nyh media [A combined term extraction method for the problem of monitoring thematic discussions in social media]. *Informatika i Avtomatizaciya* [Informatics and Automation], 2024, vol. 23, no. 4, pp. 1110–1138.
11. Gensim: Topic modelling for humans. Available at: https://radimrehurek.com/gensim/ (accessed 05.10.2024).
12. NetworkX. Available at: https://github.com/networkx/networkx (accessed 05.10.2024).

Информация об авторах

В. К. Пимешков — аспирант, стажер-исследователь;
М. Л. Никонорова — аспирант, инженер-исследователь;
М. Г. Шишаев — доктор технических наук, главный научный сотрудник;
И. Г. Вишняков — аспирант, системный администратор.

Information about the authors

V. K. Pimeshkov — PhD student, Research Assistan;
M. L. Nikonorova — PhD student, Research Engineer;
M. G. Shishaev — Doctor of Science (Tech.), Chief Research Fellow;
I. G. Vishnyakov — PhD student, System Administrator.

Статья поступила в редакцию 14.10.2024; одобрена после рецензирования 28.10.2024; принята к публикации 06.11.2024.
The article was submitted 14.10.2024; approved after reviewing 28.10.2024; accepted for publication 06.11.2024.