МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 7** **Москва 2023**

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.75:004.455:004.85

К.И. Белоусов, Р.К. Баширов, Н.Л. Зелянская, И.А. Лабутин,
К.В. Рябинин, Р.В. Чумаков

Профилирование концептуальных систем на основе комплекса методов психосемантики и машинного обучения

«О боже! Я мог бы заключиться в ореховую скорлупу
и считать себя королем необъятного пространства,
если бы не злые сны мои».

Уильям Шекспир*

Представлены новый подход к профилированию пользователей социальных Интернет-сервисов и базирующаяся на нем концепция рекомендательных систем. В основе подхода – интеграция методов и моделей машинного обучения с методами психосемантики и визуальной аналитики, реализация которых осуществляется в виде программного комплекса, включающего три клиент-серверные системы сбора, обработки, анализа и визуализации данных.

* Шекспир У. Гамлет, Принц Датский // Трагедии. Комедии. Сонеты. Коллекционное иллюстрированное издание. – Москва: Алгоритм, 2018. – 976 с.

Ключевые слова: *профилирование пользователей, машинное обучение, метод психосемантической локализации, Word2Vec, текстовые массивы, концептосферы, графовые модели*

DOI: 10.36535/0548-0027-2023-07-1

ВВЕДЕНИЕ

Профилирование пользователей социальных интернет-сервисов – это актуальная научная проблема, связанная с безопасностью [1, 2], политикой и избирательными процессами [3, 4], поведением социальных групп [5, 6] и индивидуальными потребительскими предпочтениями и поведением [7, 8]. Материал для построения профилей довольно широк и включает "статические" и "динамические" подходы к собираемым данным: письменный опрос, стенограммы интервью, анкеты с развернутыми ответами, тексты комментариев, социально-демографические параметры, формы активности в сетевой коммуникации (активный или пассивный участник или наблюдатель), время, проведенное в сети, историю выполненных действий, социальное окружение, количество “друзей”, интересы и предпочтения и др. [9, 10]. Профилирование в том или ином виде лежит в основе создания рекомендательных систем: пользователю предлагаются продукт или контент, созданный на базе обобщенной информации о его поведении в цифровом пространстве [11, 12]. Все эти методы профилирования основываются на сборе информации от пользователей, обычно в форме так называемых "цифровых следов"; при этом, как правило, сам пользователь имеет очень ограниченные возможности в изменении своего профиля. Опросные методы, например, определяют "раз и навсегда" профиль человека по его ответам, будь то интервью или психологический опрос. Методы, основанные на автоматизированном сборе данных, имеют инерционный характер, при котором новая информация о поведении дополняет уже существующую. Тем самым пользователь, с одной стороны, находится под постоянным отслеживанием всех своих действий, а с другой – не имеет возможности изменить часто ошибочные или неточные результаты работы алгоритмов и оказывается в создаваемом самим собою так называемом информационном пузыре. В нашей работе предлагается иной способ профилирования пользователей, названный методом психосемантической локализации, и представляется концепция рекомендательных систем, основанная на возможности пользователя оказывать влияние на результаты работы этих систем.

ОПИСАНИЕ КОНЦЕПЦИИ ПРОФИЛИРОВАНИЯ КОНЦЕПТУАЛЬНЫХ СИСТЕМ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-СЕРВИСОВ: ОБЩИЙ ПОДХОД

Цифровая коммуникация пользователей в социальных интернет-сервисах вследствие отдаления от телесно переживаемой реальности становится для человека реальностью, "основанной на собственных системных правилах и механизмах" [13, с. 83-84]. При этом человек, превращаясь в медиаюзера (пользователя, участника медиакоммуникации), со временем становится субъектом этой отдельной реальности, и представление о нём у других медиаюзеров, а также об алгоритмах/программах складывается на основе оставляемых пользователем цифровых следов.

Цифровые следы связаны с формируемым представлением о человеке (его аватаре) и могут как соответствовать "истинному порядку вещей" (т.е. человеку как индивиду со своими социально-, психо-, концептуально- и биологическими характеристиками), так и представлять совершенно иной конструируемый феномен.

Рациональное поведение пользователя в социальных интернет-сервисах сейчас в б*о*льшей степени рефлексивно, чем поведение в обычной коммуникации (пользователь задумывается о последствиях, которые повлечет за собой его пост/комментарий, лайк, посещение страницы и др.). Рядовой медиаюзер находится под прессом многочисленных (и часто противоречащих друг другу!) запретов, генерируемых государственными институтами, социальными интернет-сервисами и медиасообществом. Результатом такого давления становится "закрытие в концептуальной раковине" человеческой индивидуальности, создание множества ложных цифровых следов, трансляция типичных (внешних) поведенческих фреймов и вытеснение из коммуникативного пространства обсуждения личностно значимых ценностей, мотивов и смыслов. Таким образом, опора на цифровые следы медиаюзера, в том числе декларируемые (например, указание социально-демографических данных), при его профилировании может приводить к существенным ошибкам.

В этой связи становится актуальным поиск новых способов профилирования медиаюзеров. Один из возможных подходов предлагается в настоящей работе. Мы исходим из следующих посылок:

1) коммуникацию в социальных интернет-сервисах можно представить в виде обмена текстовыми (в широком смысле слова) сообщениями;

2) в качестве источников медиатекстов могут выступать как другие медиаюзеры, так и ресурсы, целенаправленно создаваемые медиаагентами (журналистами, блогерами, PR- SMM и другими специалистами). Эти ресурсы мы и будем рассматривать;

3) в процессе электронной коммуникации медиаконтент, генерируемый медиаагентами в виде медиапродуктов, проходит через фильтры, в качестве которых выступают концептуальные системы [14] медиапользователей, в результате чего медиаконтент либо принимается (в том числе с какими-то ограничениями или трансформациями), либо отвергается той или иной концептуальной системой;

4) наиболее значимые концепты (слова или словосочетания, находящиеся в некоем концептуальном пространстве и интерпретируемые ближайшим концептуальным окружением) и связи между ними в концептуальной системе медиаюзера являются ориен-

тирами для поиска медиаконтента. Таким образом помимо **1) концептуальной системы пользователя,** можно выделить **2) медиапродукты как структурированные наборы концептов** (концептуальные домены), отражающие понимание/представление человеком какого-то фрагмента действительности и 3) **концептосферу медиаресурса** (как агрегатора медиаконтента), имеющую множественные структурные связи между концептами.

