УДК 004.82 **DOI** 10.15622/ia.25.3.6

В.К. Пимешков, М.Л. Никонорова, М.Г. Шишаев

ИНФОРМАЦИОННЫЙ ПОИСК В СОЦИАЛЬНЫХ МЕДИА НА БАЗЕ МУЛЬТИДОМЕННОЙ ДИНАМИЧЕСКОЙ СИСТЕМЫ ЗНАНИЙ

Пимешков В.К., Никонорова М.Л., Шишаев М.Г. **Информационный поиск в социальных медиа на базе мультидоменной динамической системы знаний.**

Аннотация. Задача информационного поиска заключается в нахождении информации, наилучшим образом удовлетворяющей информационную потребность пользователя. В условиях социальных медиа информационный поиск осложняется высокой динамичностью контента, тематической разнородностью и разнообразием ментальных моделей пользователей. В данной работе предлагается подход к решению задачи информационного поиска в таких условиях путем построения мультидоменной динамической системы знаний. Ее новизна заключается в объединении трех уровней семантики: проблемно-ориентированной, представленной онтологией метазадачи (описывающей цели поиска); доменно-специфической, реализованной через динамический многослойный граф знаний, построенный на основе пользовательского контента социальных медиа; и доменно-независимой, основанной на лексической базе данных и большой языковой модели. Граф знаний позволяет отразить различные контексты употребления понятий, соответствующие тематическим кластерам в коллекции документов. Такая интеграция позволяет учитывать эволюцию понятий, особенности дискурса и ментальные стереотипы участников коммуникации. Для оценки эффективности предложенной системы проведен эксперимент с использованием датасета публикаций из социальной сети “ВКонтакте” по проблемно-ориентированному мониторингу публикаций, где требуется отбор релевантных публикаций из нетематизированных источников. Для решения данной задачи предложена технология, основанная на использовании метрики расстояния между терминами запроса и терминами публикации в многослойном графе знаний. Результаты эксперимента с применением данной технологии подтверждают эффективность предложенной модели для задач информационного поиска по сравнению со стандартным поиском по ключевым словам и эмбеддинговыми моделями. В продолжении данного исследования планируется сформировать лексическую базу данных, а также рассмотреть возможность расширения модели за счет применения меры точечной совместной информации и методов векторного представления графов.

Ключевые слова: информационный поиск, социальные медиа, системы знаний, многослойный граф знаний, релевантность.

1. Введение. Информационный поиск (ИП) – процесс поиска информации, удовлетворяющей информационной потребности пользователя, является одной из главных составляющих современных процессов обработки данных в различных приложениях, в том числе работающих с данными социальных медиа. Общая схема ИП заключается в том, что пользователь адресует информационной системе запрос и получает в ответ на него некоторую выдачу, содержащую

запрошенную (с точки зрения системы) информацию. Объективным критерием успешности поиска является пертинентность сгенерированной системой выдачи. Пертинентность является трудноформализуемым параметром информационной выдачи, характеризующим разницу между содержимым последней и ожиданиями пользователя, сформулированными посредством запроса. Подчеркнем, что речь идет о разнице между информационными свойствами выдачи и именно ожиданиями, а не сформулированным запросом пользователя (в последнем случае говорят о релевантности выдачи запросу). Практика показывает, что оценка качества поиска при идентичных запросах и выдаче, но для разных пользователей, сформировавших запрос, может существенно отличаться.

Таким образом, для объективной оценки качества ИП нужно оценить пертинентность выдачи, а для адекватной оценки последней необходима модель знаний, описывающая ментальные стереотипы пользователя. С помощью такой модели можно учесть ожидания пользователя при оценке качества поиска. В случае социальных сетей модель стереотипов может описывать пользователя либо непосредственно, либо транзитивно через модель сообщества, частью которого он является. Заметим, что в качестве сообщества может рассматриваться аудитория некоторой тематической дискуссии, и тогда под пользователем будет подразумеваться не субъект, как таковой, а участник дискуссии, то есть текущая роль субъекта.

Существенная проблема заключается в том, что разные пользователи и сообщества характеризуются различными ментальными моделями (набор и способы интерпретации понятий), которые неоднозначны в том смысле, что способы интерпретации зависят от контекста использования, и которые, к тому же, способны меняться со временем. Контекст использования при этом определяется, в первую очередь, метазадачей поиска или, другими словами, прикладной задачей, решаемой пользователем в рамках взаимодействия с информационной системой. Построение исчерпывающе точных моделей знаний в таких условиях маловероятно. Однако ментальные стереотипы находят отражение в генерируемом пользователями социальных медиа контенте и коммуникативных актах, что создает возможность построения моделей, позволяющих учесть ожидания пользователя при формировании результатов поиска и обеспечить тем самым улучшение их качества с точки зрения пертинентности.

В данной работе в качестве такой модели, ориентированной на обеспечение приближения ИП к пертинентному результату,

рассматривается мультидоменная динамическая система знаний (МДСЗ). В предлагаемой модели проблема динамичности знаний решается за счет включения в модель графов знаний (ГЗ), формируемых и обновляемых на основе контента социальных медиа. То есть, в модели отражаются актуальные системы понятий и способы их интерпретации, оперируемые пользователями. Для решения проблемы разнородности знаний (неоднозначности интерпретаций понятий) ГЗ формируется в виде многослойной структуры, где каждый слой соответствует различным контекстам использования или доменам. В качестве связующего компонента в предлагаемой модели выступает онтология метазадачи (предметной области), задающая базовую, независимую от домена систему понятий. Терминологическая база при этом задается с помощью лексической базы данных (БД) и большой языковой модели. Таким образом, МДСЗ интегрирует в себе относительно стабильную проблемно-ориентированную семантику, представленную онтологией, динамичную доменно-специфическую семантику, представленную с помощью многослойного ГЗ, и доменно-независимую семантику, формируемую с помощью лексической БД и большой языковой модели.

Отметим, что в конечном итоге качество результатов поиска, по-прежнему, будет выражаться некоторой оценкой разницы между выдачей и поисковым запросом, поэтому далее в статье будет использоваться для обозначения этой разницы более традиционный термин “релевантность”. Однако предполагается, что в случае учета ментальных стереотипов автора запроса с помощью МДСЗ, такая оценка релевантности будет более близка к пертинентности, то есть к объективной оценке качества ИП. В этом заключается основная гипотеза представленного исследования.

Эффективность использования предложенной модели в ИП проверялась на примере задачи проблемно-ориентированного мониторинга социальных медиа, где требуется отбор релевантных публикаций из нетематизированных источников. Суть задачи состоит в том, чтобы отобрать из потока сообщений те (в данном случае – постов и комментариев текстового формата социальной сети “ВКонтакте”, опубликованных в пределах некоторого ограниченного набора пабликов), которые в наибольшей степени удовлетворяют целям мониторинга (метазадаче поиска). Для проведения эксперимента были использованы данные наблюдений за один год, в которых, изначально, в качестве критерия релевантности использовался простой факт вхождения в тело сообщения ключевых слов из заданного набора. Суть эксперимента заключалась в сравнении исходных результатов отбора с результатами отбора релевантных

сообщений с использованием МДСЗ. Для дополнительной верификации результатов авторского подхода к оценке релевантности использовались экспертно размеченные наборы данных.

Основной вклад (значимость) работы заключается в следующем:

- разработана концептуальная модель предметной области, отражающая мультидоменность и динамичность последней и позволяющая учесть специфику задачи ИП в контексте поиска данных в социальных медиа;
- предложен способ построения МДСЗ, интегрирующей доменно-специфичную, доменно-независимую (общую) и проблемно-ориентированную семантику;
- подтверждена эффективность предложенной модели на примере задачи мониторинга социально-политической обстановки на территории активности крупного предприятия, реализующего социально-ответственную политику;
- предложен подход к формированию многослойного ГЗ на базе тематического моделирования, позволяющий автоматизировать процесс построения мультидоменной модели знаний в условиях отсутствия явной тематизации источников данных;
- предложена технология оценки релевантности документов с помощью многослойного ГЗ, продемонстрировавшая эффективность на примере тестовой прикладной задачи мониторинга социально-политической обстановки на территории.

Наряду с рассмотренным здесь мониторингом социальных сетей, разработанные модели и технологии могут найти применение в задачах пополнения онтологий (пополнение начальной онтологии за счет ГЗ, полученного на данных, предположительно содержащих необходимые знания), а также в широком спектре задач поиска знаний (knowledge mining).

Дальнейший материал имеет следующую структуру. В разделе 2 представлено позиционирование данного исследования относительно современных подходов и направлений развития методологии ИП. В разделе 3 рассмотрены особенности задачи ИП в социальных медиа и формализована постановка данной задачи. В разделе 4 представлена модель МДСЗ, состоящая из проблемно-ориентированной онтологии, динамического многослойного ГЗ и лексической БД. В разделе 5 представлена технология определения релевантности публикаций с помощью многослойного ГЗ, в разделе 6 приведены результаты её экспериментального тестирования.

2. Обзор релевантных исследований в области технологий информационного поиска. Основной целью современного ИП является предоставление конечному пользователю максимально релевантной информации по его запросу. Этот процесс можно разделить на два этапа: поиск и ранжирование [1]. На первом этапе извлекается набор исходных документов, наиболее релевантных запросу, а на втором этапе происходит ранжирование полученных документов на основе их оценки релевантности. На этапе поиска могут использоваться как традиционные методы, например, логический (Boolean) поиск и TF-IDF, так и методы семантического поиска, например, расширение запроса, латентно-семантический анализ и тематическое моделирование, а также методы глубокого обучения и гибридные. На этапе ранжирования также применяются различные методы классического и глубокого машинного обучения, включая обучение с подкреплением, контекстуальные эмбеддинги и механизмы внимания, чтобы научиться ранжировать документы на основе критериев запроса пользователя. Одной из распространенных количественных оценок релевантности являются показатели точности (доля правильных результатов в их общем количестве) и полноты (доля правильных результатов, попавших в результирующую выдачу).

Рассматриваемую в данной работе технологию определения релевантности с помощью многослойного ГЗ можно отнести к методам семантического поиска, а именно к методам расширения запроса. Подобные методы модифицируют исходный поисковый запрос для улучшения точности и полноты результатов поиска [2, 3]. Такая модификация может включать добавление синонимов, использование нормализованных и словообразовательных вариантов терминов из запроса, добавление терминов, встречающихся в том же контексте и т.д. Данные методы, как правило, позволяют повысить качество и полноту поиска, но существует риск избыточного расширения, что приводит к большому числу ложноположительных результатов.