Значимость конкретных концептов и их структурное окружение, преломленное в медиатекстах, обусловливает тип читателя (пользователя): **в концептосфере ресурса должны присутствовать концептуальные домены в том виде, который пропускают фильтры концептуальной системы пользователя**. Медиапользователь посещает те ресурсы, на которых он ожидает получить контент (структурированный набор концептов определенного концептуального домена), удовлетворяющий его концептуальной системе. Таким образом, профилирование медиаюзеров можно свести к сопоставительному анализу концептосферы ресурса, концептуального домена и концептуальной системы пользователя.

ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ИССЛЕДОВАНИЯ

В нашей работе используются три платформы для сбора и анализа данных, разработанные коллективом авторов. В данном разделе мы представим краткое функциональное описание каждой из них.

Платформа визуальной аналитики SciVi (https://scivi.semograph.com/) – клиент-серверная система для комплексного анализа данных, построенная на принципах микросервисной архитектуры и онтологического инжиниринга [15]. В основе платформы лежит онтологическая база знаний, описывающая набор поддерживаемых операторов загрузки, трансформации, визуализации и анализа данных, а также протоколы взаимодействия этих операторов. Каждый оператор характеризуется набором типизированных входов, выходов и настроечных параметров, а также набором конкретных реализаций на разных языках программирования (чаще всего – JavaScript, TypeScript или Python) для различного вычислительного оборудования. Во время работы каждый оператор функционирует как изолированный микросервис, а платформа SciVi играет роль шины данных и управления, обеспечивая динамическую загрузку операторов и пересылку данных между ними. Для пользователя платформа SciVi предоставляет высокоуровневый графический интерфейс, позволяющий задать цепочку преобразования и анализа данных в виде диаграммы потока данных (Data Flow Diagram, DFD) [16] – графа, вершины которого – это отдельные операторы, а рёбра – связи по данным. DFD здесь выступает, фактически, высокоуровневым визуальным языком программирования для описания алгоритмов обработки данных, а платформа SciVi, соответственно, может быть отнесена к категории low-code платформ [17], обладающих низким порогом вхождения и не требующих от пользователя продвинутых навыков программирования.

Важной особенностью платформы SciVi является её расширяемость и гибкость. При необходимости, новые операторы могут добавляться в состав платформы лишь путём пополнения её базы знаний, без модификации исходного кода ядра. Механизмы исполнения и взаимодействия операторов полностью автоматизированы и не требуют ни от пользователя, ни от разработчика никаких настроек [18]. Таким образом, пополнение платформы SciVi новой функциональностью оказывается процессом очень небольшой трудоёмкости, что позволяет быстро развивать эту платформу, адаптируя её к решению новых аналитических задач.

Редактор когнитивной графики Creative Maps Studio (https://creativemaps.studio/) – клиент-серверное приложение, предназначенное для создания, обработки и анализа цифровых репрезентаций ментальных карт [19]. Приложение имеет модульную архитектуру. Основной модуль – векторный графический редактор, позволяющий создавать двухмерную карту произвольного пространства (можно также создавать различные схемы). Редактор разрешает изображать на карте объекты нескольких типов: а) контуры, б) замкнутые кривые с заливкой (текстура с произвольным цветом), в) точечные объекты – заранее заготовленные векторные изображения, параметры (например, цвет, размер) которых можно изменять в заданных пределах. Объекты любого из типов могут быть названы и описаны в виде текста. Текст может использоваться и как самостоятельный объект. С помощью объектов этого типа можно осуществлять классификацию языкового материала.

Для обработки карт и настройки редактора в нашей работе использовался модуль управления проектами, позволяющий загружать карту, которая будет предустановлена для всех участников исследования, а также скрывать инструменты, которые не нужны для исследования, и устанавливать так называемую награду за выполнение работы. В данном случае, информантам предлагалось после прохождения исследования скачать открытку, на которой изображается финальное состояние карты и результат классификации карты по данным, хранящимся на сервере.

Для классифицирования собранных карт мы использовали модуль аналитики Creative Maps Studio, который представляет собой набор алгоритмов и ядро, управляющее порядком их выполнения и передачей данных между ними по модели DFD (Data Flow Diagram), реализованный по аналогии с тем, как это сделано в платформе визуальной аналитики SciVi.

Платформа ML-моделей – клиент-серверная система, предназначенная для визуализации векторных представлений выбранных концептов (эмбеддингов) определенного концептуального домена в моделях дистрибутивной семантики с использованием графового подхода, реализованного на платформе информационной системы Семограф (https://concept.semograph.com/). Пользовательский интерфейс платформы позволяет делать запросы к хранящимся в репозитории векторным представлениям текстовых массивов социальных сетей, СМИ, художественной литературы, публичных выступлений. В результате запроса генерируются: 1) таблица косинусных расстояний между концептами и 2) таблица встречаемости концептов в текстовом массиве. На рис. 1 представлен скриншот фрагмента результатов запроса к Word2Vec-модели.