Примером работы, реализующей методы расширения запроса, является [4], где применяют контролируемые словари, онтологии и контекстно-свободные дистрибутивные языковые модели для расширения поискового запроса, исходящего от информационных работников. Для онтологических ресурсов применяется простой поиск связанных терминов в виде гиперонимов, гипонимов и т.д., а языковые модели используются для генерации запросов из связанных терминов.

Во многих системах ИП активно используются ГЗ. Они позволяют лучше понять информационную потребность пользователя, его запрос или расширить представления о документах на таком

уровне, который нельзя достичь с помощью одних лишь ключевых слов как таковых [5]. Дополнительно ГЗ открывают возможность сразу отвечать на запрос пользователя в рамках поиска сущностей (когда данный ответ заложен в понятиях и/или структуре графа в явном виде), а также дают возможность исследовать связанные сущности из запроса или документов, включая справочную информацию по ним или их связям [6].

С вовлечением в процессы обработки сложных, семантически неоднородных данных возникает потребность в структуризации знаний, представленных графом. Это осуществляется с помощью многослойных ГЗ, где их элементы разделены на несколько слоев или измерений, отражающих различные типы связей между сущностями (например, социальные, профессиональные или контекстуальные) или разные точки зрения на предметную область (структура понятий и связей на каждом слое различна). Известным примером многослойного ГЗ является KnowWhereGraph [7], включающий около 13 млрд. триплетов и более 30 слоев (пространственные, демографические, инфраструктурные, сельскохозяйственные данные и др.), что обеспечивает его применимость в широком спектре задач.

Таким образом, можно отметить, что основным подходом к улучшению качества ИП являются различные манипуляции с поисковым запросом (как правило – расширение), а ГЗ предоставляют технологическую базу для формирования в результате таких манипуляций поисковых запросов, в большей степени соответствующих ожиданиям пользователя. Это достигается за счет адекватного формального представления знаний о предметной области, которые используются в процессе генерации поискового запроса и его результата. При этом одним из способов повышения эффективности использования ГЗ в случае сложных неоднородных предметных областей является его разбиение на слои, каждый из которых представляет собой специфическую точку зрения на предметную область. В данной работе развивается подход к ИП с использованием ГЗ в направлении повышения адекватности структуры знаний, представленной многослойным графом, за счет применения тематического моделирования. Использование тематического моделирования позволяет структурировать знания в ситуации, когда нет априорной тематизации источников данных, на базе которых строится ГЗ.

Еще одним способом улучшения качества поиска является использование дополнительной, внешней по отношению к запросу, информации при формировании результатов ИП. Примером такой

информации является структура коммуникаций и различные метаданные, представленные в социальных медиа. Область исследований, в рамках которой для улучшения работы систем поиска используются также показатели взаимодействия пользователей с контентом (лайки, просмотры), поведение пользователей и знания сообществ, называется “социальный информационный поиск” (Social Information Retrieval) [8]. Примером работы в данной области является [9], где авторы применяют технологию расширения запроса, включающую информацию о пользователях для улучшения работы ИП. В начале из документов выбираются важные термины, которые затем кластеризуются с использованием k-means. На первом этапе расширения запроса выбирается самый похожий на запрос кластер, который расширяется терминами из этого кластера. Затем оцениваются комментарии пользователей к документам в выбранном кластере и эти оценки используются для последующего расширения запроса.

Таким образом, существующие работы по проблематике социального информационного поиска показывают, что учет дополнительной информации, доступной при ИП в социальных медиа, позволяет повысить качество результата. В данной работе для повышения эффективности ИП использовалась информация о структуре коммуникаций (причинно-следственная структура публикаций и комментариев) в социальных медиа, представленная в виде соответствующей формальной модели.

Можно также отметить, что в целом современное развитие методов ИП фокусируется на оперировании семантикой запросов и поисковых выдач, как альтернативе более примитивным интерпретациям запроса и его результата как последовательностей символов. Важной проблемой, возникающей при реализации семантического поиска в социальных медиа, является разнородность используемых понятийных систем, представленных предметными (а в ряде случаев – мультипредметными) и общими знаниями, а также соответствующей терминологией. Для решения этой проблемы широко используются модели машинного обучения. Например, в работе [10] применяется глубокое обучение для идентификации разговорных выражений, описывающих симптомы COVID-19, в тексте постов Twitter. Затем они нормализуются и сопоставляются с внешним источником данных для получения словаря, который может использоваться врачами и неспециалистами для лучшего понимания того, как пользователи описывают свои симптомы в социальных медиа.

В работе использовались переобученные модели глубокого обучения для устранения лексической многозначности, характерной для коммуникаций в социальных сетях, при этом устанавливается соответствие между терминами общей лексики и терминами, используемыми в различных доменах.

3. Особенности и формальная постановка задачи информационного поиска в социальных медиа. Решаемая метазадача ИП в социальных медиа задает границы предметной области и таким образом определяет некоторый стабильный домен (задающий способы интерпретации понятий и связей между ними), в рамках которого осуществляется поиск и интерпретация результатов.

В то же время коммуникации в социальных медиа, как правило, охватывают множество доменов. Одно и то же понятие может встречаться в двух доменах, но иметь различные ассоциативные связи (различную семантику). Например, “вакцинация” в контексте COVID-19 будет связана с одними понятиями, а в контексте ветеринарии – с другими. Какие-то домены при этом существуют практически постоянно в социальном дискурсе без особых изменений, другие могут появляться, исчезать и видоизменяться достаточно быстро вместе с тем, как разгорается и угасает интерес к социальной дискуссии на какую-либо тему. Таким образом, домены, по существу, являются отражением ментальных моделей сообществ и пользователей социальных медиа.

Исходя из этого, модель предметной области должна обладать свойствами мультидоменности и динамичности, а точка зрения на решаемую проблему – задавать некоторый постоянный домен:

$$S\big(D(t)\big) = \{Domain^P, Domain(t)_1, \dots, Domain(t)_n\},$$
$$\forall\, Domain_i(t), Domain_j(t) \in S(t), Domain_i(t) \neq Domain_j(t),$$

где $Domain^P$ задает стабильную часть предметных знаний.

Формально определим домен как систему, состоящую из концептов, связей, классов связей, функции, определяющей связи между концептами, и функции, определяющей вес связи и существующую в статичном виде только в некоторый момент времени t:

$$Domain(t) = (Concept^D(t), Relations^D(t),$$
$$RelTypes, r\_assignment, w\_estimating),$$

где $Concept^D(t)$ – множество концептов/терминов домена в момент времени t; $Relations^D(t) \subseteq Concept^D(t) \times Concept^D(t) \times RelTypes$ –

множество связей между вершинами, где каждая связь rel – это тройка (c_i, c_j, rt), где $rt \in RelTypes$ в момент времени t, $RelTypes$ – множество типов связей; $w\_estimating: Relations^D(t) \to R_{>0}$ – функция, присваивающая каждой связи некоторый вес; $r\_assignment: Concept^D(t) \times Concept^D(t) \to B(RelTypes)$ – функция, присваивающая каждой паре концептов некоторый тип или типы неповторяющихся связей из $RelTypes$: $r\_assignment(c_i, c_j) \subseteq RelTypes$ – определяет набор отношений между вершинами c_i и c_j; $r\_assignment(c_i, c_j) = \emptyset$, если связей между терминами нет. $B(RelTypes)$ – множество всех подмножеств $RelTypes$.

Структурные ограничения:

- концепты не могут быть связаны сами с собой:

$$\forall c_i \in Concept^D(t), \forall r \in RelTypes, r\_assignment(c_i, c_i) = \emptyset;$$

- одновременное существование между двумя вершинами только уникальных связей:

$$\forall c_i, c_j \in Concept^D(t), \forall r \in RelTypes, \left|\{c_i, c_j, rt\}\right| \leq 1.$$

Учет времени:

$$Concept^D(t + \Delta t) = Concept^D(t) \cup Concept_{new}(t) - Concept_{removed}(t), Relations^D(t + \Delta t) = Relations^D(t) \cup Relations^D{}_{new}(t) - Relations^D{}_{removed}(t).$$

Различие доменов. Пусть есть два домена – $Domain_1(t)$ и $Domain_2(t)$:

$$Domain_1(t) = (Concept_1^D(t), Relations_1^D(t), RelTypes_1, r\_assignment, w\_estimating),$$

$$Domain_2(t) = (Concept_2^D(t), Relations_2^D(t), RelTypes_2, r\_assignment, w\_estimating).$$

Будем утверждать, что $Domain_1(t) \neq Domain_2(t)$, если выполняется хотя бы одно из следующих условий:

1. Домены различаются по составу множества концептов:

$$Concept_1^D(t) \neq Concept_2^D(t);$$

2. Домены различаются по составу множества связей:

$$Relations_1^D(t) \neq Relations_2^D(t);$$

3. Домены совпадают по составу множеств концептов и связей, но различаются веса этих связей:

$$\exists rel \in Relations_1^D \cap Relations_2^D,$$
$$w\_estimating(rel, Domain_1) \neq w\_estimating(rel, Domain_2).$$

Коммуникации в социальных медиа материализуются в форме публикаций и комментариев, которые далее именуются документами. Все множество документов, являющихся объектами поиска в контексте рассматриваемой метазадачи, образуют коллекцию. Процесс появления публикаций в социальных медиа не случаен и характеризуется наличием причинно-следственной структуры. Так, например, комментарии, как правило, являются реакцией на публикацию или ассоциированные с ней понятия. Или, например, новостная публикация может быть обновлением информации касательно события, уже произошедшего и получившего упоминание в социальных медиа. Для адекватного представления специфики социальных медиа в контексте задач ИП в модельном представлении необходимо отобразить структуру коммуникации на коллекцию документов.

Соответственно, определим мультидоменную динамическую коллекцию документов как:

$$MDDC(t) = \{D(t), L, S(t)\},$$

где $D(t) = \{d_1, \ldots, d_n\}$ – множество документов источника в некоторый момент времени t. Каждый документ $d \in D(t)$ представлен множеством предложений $d = \{snt_1, \ldots, snt_i, \ldots, snt_n\}$, где $snt \in d$, $snt = \{wrd_{i1}, sep_{i1}, \ldots, wrd_{im}, sep_{ik}\}$ – предложение, состоящее из слов wrd_{ij} и знаков препинания sep_{ib}; $L \subseteq D \times D$ – асимметричное транзитивное отношение, задающее последовательность документов в иерархии (причинно-следственную структуру). $d_1 L d_2$ означает, что документ d_2 следует после документа d_1. Такая цепочка состоит из веток, представляющих собой последовательности Br документов из $D(t)$ ($Br \subseteq D(t)$), удовлетворяющие условию:

$$Br = \{d_1, \ldots, d_N\}: \forall i < j, d_i L d_j .$$

Полная цепочка (полное дерево, CS) задается корневым документом d и всеми транзитивно связанными с ним документами $Br(d)$.

$S(t)$ – множество доменов в момент времени t.

Функция определения принадлежности документа d к домену s в момент времени t имеет следующий вид:

$$d\_entry(d,s,t)\colon D(t) \times S(t) \to [0,1].$$

Таким образом, решаемая задача ИП в социальных медиа может быть формально описана следующим образом. Имеется мультидоменная динамическая коллекция документов $MDDC$ и запрос q, сформулированный пользователем. Система должна вернуть упорядоченное подмножество релевантных документов следующего вида:

$$(D', \preceq) = \{d \in D(t) | f_{rel}\,(d,q) > th\},$$
$$f_{rel} : D \times \{q\} \to E,$$
$$d_i \preceq_q d_j \Leftrightarrow f_{rel}\,(d_i\,, q) \leq f_{rel}\,(d_j\,, q),$$

где th – некоторое пороговое значение релевантности, E – множество оценок (например, множество вещественных чисел).

4. Модель мультидоменной динамической системы знаний. Для решения исследуемой, в рамках ИП, задачи определения релевантных публикаций предлагается использовать МДСЗ, состоящую из трех основных частей (Рис. 1):

1. Проблемно-ориентированная часть, представленная онтологией. Она задает некоторую точку зрения на решаемую проблему и данные соответственно, при этом не ограничивая остальную систему.

2. Мультидоменная часть системы, представленная мультидоменным динамическим ГЗ, отражает знания, заложенные в динамических мультидоменных структурированных коллекциях документов, и ментальные стереотипы сообществ или пользователей, породивших эти коллекции соответственно.

3. Доменно-независимая часть, представленная лексической БД, необходима для разрешения лексической многозначности (word sense disambiguation) и является отправной точкой для согласования понятий проблемно-ориентированной и мультидоменной частей. Согласование концептов, хранящихся в системе, предлагается производить с помощью большой языковой модели (LLM), так как она в некотором роде содержит усредненную доменно-независимую семантику.

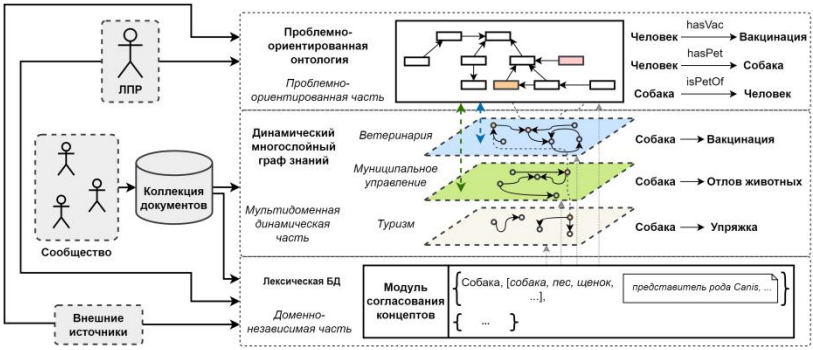

Рис. 1. Концептуальная модель МДСЗ

Процесс построения МДСЗ начинается с формирования лексической БД, которая задается с помощью внешних источников данных, и проблемно-ориентированной онтологии, задаваемой с помощью эксперта или внешних источников. Лексическая БД пополняется также концептами, входящими в онтологию, и наоборот – при создании онтологии могут использоваться концепты, уже заданные ранее в лексической БД. Далее осуществляется построение ГЗ на основе мультидоменной коллекции документов социальных медиа, которые предварительно преобразовываются в структуры коммуникаций с последующим выделением в них концептов. Данный процесс продемонстрирован на рисунке 2.

Пополнение МДСЗ новой информацией происходит при необходимости в процессе её применения. Онтология обновляется при обнаружении неполноты или несоответствия текущей структуры или экземпляров по отношению к решаемой прикладной задаче и ожиданиям пользователя. При этом новые экземпляры могут происходить как и из ГЗ, так и из каких-либо внешних источников, если того требует решаемая задача при условии, что они удовлетворяют заданным в онтологии свойствам и ограничениям. Онтология перестраивается полностью при кардинальной смене предметной области мониторинга, например, при смене пользователя. ГЗ пополняется новой статистической информацией из наблюдаемых в рамках мониторинга документов в пакетном режиме. Перестройка или дополнение ГЗ новыми слоями происходит при значительном изменении семантических зависимостей в наблюдаемых данных, что, как правило, происходит со временем или при смене наблюдаемого множества источников. Лексическая БД пополняется новыми терминами и/или концептами, встречающимися

в обрабатываемых документах, по мере необходимости в течение всего времени работы системы.

Таким образом, определение релевантности с помощью предлагаемой системы может производиться как отдельно на каждом из её уровней, так и совместно для достижения наилучшего результата. Так, лексическая БД обеспечивает решение проблемы лексической многозначности, что позволяет использовать систему для простого поиска по ключевым словам, принимая во внимание различные лексикализации понятий. Многослойный ГЗ позволяет взглянуть на понятия и отношения между ними с точки зрения актуальных тематических представлений сообществ и, при необходимости, комбинировать таковые без перестройки системы. Для определения релевантности с его помощью предлагается использовать метрику расстояния между терминами запроса и документов (см. раздел 5). Онтология, в свою очередь, задает проблемно-ориентированную точку зрения на решаемую проблему с осмысленными отношениями и правилами вывода, с помощью которых, в том числе, можно определять релевантность документов. Далее рассмотрим компоненты МДСЗ более подробно.

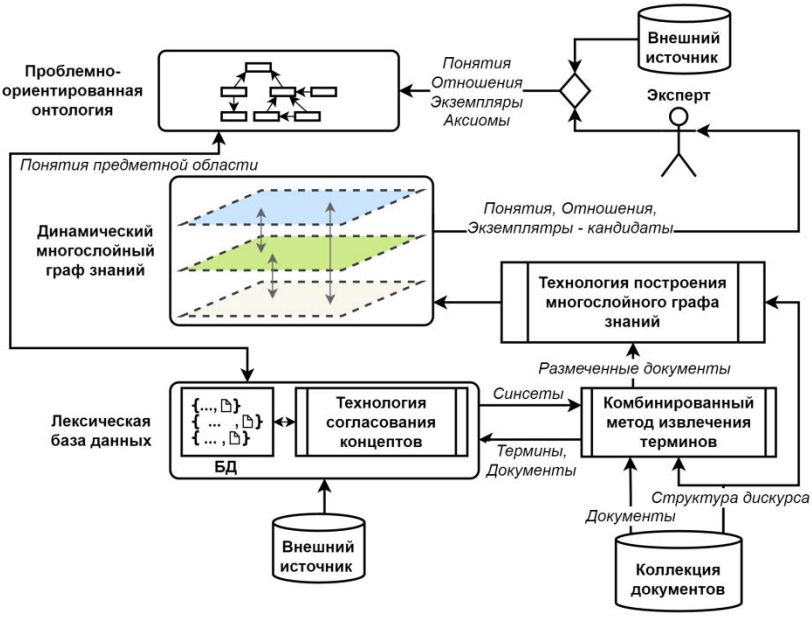

Рис. 2. Функциональная структура МДСЗ

4.1. Лексическая БД. Лексическая БД содержит структурированную информацию о различных словах и отношениях между ними, а также сопутствующую лингвистическую информацию, например, о частоте использования слов или морфологическую информацию. От тезауруса или словаря её отличает ряд признаков (таблица 1).

Таблица 1. Различия между лексической БД и тезаурусом

Признак	Лексическая БД	Тезаурус
Цель	Хранение лексических единиц, отношений между ними, а также сопутствующей лингвистической информации без привязки к конкретной области знаний, чаще с привязкой к конкретному языку.	Хранение лексических единиц конкретной области знаний, семантических отношений между ними (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.п.) и определений для них.
Структура	Сложная сеть из лексических единиц, их отношений, признаков и сопутствующей лингвистической информации.	Простые наборы лексических единиц с их определениями.
В каких задачах используется	Обработка естественного языка, исследования, касающиеся лингвистики.	Написание различных текстов, для получения справочной информации при переводе, для описания предметных областей с помощью понятий.

Исходя из этого, можно заключить, что тезаурусы предназначены для повседневного использования, в то время как лексические БД используются в исследовательской работе. Например, можно выделить лексическую БД WordNet [11], которая содержит обширную информацию об английских частях речи, включая существительные, глаголы, прилагательные и наречия. Основой структуры WordNet являются синсеты – специальные группы слов, принадлежащих к одной части речи и обладающих высокой степенью синонимичности в конкретном контексте. В качестве примера можно привести синсет, включающий лексемы {car, auto, automotive, machine, motorcar}, которые взаимозаменяемы при обозначении транспортного средства определенного типа. Каждый синсет сопровождается пояснительным комментарием, который содержит краткое описание обозначаемого понятия. Например, для вышеупомянутого синсета приводится определение: “четырехколесный” – обычно приводимый в движение двигателем внутреннего сгорания. Особую значимость в архитектуре WordNet представляет система семантических связей между синсетами, которая включает такие типы отношений, как

гипонимия, гиперонимия и меронимия, что позволяет выстраивать сложные семантические сети между лексическими единицами. Также в качестве примера можно привести CELEX [12], включающий в себя три лексические БД для голландского, английского и немецкого языков, содержащие орфографическую, фонологическую, морфологическую, синтаксическую и частотную информацию.

Формально разрабатываемую лексическую БД можно описать следующим образом:

$$LD = \{Concept^{LD}, Vocabulary, Context\},$$

где $Concept^{LD}$ – множество концептов БД (включая концепты из внешних источников, онтологии и полученные в процессе обработки документов), где каждый отдельный концепт c представлен четверкой $(id, name, synset, context)$, где $id \in ID$ (ID – множество уникальных меток) – уникальная метка концепта в рамках БД; $name \in Vocabulary$ – наименование концепта на естественном языке; $synset \subset Vocabulary$ – множество терминов (словоформ и синонимов) данного концепта; $context$ – некоторый контекст употребления концепта. Для любого термина может существовать несколько контекстов и для любого контекста может существовать несколько терминов, но для одного синсета и, соответственно, концепта существует только один контекст; $Vocabulary$ – множество терминов или лексических представлений (слов или словосочетаний); $Context$ – множество контекстов лексической БД.