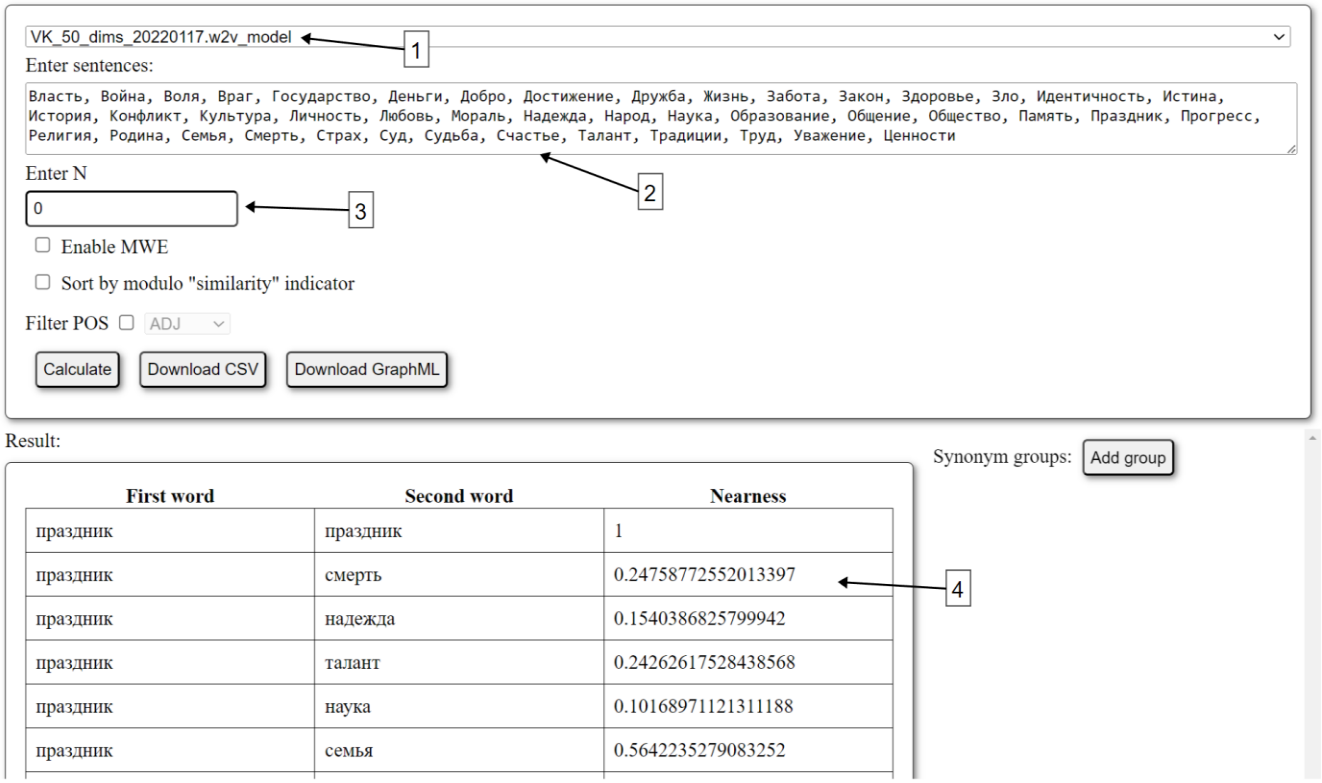

Рис. 1. Скриншот фрагмента результатов запроса к W2V-модели корпуса ВКонтакте: в поле (1) выбирается модель – в нашем случае модель Word2Vec, построенная на основе контента ВК; в поле (2) помещаются концепты-стимулы – в нашем случае 44 слова; в поле (3) выбирается количество дополнительных концептов между стимульными словами. В случае N = 0 в модели остаются только слова-стимулы (дополнительные концепты не включаются); в поле (4) выводится список расстояний между всеми словами-стимулами (на рисунке видно, что одинаковые слова имеют косинусное расстояние, равное 1, т.е. способны полностью замещать друг друга во всех контекстах).

Сгенерированные таблицы выгружаются в виде scv-файла для последующей графовой визуализации в SciVi.

ПРОФИЛИРОВАНИЕ КОНЦЕПТУАЛЬНЫХ СИСТЕМ С ПОМОЩЬЮ МЕТОДОВ ПСИХОСЕМАНТИЧЕСКОЙ ЛОКАЛИЗАЦИИ И МАШИННОГО ОБУЧЕНИЯ

Для выявления способов представления концептуального домена в конкретной концептуальной системе, мы предлагаем метод **психосемантической локализации**. Цель данного метода заключается в получении матрицы семантических расстояний между словами (в качестве единиц анализа могут выступать не только языковые, но и графические, и звуковые объекты). Метод реализован платформе веб-редактора когнитивной графики "Студия креативных карт" (https://creativemaps.studio/).

От информантов требуется расположить слова-стимулы, имеющие социально значимые концепты, в выделенном рабочем пространстве (см. https://vk.cc/cp9JA4). Пространственная близость слов свидетельствует об их субъективной концептуальной близости, а отдаленность – об их удаленности в концептуальной системе индивида, т.е. физическое расстояние в данном случае выполняет роль семантического расстояния. Несомненное достоинство эксперимента в том, что а) все стимулы даны одновременно и при этом не требуется никаких дополнительных действий типа шкалирования и б) каждый из стимулов интерпретирует остальные стимулы.

Информант, выполнив задание, отправляет классифицированный набор концептов на сервер (сохраняет карту).

В модуле аналитики веб-редактора создан оператор, извлекающий координаты расставленных в рабочем пространстве концептов и вычисляющий матрицу семантических расстояний между ними. На рис. 2 в правой его части отражена последовательность операций (реализованных в форме программных средств), направленная на генерацию матрицы семантических расстояний. Видно, что наборы концептов, представляющих концептуальные домены, могут меняться в зависимости от интересов исследователей. При этом сами наборы могут быть как в виде слов/словосочетаний на естественном языке, так и в виде графических символов (векторных иконок или изображений). Информанты используют метод психосемантической локализации, который является одним из методов сбора данных в "Студии креативных карт".