Контексты в рамках данной системы выполняют функцию определения некоторого смысла для концепта в виде, приспособленном для обработки компьютером. Соответственно, в качестве способа представления концептов предлагается использовать векторные представления, полученные из больших языковых моделей. Такой выбор обоснован богатым семантическим кодированием, которое обеспечивают подобные векторные представления, фиксируя не только поверхностное лексическое сходство, но и более глубокие контекстные и функциональные связи между концептами. Формирование концептов (сопоставление и распределение терминов по смыслу) в рамках лексической БД можно рассматривать как задачу кластеризации в многомерном семантическом пространстве, что соответствует теории дистрибутивной семантики. Таким образом, для динамической группировки связанных терминов под общими концептами и автоматизированной адаптации к ранее не встретившимся терминам

можно использовать метрики векторного сходства, например, косинусное расстояние. Последние исследования [13 – 15] применения больших языковых моделей и, в частности, векторных представлений из них в области обработки естественного языка подтверждают эффективность их использования для решения подобных задач.

Взаимодействие внутри системы с лексической БД предусматривает наличие функций пополнения, разметки документа, поиска концепта и получения информации.

Пополнение лексической БД необходимо для её поддержания в соответствии с обрабатываемыми данными и производится следующим образом:

1. Если такого термина и контекста еще нет в БД, то добавляем термин в словарь (вместе с его контекстом) и добавляем соответствующий ему концепт.

2. Если такой контекст уже существует, а термин – нет, то добавляем термин в словарь, находим синсет, содержащий этот контекст, и добавляем в этот синсет новый термин.

3. Если такой термин уже существует, а контекст – нет, то добавляем контекст в БД, связываем термин с этим концептом и создаем соответствующий концепт.

Основной задачей лексической БД является разметка документов, которая заключается в сопоставлении терминов документа с известными концептами. Разметку также предполагается производить при помощи векторных представлений аналогично функции пополнения. Альтернативно, подобное сравнение терминов возможно с использованием простых правил сопоставления текстовых представлений, статистических мер или, например, тематического моделирования. При этом, если встречается термин или концепт, не присутствующий в лексической БД, то он добавляются в нее.

Получение информации из лексической БД предполагает обработку конкретных запросов, например, получение конкретного концепта, синсета или набора терминов.

Для управления такой БД необходима соответствующая специфике задачи система управления БД. В нашем случае основными кандидатами являются векторные БД (например, ChromaDB [16], Milvus [17]), созданные специально для эффективной работы с многомерными векторными представлениями и позволяющие интегрировать в процесс обработки данных большие языковые модели.

4.2. Комбинированный метод извлечения терминов. Метод нацелен на извлечение как предметных, так и общеупотребительных терминов. На вход метод принимает экземпляр структуры

коммуникации, состоящий хотя бы из одного корневого документа. Затем для каждого документа производится извлечение терминов онтологии, именованных сущностей, осмысленных словосочетаний и контекстно значимых существительных. Осмысленность словосочетаний оценивается комбинацией извлекающего правила на основе контекстно-свободной грамматики и меры коллокации. Контекстная важность существительного определяется исходя из статистической оценки его употребления в рамках соответствующего экземпляра структуры коммуникации. После извлечения всех терминов-кандидатов (все термины, кроме терминов онтологии, считаются кандидатами на этом этапе), происходит их фильтрация по шагу извлечения, стоп-словам, длине и частоте употребления. Данный процесс представлен на рисунке 3 и более подробно описан в работе [18].

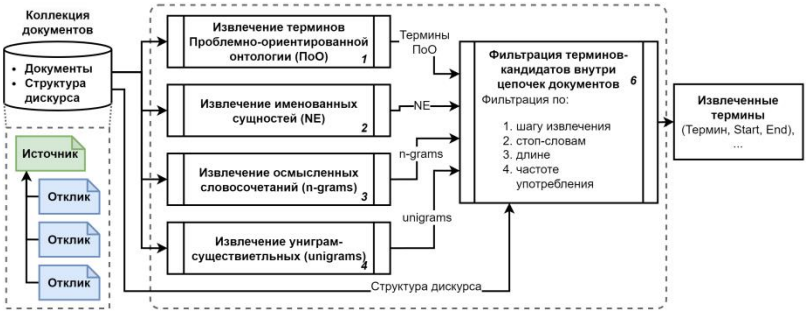

Рис. 3. Схема работы комбинированного метода извлечения терминов

4.3. Динамический многослойный ГЗ. В данной работе используется контекстно-зависимый подход к разбиению ГЗ на слои, основанный на тематической кластеризации документов, полученной в результате тематического моделирования [19]. Такой подход позволяет интерпретировать данные с помощью скрытых тем, идентифицированных в исходном наборе документов. Полученный граф представляет концепты как вершины со взвешенными направленными ребрами, захватывающими статистически значимые паттерны их совместного употребления, улучшенные выходными данными тематического моделирования. Реализуя разбиение на основе контекста, рассматриваемое как отдельные слои графа или взаимосвязанные подграфы, подход облегчает выявление значимых связей даже для тем с низкой распространенностью. Отметим, что эффективность предлагаемого подхода будет зависеть, в том числе от свойств и качества используемой тематической модели.

Формально разработанную технологию, использующую тематическую модель в качестве основы для разбиения на контексты, можно представить следующим образом.
Дано:
$MDDC$ – мультидоменная динамическая коллекция документов.
Найти:

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

где V – множество вершин – уникальные концепты из коллекции документов $Concept^{MDDC} \subset Concept^{LD}$; E – множество ребер – взвешенные направленные ассоциативные отношения между концептами; 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.

Процесс построения ГЗ включает в себя три основных этапа:

1. Получение размеченных экземпляров структур коммуникаций.
2. Обучение тематической модели.
3. Построение многослойного ГЗ (Рис. 4).

На первом этапе исходная коллекция документов подвергается предварительной обработке, в рамках которой из нее исключаются несодержательные документы и мусорные конструкции. После чего коллекция, в соответствии со структурой, преобразуется в набор экземпляров структур коммуникаций CS, где каждый экземпляр задает полное дерево документов (например, публикация и все связанные с ней комментарии). Затем, с помощью авторского комбинированного метода извлечения терминов, каждый экземпляр размечается терминами:

$$TermExtrMethod(CS) \to CS^{term\ marked}.$$

В результате исходный набор экземпляров структур коммуникаций дополняется размеченными терминами. После этого, с помощью лексической БД, данный набор размечается концептами:

$$annotate(CS^{term\ marked}) \to CS^{concept\ marked},$$

что обеспечивает устранение лексической многозначности. Затем концепты в каждом экземпляре фильтруются для выявления контекстно важных. На данном этапе работы контекстная важность термина определяется исходя из статистической оценки его употребления в рамках соответствующего экземпляра структуры коммуникации.

На втором этапе отфильтрованные экземпляры структур коммуникаций используются при формировании датасета для тематического моделирования, и производится само тематическое моделирование, результатом которого являются тематические распределения этих экземпляров и концептов:

$$TopicModelConst(CS, K) \rightarrow \{\Theta, \Phi\},$$

где Θ – матрица распределения экземпляров структур коммуникаций по темам, Φ – матрица распределения концептов по темам.

На третьем этапе производится подсчет статистики совместного употребления концептов с учетом полученных тематических распределений и заполнение соответствующих тематизированных матриц: $f(CS, \Theta, \Phi) \rightarrow CM$ – набор матриц совместного употребления, где элемент матрицы совместного употребления по k-ой теме вычисляется по формуле:

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

где $\theta_{CS^P}^{k}$ – значение принадлежности экземпляра CS^P теме k; CS_{ij}^{P} – значение совместного употребления концептов i и j в рамках CS, получаемое функцией подсчета совместного употребления в экземпляре структуры коммуникации: $fp(CS^T) \rightarrow CS^P$ вида $< i, j, c_{ij} >$, где i, j – концепты, c_{ij} – значение их совместного употребления в конкретном экземпляре CS; ϕ_i^k – значение принадлежности концепта i теме k; $\max(\phi_i)$ – максимальное значение принадлежности концепта i темам.

Далее выделяются значимые отношения из полученных матриц следующим образом:

1. Для каждого слоя, соответствующего теме из тематической модели, выбирается процент самых сильных связей, соответствующий распределению концептов по темам в тематической модели;

2. Удаляются незначимые значения в соответствии с эвристически заданным порогом:

$$CS_k^P(i,j) > \overline{x}\left(\overline{x}\left(CS_k^P(i,)\right), \overline{x}\left(CS_k^P(,j)\right), \overline{x}\left(CS_k^P\right)\right), \quad (2)$$

где $\overline{x}\left(CS_k^P(i,)\right)$ – среднее значение совместного употребления для i, $\overline{x}\left(CS_k^P(,j)\right)$ – среднее значение совместного употребления для j, $\overline{x}\left(CS_k^P\right)$ – среднее значение совместного употребления для всех концептов в соответствующей матрице (теме) k.

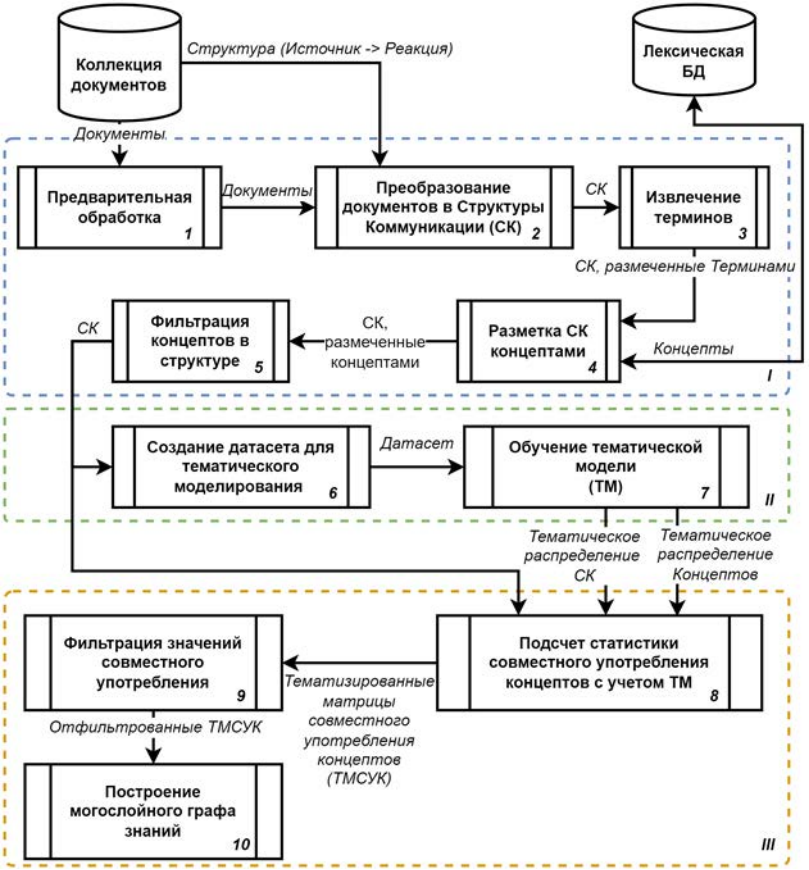

Рис. 4. Технология формирования многослойного ГЗ с применением тематического моделирования

В результате получается набор матриц со значениями совместной встречаемости концептов, соответствующий тематической модели. Набор уникальных концептов интерпретируется как вершины графа, а ненулевые значения матриц совместной встречаемости – как взвешенные ассоциативные отношения между ними.