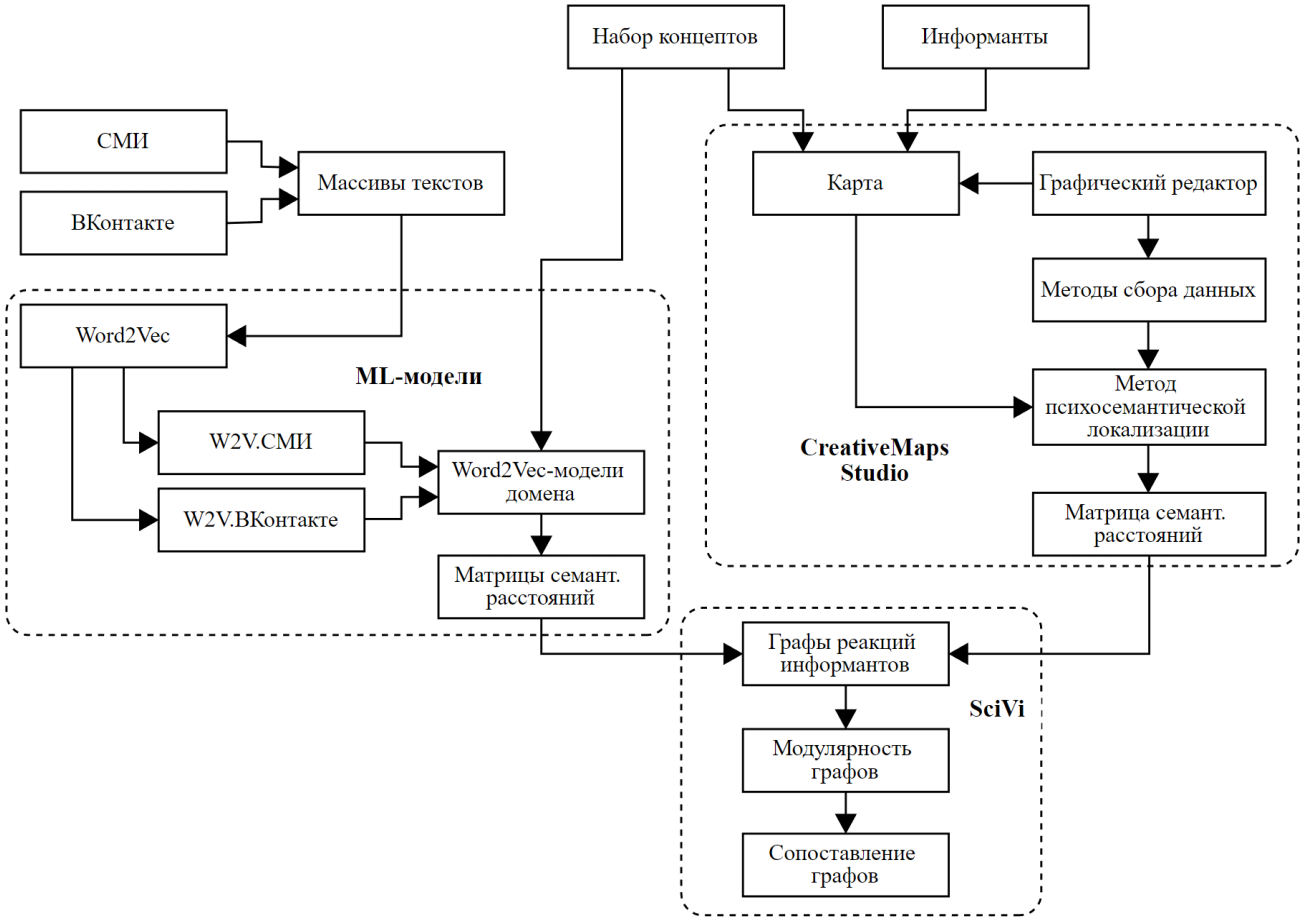

Рис. 2. Схема программного комплекса, включающего три клиент-серверные системы, использованных в исследовании (стрелками обозначена связь по данным)

После генерации матрицы семантических расстояний концептуального домена информанта следует сопоставление его матрицы с другой матрицей, которая может быть как той же природы (матрица другого испытуемого), так и иной природы, относящейся к концептосфере ресурса. Далее будем рассматривать второй вариант.

НАЦИОНАЛЬНАЯ КОНЦЕПТОСФЕРА, ЕЁ ФРАГМЕНТАЦИЯ И МОДЕЛИРОВАНИЕ

Глобальная иерархическая сеть концептов, обладающая полиструктурностью, представляет собой национальную концептосферу [20]. Полиструктурность а) вытекает из понимания концепта как семантической единицы, определяемой (интерпретируемой) своим окружением, и б) обладает множественностью состояний/проекций национальной концептосферы, при которых используются те же самые концепты, но с разной частотой и в различающихся концептуальных окружениях. Условно можно говорить о существовании множества разных концептосфер (например, медийной, обыденной, художественной, Серебряного века, женских романов, юридической, политической, индивидуально-авторской, специализированных интернет-порталов и многое другое), понимая при этом, что все они являются проекциями национальной концептосферы, воплощенными в определенных типах текстовых документов.

Начальный этап моделирования концептосфер – подготовка однородных массивов данных или корпусов. В нашем исследовании использовались: 1) корпус хедлайнов (заголовок и видимая часть текста) российских СМИ (65 млн хедлайнов) и 2) корпус комментариев пользователей социальной сети ВКонтакте (20 млн комментариев).

Одно из возможных решений моделирования концептосфер – построение на основе текстовых корпусов их математических моделей. В настоящее время оптимальными с точки зрения учета контекста слова и вычислительной сложности являются “векторные представления слов” (Word Embeddings). В отличие от ранее используемых методов (BOW или TF-IDF), Word Embeddings отображают текст на векторное пространство значительно меньшей размерности, чем объём словаря. Такое отображение строится с помощью алгоритмов машинного обучения без учителя (unsupervised), как правило, нейросетевых или статистических. Считается, что подобные модели способны определять семантические отношения между словами, и их применение позволяет значительно повышать эффективность решения многих задач компьютерной лингвистики. Примерами по-

добных моделей являются Word2Vec, GloVe (Stanford) и fastText (Facebook).