4.4. Проблемно-ориентированная онтология. Как было отмечено ранее, данная онтология задает некоторую отправную точку зрения на решаемую проблему. Соответственно, в ней содержатся базовая система понятий (классы), их свойства, отношения, экземпляры и аксиомы.

Проблемно-ориентированная онтология задается некоторым внешним источником знаний EKS или экспертными знаниями EK:

$$f: EKS \cup EK \rightarrow PoO$$
$$PoO = \{C, P, R, I, A\},$$

где: C – множество классов; P – множество свойств классов; R – множество отношений между классами; I – множество экземпляров. При этом $\forall r \in R: I(r) \subseteq I \times I,$ где $I(r)$ – множество пар экземпляров, связанных отношением r; A – множество аксиом.

В рамках разрабатываемой системы знаний онтология описывает предметную область поиска или, иначе говоря, контекста, в котором находится пользователь системы. Для работы системы в целом наибольшую роль играет наполнение онтологии экземплярами, что повышает точность и полноту извлечения заведомо релевантных терминов с помощью метода извлечения терминов. Далее эти термины встраиваются в ГЗ, где играют важную роль при определении релевантности документов на его уровне. Отношения, заложенные в онтологию, потенциально могут быть использованы для расширения запроса или документов или для определения/уточнения релевантности на её уровне на основе заложенных в нее аксиом. Однако на данном этапе работы определение релевантности на основе аксиом онтологии не производится и является возможным направлением дальнейших исследований.

На текущий момент в онтологии описана базовая структура классов, представляющих собой основные категории именованных сущностей (Локации, Организации, Личности), включающих подклассы, определены отношения между ними, некоторые свойства классов, а также заданы наиболее значимые экземпляры.

5. Технология определения релевантности публикаций с помощью многослойного графа знаний. Для определения

релевантности с помощью полученного многослойного ГЗ предлагается использовать метрику расстояния между терминами запроса и терминами публикации в наиболее релевантных слоях графа. Релевантность слоя оценивается исходя из тематического распределения запроса, полученного на основе используемой тематической модели. Таким образом, наиболее релевантными считаются слои, тематически соответствующие запросу (Рис. 5). Подобную технологию можно интерпретировать как расширение запроса при помощи ГЗ, содержащего объективные представления о стереотипах пользователей.

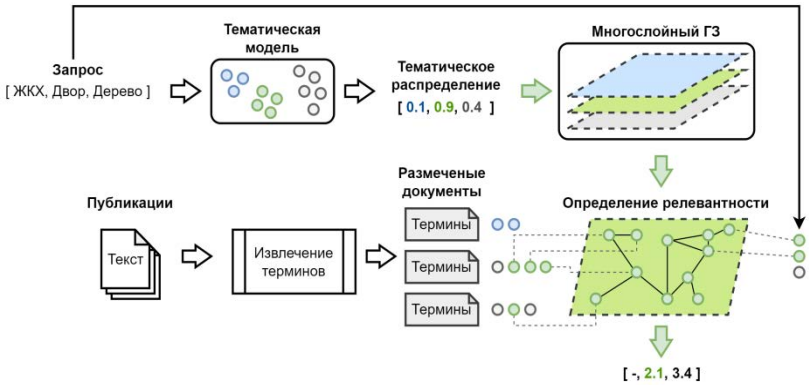

Рис. 5. Определение релевантности с помощью расстояния на графе

Перед определением расстояния на графе имеющиеся взвешенные ассоциативные связи преобразуются в расстояния, обратно пропорциональные весу и нормализованные в диапазоне значений (0, 1). Данное преобразование проводится на матрицах, на основе которых строятся графы, в соответствии с формулой:

$$N_{ij}^{k} = \begin{cases} 1 - \dfrac{M_{ij}^{k} - M_{adjmin}^{k}}{M_{adjmax}^{k} - M_{adjmin}^{k}}, if\ M_{ij}^{k} \neq 0 \\ 0\ , if\ M_{ij}^{k} = 0 \end{cases},$$

где M_k – матрица весов слоя k; $M_{adjmin}^{k} = M_{min}^{k} \cdot 0{,}99999$ – скорректированный минимум матрицы M для слоя k; $M_{adjmax}^{k} = M_{max}^{k} \cdot 1{,}00001$ – скорректированный максимум матрицы M для слоя k.

Соответственно, для оценки релевантности публикации запросу в рамках одного слоя предлагается использовать среднее значение минимальных расстояний от терминов публикации P^t до терминов запроса Q^t:

$$D(P^t, Q^t) = \frac{1}{m}\sum_{i=1}^{m} \min_{j=1,\dots,n} d(p_i, q_j), \quad (3)$$

где $d(p_i, q_j)$ – расстояние (вес кратчайшего пути) от $p_i \in P^t$ до $q_j \in Q^t$.

6. Эксперимент (определение релевантности публикаций с помощью многослойного графа знаний). Рассматриваемые в данной работе модели и методы ориентированы, прежде всего, на задачи мониторинга потоков сообщений, в рамках которых требуется в потоковом режиме определять релевантность появляющихся сообщений. При этом задача осложняется изменчивостью лексикализации семантики предметной области, а также тем, что критерии отбора могут формироваться еще до появления первых релевантных сообщений, что ограничивает возможность использования подходов на базе “эталонных” сообщений, в том числе представленных в виде эмбеддингов. Кроме того, в практических задачах ИП запрос формируется, как правило, неподготовленным пользователем. В таких условиях фактически единственным способом задания критериев релевантности информационной выдачи являются ключевые слова. При этом опыт эксплуатации систем мониторинга сообщений показывает, что начальный поисковый запрос в виде набора ключевых слов претерпевает постепенное расширение, обусловленное уточнением ожиданий пользователя в отношении релевантных результатов.

Исходя из вышеизложенного, в рамках эксперимента проверялась гипотеза о том, что путем формирования многослойного ГЗ из данных социальных медиа возможно решить задачу определения релевантности публикаций не хуже технологии, использующей только ключевые слова.

В рамках решаемой задачи мониторинга социально-политической обстановки набор ключевых слов, по существу, является запросом для поиска релевантных публикаций. Получение и поддержание такого набора в актуальном состоянии является достаточно трудоемкой задачей в силу того, что необходимо обладать не только знанием темы, в рамках которой осуществляется поиск, но иногда и интернет-жаргоном, связанным с ней. С одной стороны,

такой набор может позволить достаточно точно определить релевантность некоторых документов, но, с другой стороны, он требует постоянной модификации для включения новых терминов и их лексикализации или исключения более нерелевантных терминов. Однако стоит отметить, что расширение данного набора терминов может приводить к получению большого числа ложноположительных результатов, т.е. нерелевантных публикаций. Так, например, набор ключевых слов из реального проекта (по проблемно-ориентированному мониторингу публикаций), используемый в рамках данного эксперимента, в начале содержал около 15 терминов, затем пополнялся в процессе использования, в течение нескольких лет. В итоге, расширился до 133 терминов (включая словосочетания и аббревиатуры, которые покрывают ряд различных тематик) на момент проведения эксперимента.

Суть эксперимента заключалась в оценке эффективности предлагаемой технологии определения релевантности с помощью многослойного ГЗ посредством сравнения с технологией, основанной на ключевых словах (точное совпадение и основы слов) (Рис. 6).

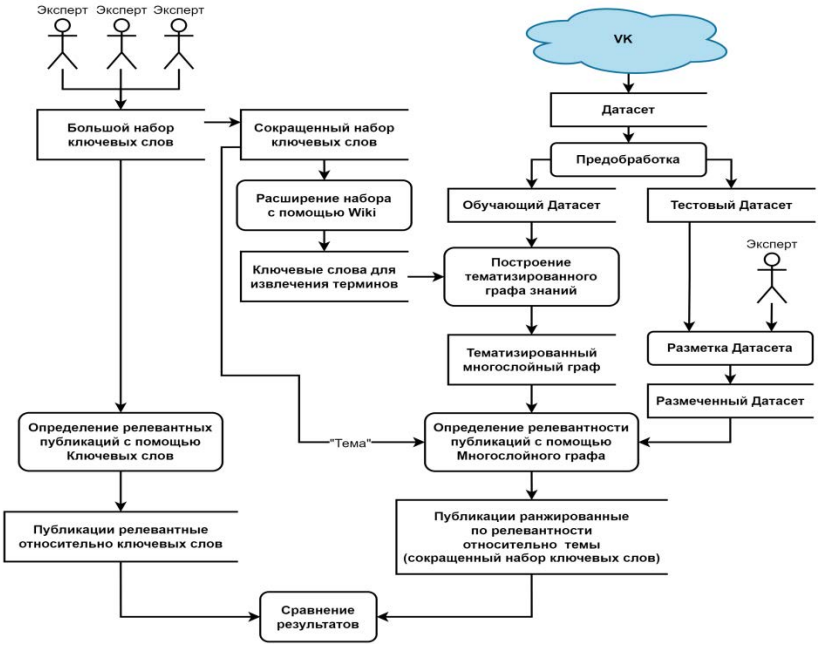

Рис. 6. Схема эксперимента по оценке эффективности предлагаемой технологии определения релевантности с помощью многослойного ГЗ

6.1. Формирование многослойного ГЗ. На первом этапе из упомянутого ранее набора ключевых слов (133 термина) было экспертно отобрано 11 терминов, наиболее полно отражающих общую тему запроса.

Для проведения эксперимента использовался датасет публикаций из 25 групп социальной сети “ВКонтакте” за период примерно в один год (начало 2024 – начало 2025), содержащим более 73 тыс. публикаций и 107 тыс. комментариев к ним. Данный датасет был очищен от повторяющихся и коротких публикаций, содержащих менее 76 символов. Как показали экспертные наблюдения, такие публикации крайне редко были содержательными или имели какой-то смысл для проводимого эксперимента. Данное число символов было выбрано исходя из средней длины слов для разговорной речи и средней длины предложения в русском языке. В результате осталось 35469 публикаций, которые использовались для построения ГЗ. Для оценки технологии использовался размеченный на основе исходного набора ключевых слов датасет из 617 публикаций, 100 из которых считались релевантными и 517 нерелевантными.

Для формирования ГЗ, обучающий датасет был преобразован в набор структур коммуникаций, где каждая структура является представлением публикации и её откликов (комментариев). Вместе с этим проводилось выделение контекстно важных терминов в этих структурах с помощью авторского комбинированного метода извлечения терминов. Термины для словаря, являющегося частью метода, были получены на основе малого набора ключевых слов, который был расширен экспертом с помощью инструмента для поиска связанных терминов, полагающегося на информацию из Википедии [20], и в итоге содержал 26 терминов. 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 32883 структур, содержащих хотя бы два уникальных термина.

На основе полученных структур был обучен ряд тематических моделей, количество тем которых варьировалось от 5 до 20, однако в дальнейшем было решено использовать модель с 11 темами исходя из её показателя согласованности тем. Далее на основе полученных структур и выбранной тематической модели был произведен подсчет статистики совместного употребления терминов с учетом тематического

распределения структур (формула (1)). В результате были получены 11 матриц (по количеству тем выбранной тематической модели) значений совместного употребления терминов. Каждая такая матрица отражает некоторую меру совместного употребления терминов в конкретной теме. Полученные значения дополнительно были отфильтрованы от незначительных в соответствии с формулой (2). Таким образом, был получен ГЗ, состоящий из 11 слоев, вершинами которого стали уникальные термины, имеющие хотя бы одну связь в любом из тематических слоев, а значения совместного употребления, соответственно, определили наличие и силу связи между терминами в каждом слое графа.

6.2. Определение релевантности публикаций. Ранее полученный малый набор из 11 ключевых слов использовался как запрос пользователя для получения оценок релевантности на основе ГЗ. При этом, используя ранее полученную тематическую модель, были получены векторы тематического распределения терминов из этого набора, что позволяет понять, какие темы (они же слои в графе), являются наиболее релевантными запросу. Соответствующее тематическое распределение запроса, полученное как сумма векторов его терминов, приведено в таблице 2 (для удобства значения были нормализованы). Соответственно, релевантными слоями графа считались слои 1, 5 и 9.

Таблица 2. Тематическое распределение запроса, заданного малым набором ключевых слов

1	2	3	4	5	6	7	8	9	10	11
0,9751	0,0003	0,0958	0,0016	**0,7463**	0,0001	0,0003	0,4756	**0,9999**	0,3014	0,0505

Релевантность публикаций в каждом слое графа определялась в соответствии с формулой (3). Для сравнения с методами на основе ключевых слов, которые дают оценку вида да/нет, был выбран порог значения релевантности 0,9, чтобы представить результаты в аналогичном виде. Стоит отметить, что выбор наиболее подходящего порогового значения осложняется из-за несбалансированности имеющегося тестового набора. Соответствующее значение было выбрано исходя из баланса значений метрик точности и полноты с учетом истинно положительных и истинно отрицательных оценок (Рис. 7). Таким образом, публикация считалась релевантной, если она получала минимальное значение метрики в одном из релевантных слоев менее или равное 0,9,

и нерелевантной, если полученное минимальное значение метрики превышало порог или было в любом другом слое графа.

В рамках данного эксперимента проводилось сравнение предлагаемой технологии с технологиями на основе точного (ExactMatch) и частичного (StemmerMatch) совпадений. При определении релевантности с помощью этих технологий публикация считалась релевантной, если содержала хотя бы одно точное совпадение с ключевым словом или вхождение основы слова соответственно. В случае словосочетаний такая проверка проводилась отдельно для каждого слова. В таблице 3 и на рисунках 8-9 представлены результаты эксперимента.

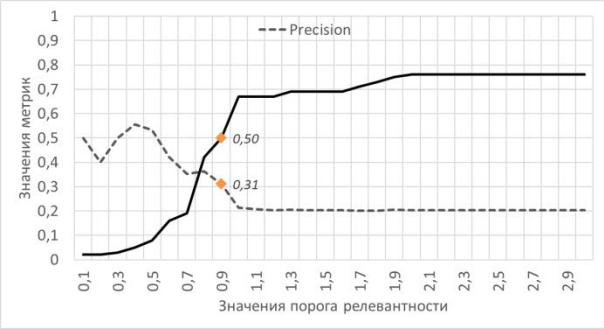

Рис. 7. Зависимость значений метрик Precision и Recall от порога релевантности

Таблица 3. Результаты проверки работы метода на тестовом наборе

	Малый набор ключевых слов (11)			**Большой набор ключевых слов (133)**	
	Многослойный ГЗ	**Exact-Match**	**Stemmer-Match**	**Exact-Match**	**Stemmer-Match**
TP	50	1	42	72	97
TN	406	517	471	402	339
FP	111	0	46	115	178
FN	50	99	58	28	3
Precision	**0,31**	1,00	0,48	**0,39**	**0,35**
Recall	0,50	0,01	0,42	0,72	0,97

Предлагаемая технология практически не уступает по показателям точности и полноты технологиям на основе ключевых слов, использующих полный набор таковых, и превосходит эти же технологии, работающие на неполном наборе ключевых слов. Если смотреть с позиции необходимости идентификации релевантных

публикаций с наименьшим числом ложноположительных результатов, то в целом подтверждается гипотеза о том, что предлагаемая технология может работать не хуже методов на основе ключевых слов, использующих значительно больший набор таковых. Иными словами, с помощью полученного ГЗ нам удалось добиться эффекта расширения набора ключевых слов экспертом с учетом социальных стереотипов. Дополнительно стоит заметить, что предлагаемая технология дает не просто оценку вида, в формате да/нет, а некоторую количественную оценку релевантности, что, в свою очередь, позволяет ранжировать полученные результаты.

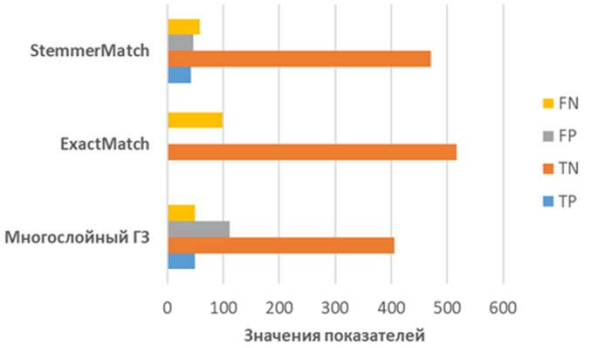

Рис. 8. Результаты эксперимента на малом наборе ключевых слов

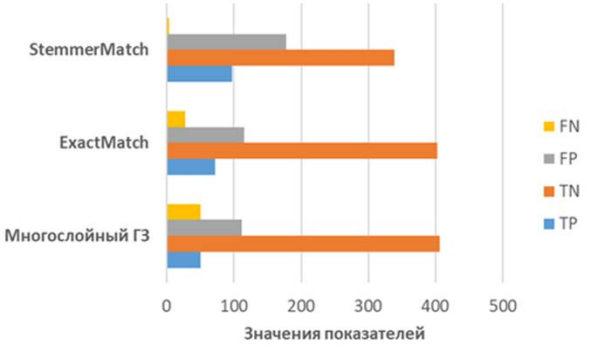

Рис. 9. Результаты сравнения многослойного ГЗ с технологиями, использующими большой набор ключевых слов

На рисунке 10 приведен анализ нескольких примеров публикаций с ложноположительными (публикации №1 и №2) и ложноотрицательными оценками (публикации №3 и №4). Сверху приведен текст публикаций и их оценки экспертами и предлагаемой технологией. Снизу приведены полученные метрики расстояний по слоям, красным выделены минимальные значения в релевантных слоях (1, 5 и 9), которые оказались ошибочными по тем или иным причинам. Для публикаций №1 и №2 наименьшая оценка расстояния находится в одной из релевантных тем, и её значение ниже заданного порога, но сами публикации не являются релевантными. При этом публикация №1 содержит крайне мало текстовой информации в целом. Для публикаций №3 и №4 наименьшая оценка релевантности находится все еще в релевантных слоях, но превышает заданный порог. Значительно превышает (более чем в 2 раза) для №3 и незначительно превышает для №4. Данные примеры демонстрируют выявленные слабые места технологии, связанные с её реализацией или выбранным пороговым значением. Таким образом, необходимо дальнейшее исследование и модернизация выбранного подхода к определению релевантности с помощью метрики расстояния на графе, чтобы избежать или минимизировать подобные ошибки.

	Текст публикации	Метка эксперта	Многослойный ГЗ
№1	Новогодние свободные катания в Ледовом дворце Апатит Арена Сегодня в 18 00 и 20 00.	0	1
№2	Кировчанин Артем Андреев стал бронзовым призером чемпионата Мурманской области по лыжным гонкам в дисциплине классический спринт Андрей Малетин и Владелинка Архипова впервые приняли участие первый в Чемпионате, получили опыт и мотивацию работать дальше.	0	1
№3	Минтранс опроверг слухи о повышении стоимости проезда до 50 рублей. Рост цен на энергоресурсы и другие услуги делает возможным увеличение тарифов, однако конкретных решений пока нет.	1	0
№4	Изменились цены на посещение природных парков в Мурманской области С 1 января 2025 года вступают в силу новые цены на посещение особо охраняемых природных территорий, таких как природный парк Териберка, Полуострова Рыбачий и Средний и Сейдъяввръ. Териберка Граждане РФ 440 рублей, иностранные граждане 880 рублей. Рыбачий и Средний 370 рублей. Сейдъяввръ 740 рублей. Для жителей Мурманской области, школьников, дошкольников, многодетных родителей и других льготных категорий вход по-прежнему бесплатный.	1	0

	Тематический слой графа										
	1	2	3	4	5	6	7	8	9	10	11
№1	**1,48**	1,96	1,47	1,87	**2,14**	1,95	1,79	1,84	**0,78**	-	1,93
№2	**0,76**	1,88	1,75	3,29	**1,69**	1,96	1,85	1,76	**0,96**	-	0,99
№3	**1,83**	2,27	2,71	2,51	**2,19**	1,99	2,27	1,88	-	-	1,98
№4	**1,73**	1,52	2,42	2,96	**1,80**	2,37	1,89	2,08	**0,96**	-	1,49

Рис. 10. Пример анализа публикаций предлагаемой технологией

6.3. Сравнение предлагаемой технологии определения релевантности с технологиями на основе векторных представлений. Определение релевантности текстовых документов с использованием эмбеддингов, как правило, производится на основе вычисления меры косинусного сходства (cosine similarity) между эмбеддингом запроса и эмбеддингами документов, которая определяет косинус угла между ними (что показывает, насколько похожи векторы независимо от их величины).

Для сравнения с предлагаемой технологией использовались три модели контекстных векторных представлений: bge-m3 [21], embeddinggemma [22], nomic-embed-text [23]. Выбор обусловлен богатой практикой использования данных моделей для решения задач ИП и их доступностью. Модели использовались локально без дообучения через платформу Ollama.