Модели концептосфер строились по схеме:

1) с использованием библиотек Natasha (https://github.com/natasha/natasha), Razdel (https://github.com/natasha/razdel), Slovnet (https://github.com/natasha/slovnet) и Pymorphy [21] была выполнена предварительная обработка корпусов текстов: а) произведена их токенизация (разбиение на предложения и слова), б) исключены стоп-слова и знаки препинания; в) выполнено приведение слов к нормальной (словарной) форме с учетом их вхождения в словосочетания;

2) далее на основе предобработанных текстовых корпусов с использованием библиотеки Gensim [22] были построены векторные модели по алгоритму Word2Vec [23]. Исходя из соображений относительно компактного размера корпусов, в качестве гиперпараметра размерности векторного пространства было выбрано значение 100.

Созданные Word2Vec-модели концептосфер позволяют эксплицировать (в том числе и в системе визуализации) их участки, состоящие из используемых в эксперименте выбранных концептов. Полученные векторные представления слов (эмбеддинги) можно подвергать математическим или иным операциям, например, вычитанию из одного вектора (например, целевого состояния) другого вектора (например, актуального состояния). Между полученными векторами можно измерить семантическое расстояние, которое будет показывать близость/отдаленность этих слов в концептосфере. В нашем случае по предварительно сформированному списку концептов осуществляется запрос к Word2Vec-моделям, построенным на корпусах хедлайнов СМИ и комментариях пользователей соцсетей, для последующего сопоставления данного фрагмента Word2Vec-моделей с результатами экспериментов, основанных на методе психосемантической локализации.

В качестве метрики расстояния мы использовали косинус угла между векторами, представляющими анализируемые единицы (в нашем случае социально значимые понятия). Полученные матрицы размером N*N (N – количество отобранных концептов) визуализировались в виде графов в приложении SciVi (https://scivi.semograph.com). Порог вхождения вершин и ребер в граф был ограничен: по вершинам > 0,01% вхождений единицы в корпус, по ребрам > 0,5 семантической близости, что соответствует углу, равному 60° между векторами. Близость векторов (показатель косинуса угла между ними) свидетельствует о близости контекстного использования слов в текстах: при величине косинуса угла равной 1 два слова могут полностью замещать друг друга. Заметим, что такая модель отражает не факты совместной встречаемости знаков, а только их контекстуальную близость, обнаруженную на объемном материале.

ВИЗУАЛИЗАЦИЯ WORD2VEC-МОДЕЛЕЙ ФРАГМЕНТОВ КОНЦЕПТОСФЕРЫ

На следующем этапе нашей работы с помощью платформы визуальной аналитики SciVi (https://scivi.tools/) [15, 24] результаты запроса к Word2Vec-моделям имеющихся корпусов были визуализированы в виде графов (рис. 3 и 4), вершинами которых являются используемые в исследовании концепты, а ребрами – показатели близости между ними.

Графически встречаемость каждого концепта в корпусе и показатели близости передаются размером (для вершины) и толщиной (для ребер). С помощью метода модулярности [25] осуществлялось разбиение графа на классы модулярности – подграфы. Затем проводилась автоматическая и ручная укладка графа, в результате чего вершины, относящиеся к одному кластеру, группировались в пространстве и отделялись от вершин, образующих другой кластер. При этом каждый класс визуально отграничен замкнутым контуром и обозначен цифрой. Заметим, что выделенные классы модулярности (кластеры) как фрагменты гиперсети обособлены друг от друга лишь условно: любая из вершин класса может иметь связи с другими вершинами гиперсети, относящимися к разным классам модулярности.

Word2Vec-модель фрагмента корпуса комментариев соцсети ВКонтакте. Представленный на рис. 3 граф разделен на две подсистемы, одна из которых состоит из субъективно значимых концептов (**1 кластер**), а другая из сложной семантической сети общественно значимых понятий (**остальные кластеры**).

Субъективно значимые концепты образуют смысловое пространство «вечных ценностей», существующих помимо и вне актуальных проблем общества: «жизнь», «любовь», «счастье», «семья», «здоровье», «забота», «дружба», «уважение», «общение», «надежда».

Подсистемы связаны через концепты «жизнь» (самый встречаемый) и «любовь» (**1-й кластер**) – «зло» (**2-й кластер**). Но если концепт «жизнь» является своеобразным семантическим «хабом», используемым юзерами в качестве потенциально всеобъясняющего понятия, то связь концептов «любовь» – «зло» отрицает четкое дуальное противопоставление добра и зла, принятое в основных морально-нравственных концепциях.

Социально значимая подсистема подчиняется кластеру, объединившему концепты, указывающие на варианты адаптации личности в социуме (**3-й кластер**): «общество», «религия», «идентичность», «личность», «ценности», «культура», «конфликт», «традиция», «праздник», «талант». Причем собственным атрибутом «личности» в этом процессе является «талант».

Концепты «Общество» и «религия» (понимаемая, скорее, в широком контексте национальной культуры и традиций) становятся определяющими ориентирами для семантического согласования индивидуально-личностных стремлений с институциональными понятиями при встраивании их в историко-культурное, морально-нравственное, правовое и экзистенциальное смысловые пространства.

Системная логика структурирования смыслов ВКонтакте связывает идеи достижений, прогресса с кластером «наука» и «образование», а через него соотносит эти идеи с контекстами не регламентирующей деятельности государства, а культурного развития и религии.