Сравнение проводилось на датасете, описанном в разделе 6.1. В начале для запроса и всех документов в датасете была получена оценка их косинусного сходства. При этом для каждой модели, используемой в эксперименте, характерен свой собственный разброс значений метрики на тестовых данных, что продемонстрировано на рисунке 11. Затем для каждой модели был, подобрал подходящий порог релевантности, аналогично предшествующему эксперименту. Зависимости значений метрик точности и полноты от применяемого порога релевантности продемонстрированы на рисунке 12. Таким образом, были выбраны следующие значения порогов: 0,35 (bge-m3), 0,2 (embeddinggemma) и 0,75 (nomic-embed-text). Выбор основан на балансе между количеством истинно положительных и истинно отрицательных оценок. Соответствующие значения метрик приведены в таблице 4.

Модели bge-m3 и nomic-embed-text продемонстрировали результат хуже предлагаемой технологии, поскольку при схожих значениях истинно положительных оценок количество истинно отрицательных оценок у этих моделей значительно ниже. Модель embeddinggemma продемонстрировала схожие, но все еще уступающие результаты. Потенциально модели эмбеддингов могут быть применены для решения данной задачи, но остаются открытыми вопросы выбора подходящей модели, её дообучения (что требует наличия соответствующих ресурсов – как языковых, так и технических) и выбора пороговых значений для выбранной меры оценки релевантности. Также стоит заметить, что дообучение, вероятно, будет производиться не только в начале применения модели, но и в процессе её использования для включения в нее новых данных,

отражающих изменение и развитие социального дискурса в наблюдаемых источниках. Кроме того, как и с любыми методами определения релевантности, основанными на машинном обучении, становится затруднительным объяснение полученных результатов по сравнению с методами, полагающимися на термины и системы знаний.

Рис. 11. Распределение показателей косинусного сходства для тестового датасета по трем моделям

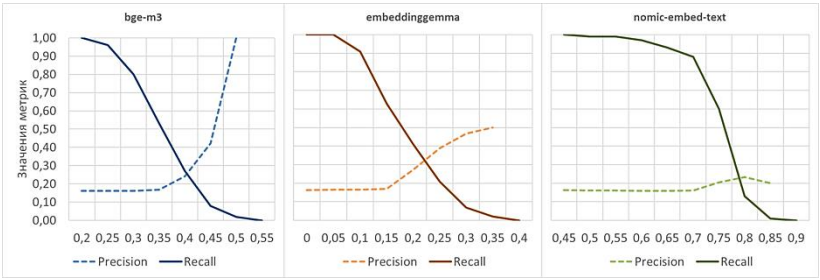

Рис. 12. Зависимость значений метрик точности и полноты от порога релевантности для моделей эмбеддингов

Таблица 4. Результаты сравнения предлагаемой технологии с технологиями на основе векторных представлений

	Многослойный ГЗ	**bge-m3**	**embeddinggemma**	**nomic-embed-text**
TP	50	53	41	60
TN	406	254	408	283
FP	111	263	109	234
FN	50	47	59	40
Precision	**0,31**	0,17	0,27	0,20
Recall	0,50	0,53	0,41	0,60

7. Заключение. В статье рассмотрена проблема повышения эффективности ИП в данных социальных медиа с точки зрения соответствия результатов поиска ожиданиям пользователя (пертинентности результата). Ключевыми проблемами ИП в социальных медиа является динамичность и разнородность знаний, представленных содержанием пользовательской коммуникации с одной стороны и метазадачей поиска, определяющей оценку качества результата – с другой. В качестве решения этих проблем, которые не способны разрешить рассмотренные в разделе 2 подходы (таблица 5), в работе предлагается МДСЗ, интегрирующая в своем составе динамичные разнородные знания и позволяющая приблизить поисковую выдачу к пертинентному результату. В частности, в статье рассмотрены следующие научные и практические результаты:

1. Представлена формализация задачи ИП в социальных медиа, включающая определение концептуальной модели предметной области и коллекции документов, и отражающая свойства мульдоменности и динамичности семантики данных.

2. Разработана модель МДСЗ, одновременно интегрирующая проблемно-ориентированную семантику с помощью онтологии, динамическую доменно-специфическую семантику с помощью многослойного ГЗ и доменно-независимую семантику с помощью лексической БД и большой языковой модели.

3. Предложена технология определения релевантности публикаций с использованием метрики расстояния между терминами пользовательского запроса и терминами публикации в многослойном ГЗ, слои которого соответствуют тематическому распределению коллекции документов.

4. Проведен эксперимент по определению релевантных публикаций с помощью предложенной технологии и её сравнения с технологиями, основанными на полном и частичном поиске вхождения ключевого слова запроса в публикацию и на эмбеддинговых моделях. Сравнительный анализ результатов эксперимента подтверждает эффективность применения разработанной технологии.

В продолжении данного исследования планируется сформировать лексическую БД, расширить функционал использования онтологии в рамках МДСЗ, а также рассмотреть возможность использования не просто меры совместного употребления при построении ГЗ, а точечной совместной информации (pointwise mutual information, PMI) как более точного способа отражения семантических зависимостей ментальной картины пользователей [24]. Кроме того,

в рамках задачи определения релевантности планируется изучить возможность применения алгоритмов распространения и векторных представлений графов [25], таких как node2vec [26].

Таблица 5. Сравнение подходов к определению релевантных документов в рамках ИП

Подход	**Возможность работы с мультидомен-ными данными**	**Возможность работы в условиях динамичности данных**	**Возможность использования дополнительной информации из социальных медиа для повышения точности**
Логический поиск	+	-	-
TF-IDF	+	-	-
Расширение запроса	± (зависит от метода)	-	±
Латентно-семантический анализ	+	-	-
Тематическое моделирование	+	-	-
Глубокое обучение (модели эмбеддингов)	+	-	+
МДСЗ	**+**	**+**	**+**

Литература

1. Hambarde K.A., Proenca H. Information Retrieval: Recent Advances and Beyond // IEEE Access. 2023. vol. 11. pp. 76581–76604. DOI: 10.1109/ACCESS.2023.3295776.1–26.
2. Carpineto C., Romano G. A Survey of Automatic Query Expansion in Information Retrieval // ACM Computing Surveys. 2012. vol. 44. no. 1. pp. 1–50. DOI: 10.1145/2071389.2071390.
3. Azad H.K., Deepak A. Query Expansion Techniques for Information Retrieval: a survey // Information Processing and Management. 2019. vol. 56. no. 5. pp. 1698–1735. DOI: 10.1016/j.ipm.2019.05.009.
4. Russell-Rose T., Gooch P., Kruschwitz U. Interactive query expansion for professional search applications. arXiv:2106.13528. 2021.
5. Zou X. A Survey on Application of Knowledge Graph // Journal of Physics Conference Series. 2020. vol. 1487. no. 1. pp. 1–11. DOI: 10.1088/1742-6596/1487/1/012016.
6. Reinanda R., Meij E., de Rijke M. Knowledge Graphs: An Information Retrieval Perspective. 2020. pp. 1–153. DOI: 10.1561/9781680837292.
7. 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. DOI: 10.1609/aimag.v43i1.19120.
8. Bouadjenek M.R., Hacid H., Bouzeghoub M. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social

information retrieval approaches and platforms // Information Systems. 2016. vol. 56. pp. 1–18. DOI: 10.1016/j.is.2015.07.008.
9. Khalifi H., Dahir S., El Qadi A., et al Enhancing information retrieval performance by using social analysis // Social Network Analysis and Mining. 2020. vol. 10. no. 1. pp. 1–7. DOI: 10.1007/s13278-020-00635-w.
10. Hua Y., et al. Streamlining social media information retrieval for public health research with deep learning // J Am Med Inform Assoc. 2024. vol. 31. no. 7. pp. 1569–1577. DOI: 10.1093/jamia/ocae118.
11. Fellbaum C. WordNet: An Electronic Lexical Database / With a preface by George Miller. Cambridge, MA, USA: MIT Press, 1998. 422 p.
12. Baayen R.H., Piepenbrock R, Gulikers L. CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium, 1995. DOI: 10.35111/gs6s-gm48.
13. Mahajan Y., Freestone M., Aakur S., Karmaker S. Revisiting Word Embeddings in the LLM Era: arXiv:2502.19607. 2025.
14. Park K., et al. The Geometry of Categorical and Hierarchical Concepts in Large Language Models: arXiv:2406.01506. 2025.
15. Nie Z., et al. When Text Embedding Meets Large Language Model: A Comprehensive Survey: arXiv:2412.09165. 2025.
16. Chroma. URL: https://github.com/chroma-core/chroma (дата обращения: 12.09.2025).
17. Wang J., et al. Milvus: A Purpose-Built Vector Data Management System // Proceedings of the 2021 International Conference on Management of Data. Virtual Event China: ACM. 2021. pp. 2614–2627. DOI: 10.1145/3448016.3457550.
18. 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. DOI: 10.15622/ia.23.4.7.
19. Pimeshkov V., Nikonorova M., Shishaev M. Technology for Forming a Multilayer Knowledge Graph to Determine the Relevance of Documents // Digital and Information Technologies in Economics and Management / In: Gibadullin, A. (eds). Cham: Springer Nature Switzerland. 2025. vol. 1422. pp. 134–152. DOI: 10.1007/978-3-031-94273-0_11.
20. Шишаев М.Г., Пимешков В.К., Никонорова М.Л. Утилита формирования структур данных из документов Википедии // Роспатент: Свидетельство о государственной регистрации программы для ЭВМ №2024661261 от 16.05.2024.
21. Chen J., et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation: arXiv:2402.03216. 2024.
22. Vera H. S., et al. EmbeddingGemma: Powerful and Lightweight Text Representations: arXiv:2509.20354. 2025.
23. Nussbaum Z., et al. Nomic Embed: Training a Reproducible Long Context Text Embedder: arXiv:2402.01613. 2025.
24. Zhu C., Liang Y., Liang X., Zhong L., Xie F. Fairness modeling for topics with different scales in short texts // PeerJ Computer Science. 2025. vol. 11. pp. 1–20. DOI: 10.7717/peerj-cs.2936.
25. Choudhary S., Luthra T., Mittal A., Singh R. A Survey of Knowledge Graph Embedding and Their Applications: arXiv:2107.07842. 2021.
26. Grover A., Leskovec J. node2vec: Scalable Feature Learning for Networks: arXiv:1607.00653. 2016.