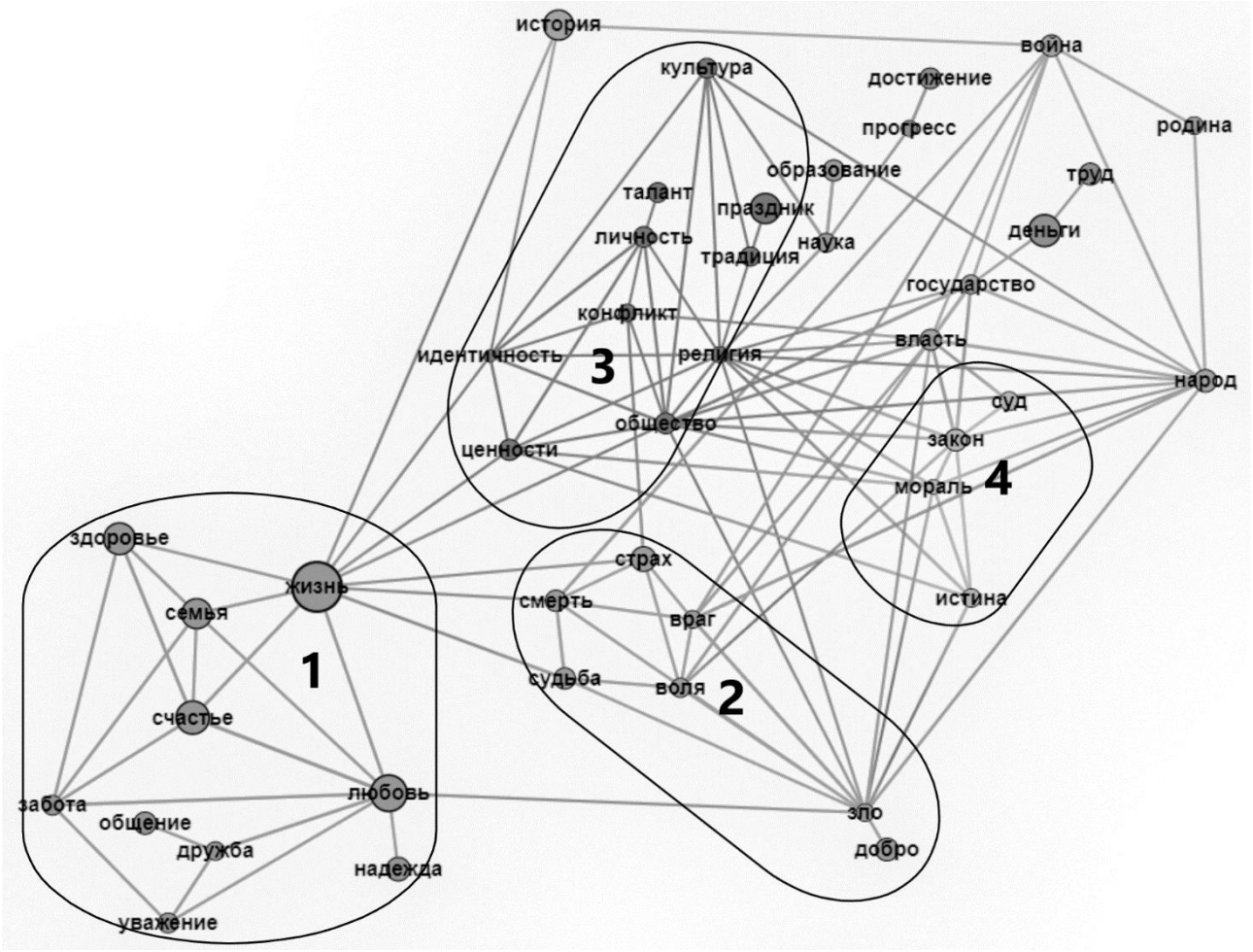

Рис. 3. Визуализация Word2Vec-модели фрагмента корпуса комментариев соцсети ВКонтакте

Противоречия взаимоотношений между концептами «народ» и «власть», «государство» также отражены в Word2Vec-модели, например, в рамках исторического контекста, в котором концепты «война» и «государство»-«власть»-«народ»-«родина» соотносятся с личностными идентификационными процессами и с концептом «жизнь» во всем многообразии ее интерпретаций (см. рис. 3). Кроме того, они образуют многочисленные связи с контекстом морально-нравственных ценностей (**2-й кластер** «воля», «зло», «враг», «смерть», «страх», «судьба», «добро»), важных для пользователей ВКонтакте и в целом, и для осмысления деятельности субъектов власти.

Нравственный контекст становится значимым и для кластера легитимации (**4-й кластер**), который строится не только на юридической составляющей – концепты: «закон» и «суд», – но «мораль» и «истина». А в общем этот кластер тесно переплетен (в силу малосоставности семантически обусловлен) с концептами «религия» и «общество» как носителями морально-нравственных ценностных установок.

Word2Vec-модель фрагмента корпуса хедлайнов федеральных СМИ

Концептуальное пространство федеральных СМИ представляет собой две подсистемы (см. рис. 4), но причиной этого становится достаточно четкая фрагментация общей смысловой системы. Это обстоятельство, должно быть, подтверждает наличие структурированных повесток официальных медиа, задающих направления организации базовых концептов. Соцсети же в основном транслируют доминантные паттерны массового сознания, которые имеют много источников формирования.

Модель демонстрирует единство концептуального пространства СМИ: есть центр (самые активные концепты) и периферия. А «обособившийся» кластер «культура», «образование», «наука», «труд» (**1-й кластер**) в данном случае указывает одновременно и на его периферийность, и на некую отграниченность тем образования, науки и культуры от актуального информационного блока.

Фрагменты подсистемы, объединившей транслируемые федеральными СМИ ценности, просты и непротиворечивы. Кластер «история» (**2-й кластер**) организует концепты, объединяющие историчность, и рассказ об актуальных происшествиях, повседневных историях. Федеральные СМИ демонстрируют потенциальную историчность предоставляемых ими актуальных сведений и встроенность в проверенный временем поток информации.

Концепты «любовь» и «история» оказываются связанными с концептом «жизнь», показывая стремление создавать канонические базовые (проверенные временем/историей) жизненные установки (**3-й кластер**).

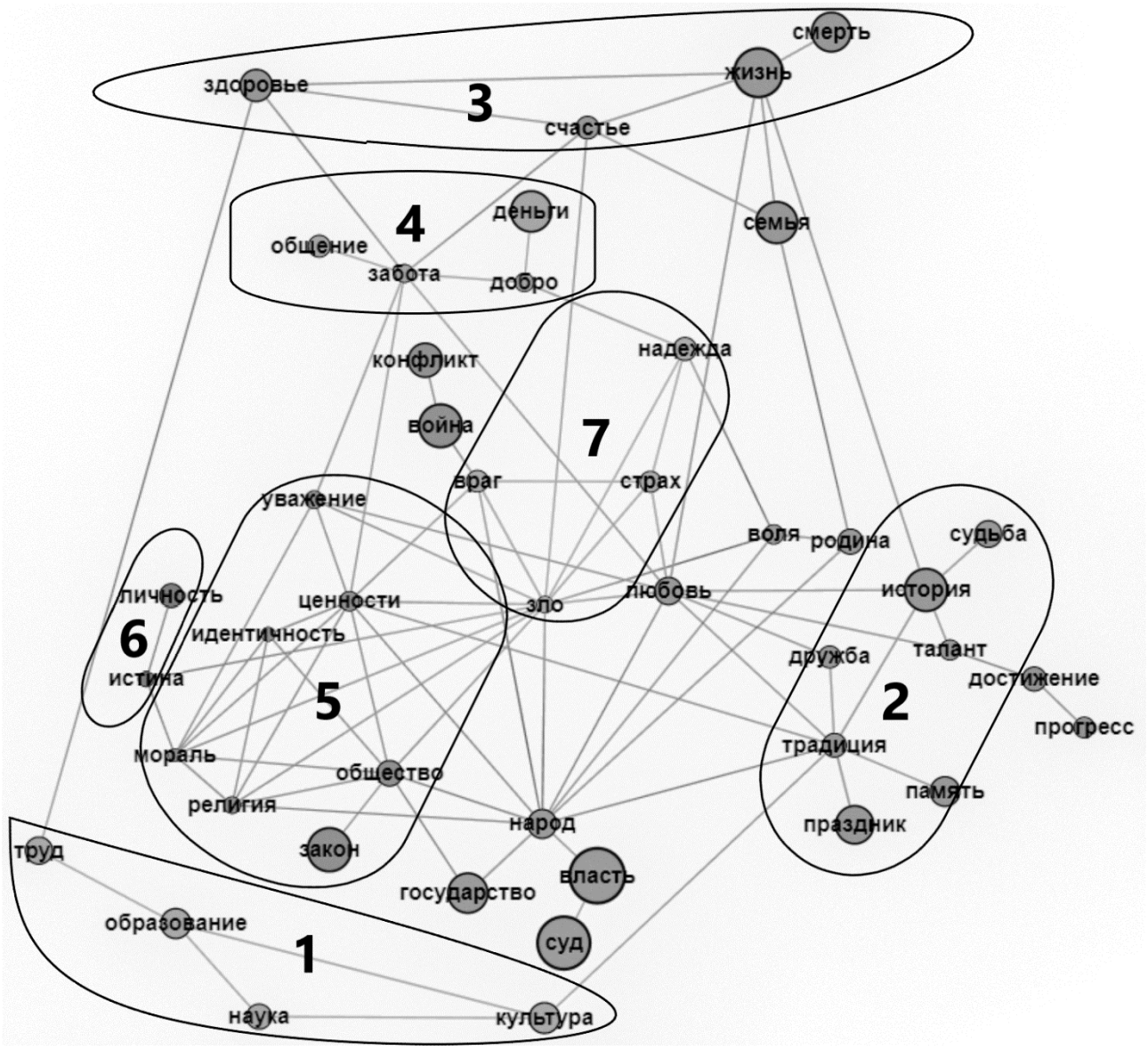

Рис. 4. Визуализация Word2Vec-модели фрагмента корпуса хедлайнов федеральных СМИ

Кластер «заботы» (**4-й кластер**) уточняет базовые понятия жизни, переводя их в контекст человеческого «общения» и установок «добра». Причем добро в федеральных СМИ, в отличие от ВКонтакте, интерпретируемо, проявляется через заботу и финансовую поддержку («деньги» – очень встречаемый концепт).

Кластер «общественно значимые ценности» (**5-й кластер**) объединил ориентиры общественной морали («ценности», «общество», «мораль», «религия», «идентичность», «уважение», «закон»). «Личность» как объект, ведомый к «истине» (**6-й кластер**), вписывается в кластер «общественно значимых ценностей», т.е. СМИ декларируют и констатируют регулируемость личностных стремлений постулатами общественной морали.

Отношения концептов «власть», «государство» и «народ» оказываются регулируемы юстицией («суд») и общественной моралью; между ними нет опосредующих смысловых элементов, свидетельствующих о противоречиях.

В системе четко выделен отрицательный полюс: концепты «зло», «страх», «враг» (**7-й кластер**). Но концепт «зло» и его атрибуты не только определенно названы и зафиксированы, они изначально встроены в рассматриваемую медиасреду как то, что должно быть побеждено: выходят через концепт «надежда» к концепту «добро» и им противостоят активные концепты «любовь», «народ», «ценности», «мораль».

Таким образом, концептуальное пространство пользователей социальной сети ВКонтакте оказывается самым насыщенным с точки зрения образования смысловых связей и возможностей расширения смыслового потенциала понятий. Основным направлением для смыслообразования является адаптация субъективно значимых концептов к общественным нормам. Такая полисемантичность концептосферы медиаюзеров обусловлена большим количеством источников формирования массового сознания – обобщенного субъекта, от которого зависят границы проанализированных понятий.

Концептуальная система, сформированная федеральными СМИ, построена на противопоставлении актуальных и периферийных для медиаповестки смыслов. Потому концептосфера СМИ имеет четкую фрагментацию, которая обусловлена социальной реальностью, конструируемой активными акторами.

СОПОСТАВЛЕНИЕ ФРАГМЕНТОВ КОНЦЕПТУАЛЬНЫХ СИСТЕМ И КОНЦЕПТОСФЕР

Сложность сопоставления результатов психосемантического эксперимента и фрагментов Word2Vec-модели, состоящей из тех же концептов двух концептосфер, состоит в различии метрических инструментов. Так, в психосемантическом эксперименте каждый концепт имеет две координаты (x, y), что позволяет вычислять между любыми парами евклидово расстояние. В Word2Vec-модели между концептами вычисляется косинусное расстояние, поэтому прямое сопоставление с евклидовой метрикой нежелательно. Кроме того в Word2Vec-моделях концепты представлены векторами чисел с плавающей точкой размерности 100, в то время как в модели психосемантической локализации – 2.