Пимешков Вадим Константинович — стажер-исследователь, лаборатория информационных технологий управления региональным развитием, Институт

информатики и математического моделирования им. В.А. Путилова — обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Кольский научный центр Российской академии наук» (ИИММ КНЦ РАН). Область научных интересов: обработка естественного языка, извлечение знаний из естественно-языковых текстов. Число научных публикаций — 11. v.pimeshkov@ksc.ru; ул. Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9689.

Никонорова Марина Леонидовна — инженер-исследователь, лаборатория информационных технологий управления региональным развитием, Институт информатики и математического моделирования им. В.А. Путилова — обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Кольский научный центр Российской академии наук» (ИИММ КНЦ РАН). Область научных интересов: анализ естественного языка, извлечение знаний, онтологическое моделирование, машинное обучение. Число научных публикаций — 15. m.nikonorova@ksc.ru; ул. Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9689.

Шишаев Максим Геннадьевич — д-р техн. наук, доцент, главный научный сотрудник, руководитель лаборатории, лаборатория информационных технологий управления региональным развитием, Институт информатики и математического моделирования им. В.А. Путилова — обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Кольский научный центр Российской академии наук» (ИИММ КНЦ РАН). Область научных интересов: информационные системы, региональное развитие, инженерия знаний, искусственный интеллект, машинное обучение, интеллектуальный анализ текстов. Число научных публикаций — 169. m.shishaev@ksc.ru; ул. Ферсмана, 24А, 184209, Апатиты, Россия; р.т.: +7(815)557-9248.

Поддержка исследований. Исследование выполнено в рамках государственного задания ИИММ КНЦ РАН Министерства науки и высшего образования РФ, тема НИР: «Методы и технологии создания интеллектуальных информационных систем для поддержки развития сложных динамических систем с региональной спецификой в условиях неопределенности и риска» (FMEZ-2025-0053). Авторы выражают благодарность Вишнякову Ивану Геннадьевичу, аспиранту и системному администратору ИИММ КНЦ РАН, за выполнение тематического моделирования.

DOI 10.15622/ia.25.3.6

V. PIMESHKOV, M. NIKONOROVA, M. SHISHAEV

INFORMATION RETRIEVAL IN SOCIAL MEDIA BASED ON A MULTI-DOMAIN DYNAMIC KNOWLEDGE SYSTEM

Pimeshkov V., Nikonorova M., Shishaev M. **Information Retrieval in Social Media Based on a Multi-Domain Dynamic Knowledge System.**

Abstract. The task of information retrieval is to find information that best satisfies the user's information needs. In the context of social media, information retrieval is complicated by the high dynamism of content, thematic heterogeneity, and the diversity of users' mental models. This paper proposes an approach to solving the problem of information retrieval under such conditions by constructing a multi-domain dynamic knowledge system. Its novelty lies in the combination of three levels of semantics: problem-oriented, represented by the ontology of the metatask (describing the search objectives); domain-specific, implemented through a dynamic multi-layer knowledge graph built on the basis of user content of social media; and domain-independent, based on a lexical database and a large language model. The knowledge graph allows us to reflect various contexts of concept usage corresponding to thematic clusters in the document collection. Such integration allows us to take into account the evolution of concepts, discourse features, and mental stereotypes of communication participants. To evaluate the effectiveness of the proposed system, an experiment was conducted using a dataset of publications from the VKontakte social network for problem-oriented monitoring of publications, where the selection of relevant publications from non-thematic sources is required. To solve this problem, a technology based on the use of the distance metric between query terms and publication terms in a multi-layer knowledge graph was proposed. The results of the experiment using this technology confirm the effectiveness of the proposed model for information retrieval tasks compared to standard keyword search and embedding models. In continuation of this study, it is planned to create a lexical database and also to consider the possibility of expanding the model by using a measure of pointwise mutual information and graph embedding methods.

Keywords: information retrieval, social media, knowledge systems, multi-layer knowledge graph, relevance.

References

1. Hambarde K.A., Proenca H. Information Retrieval: Recent Advances and Beyond. IEEE Access. 2023. vol. 11. pp. 76581–76604. DOI: 10.1109/ACCESS.2023.3295776.
2. Carpineto C., Romano G. A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys. 2012. vol. 44. no. 1. pp. 1–50. DOI: 10.1145/2071389.2071390.
3. Azad H.K., Deepak A. Query Expansion Techniques for Information Retrieval: a survey. Information Processing and Management. 2019. vol. 56. no. 5. pp. 1698–1735. DOI: 10.1016/j.ipm.2019.05.009.
4. Russell-Rose T., Gooch P., Kruschwitz U. Interactive query expansion for professional search applications. arXiv:2106.13528. 2021.
5. Zou X. A Survey on Application of Knowledge Graph. Journal of Physics Conference Series. 2020. vol. 1487. no. 1. pp. 1–11. DOI: 10.1088/1742-6596/1487/1/012016.
6. Reinanda R., Meij E., de Rijke M. Knowledge Graphs: An Information Retrieval Perspective. 2020. pp. 1–153. DOI: 10.1561/9781680837292.
7. 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. DOI: 10.1609/aimag.v43i1.19120.
8. Bouadjenek M.R., Hacid H., Bouzeghoub M. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms. Information Systems. 2016. vol. 56. pp. 1–18. DOI: 10.1016/j.is.2015.07.008.
9. Khalifi H., Dahir S., El Qadi A., et al Enhancing information retrieval performance by using social analysis. Social Network Analysis and Mining. 2020. vol. 10. no. 1. pp. 1–7. DOI: 10.1007/s13278-020-00635-w.
10. Hua Y., et al. Streamlining social media information retrieval for public health research with deep learning. J Am Med Inform Assoc. 2024. vol. 31. no. 7. pp. 1569–1577. DOI: 10.1093/jamia/ocae118.
11. Fellbaum C. WordNet: An Electronic Lexical Database. With a preface by George Miller. Cambridge, MA, USA: MIT Press, 1998. 422 p.
12. Baayen R.H., Piepenbrock R, Gulikers L. CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium, 1995. DOI: 10.35111/gs6s-gm48.
13. Mahajan Y., Freestone M., Aakur S., Karmaker S. Revisiting Word Embeddings in the LLM Era: arXiv:2502.19607. 2025.
14. Park K., et al. The Geometry of Categorical and Hierarchical Concepts in Large Language Models: arXiv:2406.01506. 2025.
15. Nie Z., et al. When Text Embedding Meets Large Language Model: A Comprehensive Survey: arXiv:2412.09165. 2025.
16. Chroma. Available at: https://github.com/chroma-core/chroma (accessed 12.09.2025).
17. Wang J., et al. Milvus: A Purpose-Built Vector Data Management System. Proceedings of the 2021 International Conference on Management of Data. Virtual Event China: ACM. 2021. pp. 2614–2627. DOI: 10.1145/3448016.3457550.
18. Pimeshkov V., Nikonorova M., Shishaev M. [A Combined Term Extraction Method for the Problem of Monitoring Thematic Discussions in Social Media]. Informatika i Avtomatizatsiya – Informatics and Automation. 2024. vol. 23. no. 4. pp. 1110–1138. DOI: 10.15622/ia.23.4.7. (In Russ.)
19. Pimeshkov V., Nikonorova M., Shishaev M. Technology for Forming a Multilayer Knowledge Graph to Determine the Relevance of Documents. Digital and Information Technologies in Economics and Management. In: Gibadullin, A. (eds). Cham: Springer Nature Switzerland. 2025. vol. 1422. pp. 134–152. DOI: 10.1007/978-3-031-94273-0_11.
20. Shishaev M.G., Pimeshkov V.K., Nikonorova M.L. [Utility for generating data structures from Wikipedia documents]. Rospatent: Svidetel'stvo o gosudarstvennoy registratsii programmy dlya EVM №2024661261 ot 16.05.2024 [Rospatent: Certificate of state registration of a computer program no. 2024661261 dated 05/16/2024]. (In Russ.).
21. Chen J., et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation: arXiv:2402.03216. 2024.
22. Vera H. S., et al. EmbeddingGemma: Powerful and Lightweight Text Representations: arXiv:2509.20354. 2025.
23. Nussbaum Z., et al. Nomic Embed: Training a Reproducible Long Context Text Embedder: arXiv:2402.01613. 2025.
24. Zhu C., Liang Y., Liang X., Zhong L., Xie F. Fairness modeling for topics with different scales in short texts. PeerJ Computer Science. 2025. vol. 11. pp. 1–20. DOI: 10.7717/peerj-cs.2936.
25. Choudhary S., Luthra T., Mittal A., Singh R. A Survey of Knowledge Graph Embedding and Their Applications: arXiv:2107.07842. 2021.

26. Grover A., Leskovec J. node2vec: Scalable Feature Learning for Networks: arXiv:1607.00653. 2016.

Pimeshkov Vadim — Research intern, Laboratory of Information Technologies for Regional Development Management, V.A. Putilov Institute for Informatics and Mathematical Modeling, a separate subdivision of the Federal Research Center "Kola Scientific Center of the Russian Academy of Sciences" (IIMM KSC RAS). Research interests: natural language processing, knowledge extraction from natural language texts. The number of publications — 11. v.pimeshkov@ksc.ru; 24A, Fersman St., 184209, Apatites, Russia; office phone: +7(815)557-9689.

Nikonorova Marina — Research engineer, Laboratory of Information Technologies for Regional Development Management, V.A. Putilov Institute for Informatics and Mathematical Modeling, a separate subdivision of the Federal Research Center "Kola Scientific Center of the Russian Academy of Sciences" (IIMM KSC RAS). Research interests: natural language analysis, knowledge extraction, ontological modeling, machine learning. The number of publications — 15. m.nikonorova@ksc.ru; 24A, Fersman St., 184209, Apatites, Russia; office phone: +7(815)557-9689.

Shishaev Maksim — Ph.D., Dr. Sci., Associate professor, Chief scientific officer, Head of the Laboratory, Laboratory of Information Technologies for Regional Development Management, V.A. Putilov Institute for Informatics and Mathematical Modeling, a separate subdivision of the Federal Research Center "Kola Scientific Center of the Russian Academy of Sciences" (IIMM KSC RAS). Research interests: information systems, regional development, knowledge engineering, artificial intelligence, machine learning, text mining. The number of publications — 169. m.shishaev@ksc.ru; 24A, Fersman St., 184209, Apatites, Russia; office phone: +7(815)557-9248.

Acknowledgements. The study was carried out within the framework of the state assignment of IIMM KSC RAS from the Ministry of Science and Higher Education of the Russian Federation, research topic "Methods and technologies for creating intelligent information systems to support the development of complex dynamic systems with regional specifics in conditions of uncertainty and risk" (FMEZ-2025-0053). The authors express their gratitude to Ivan Vishnyakov, postgraduate student and system administrator of IIMM KSC RAS, for performing the topic modeling.