Для решения проблемы сопоставления этих моделей использовалось два метода.

1. Из модели Word2Vec извлекались векторы концептов, представляющие собой векторы чисел с плавающей точкой. Далее эти векторы нормировались по длине. Это означает, что все векторы были приведены к единичной длине, чтобы уменьшить влияние на результаты анализа. Затем векторы были преобразованы в трехмерное пространство с использованием метода главных компонентов – PCA. Этот метод позволяет сократить размерность данных, сохраняя при этом основную информацию, и находит главные компоненты, которые объясняют наибольшую часть изменчивости данных, и проецирует данные на эти компоненты. Таким образом, после применения PCA, векторы концептов были представлены в трехмерном пространстве, что позволило визуализировать их и выполнять дальнейший анализ.

Карты концептов пользователей, в свою очередь, преобразовывались путем их наложения на поверхность сферы с использованием математического аппарата полярных координат. В качестве критерия близости карт концептов, полученных вследствие анализа корпусов текстов, и результатов эксперимента использовался логарифм модуля следа псевдообратной матрицы, полученной от произведения матриц расстояний карт:

$$log(|tr((X \times Y') *)|) \quad (1)$$

Скалярное произведение, определенное как след матричного произведения двух матриц, само по себе может использоваться как метрика близости, но оно требует, чтобы матрицы были невырождены (в противном случае произведение также будет вырожденным и след равен нулю), что не выполняется для матриц расстояний, имеющих нулевую диагональ. Поэтому имеет смысл находить от произведения матриц псевдообратную матрицу (она всегда существует и невырождена) и брать в качестве критерия близости ее след. Логарифмирование осуществляется для приведения результатов к более удобному для восприятия и анализа виду. Следует отметить, что большие значения функции соответствуют более схожим матрицам расстояний концептов.

Существенный недостаток данного метода сопоставления двух моделей – сохранение только 10% дисперсии при преобразовании 100-мерного векторного пространства к 3-мерному, поэтому для сопоставления моделей мы использовали другой метод, основанный на графовом представлении и анализе.

2. Графовые визуализации 100-мерной Word2Vec-модели фрагментов концептосфер ВКонтакте и российских СМИ были подробно описаны выше (см. рис. 3 и 4). Графовое представление результатов психосемантической локализации выполнялось на основе подсчета семантических расстояний между концептами на плоскости (чем меньше геометрическое расстояние, тем больше семантическое). При таком преобразовании 2D-распределений в графовую форму информация сохраняется полностью: потери данных связаны с возможной фильтрацией по порогу значения семантического расстояния между вершинами графов.

В итоге мы получаем: а) графы фрагментов концептосфер (Word2Vec-модели) и б) графы фрагментов концептуальных систем информантов (модели психосемантической локализации). При этом очевидна сложность сопоставления двух видов графов в рамках качественного подхода и вызванная этим фактом необходимость количественного сопоставительного анализа графового представления результатов.

Представленные выше интерпретации Word2Vec-моделей фрагментов корпусов комментариев ВКонтакте и хедлайнов российских СМИ основываются на следующих посылках: а) необходимость фильтрации по принятому порогу значения семантического расстояния между вершинами (для избавления от несущественных связей и структурно-семантического шума, мешающего интерпретации) и б) анализ графов с упором на выделяемые с помощью метода модулярности подграфы. Таким образом, принимая за эталон принципы качественной интерпретации, мы приходим к возможности количественного сопоставительного анализа между графами на основе сравнения подграфов, каждый из которых несет собственное концептуальное содержание.

Подобный подход для сопоставления вариантов интерпретации информантами одного текста с последующей кластеризацией этих интерпретаций был реализован нами ранее [26]. Считаем возможным взять те же принципы сопоставления двух и более графов между собой, номинируя представленные в настоящей работе понятия в соответствии с логикой новой предметной области.

1. На первом этапе исследования генерируется таблица $N*M$, где N – выбранные концепты, M – количество подграфов (выделяются с помощью метода

модулярности) во всех построенных графах. При этом каждый подграф дополнительно маркируется как принадлежащий/не принадлежащий графу. На пересечении строки и столбца располагается ячейка с числом, которое может быть либо 0 (отсутствие концепта в подграфе), либо 1 (наличие концепта в подграфе).

Затем между всеми столбцами (подграфами) вычисляются семантические расстояния:

$$r_{ij} = \frac{\sum_{a_{ik}+a_{jk}=2}^{k=n}\left(a_{ik}+a_{jk}\right)}{\sum_{a_{ik}+a_{jk}\geq 1}^{k=n}\left(a_{ik}+a_{jk}\right)} * log_2\left(\sum_{a_{ik}+a_{jk}=2}^{k=n}\left(a_{ik}+a_{jk}\right)\right), \quad (2)$$

где: r_{ij} – семантическое расстояние между i-м и j-м подграфами; a_{ik} и a_{jk} – значения (0 или 1) i-го и j-го подграфов, относящихся к k-му концепту. Иными словами, семантическое расстояние между двумя подграфами понимается равным отношению суммы всех пар значений в том случае, когда оба члена пары имеют ненулевое значение, к сумме всех возможных комбинаций, в которых хотя бы один член пары имеет ненулевое значение. Для увеличения значения близости подграфов при увеличении количества совпадающих вершин вводится повышающий коэффициент, равный логарифму количества совпадающих вершин подграфов.

Семантическое расстояние между двумя графами полагается равным сумме всех семантических расстояний комбинаций подграфов двух графов. Например, в графе G_s есть подграфы SG_{si}, а в графе G_k есть подграфы SG_{kj}. Тогда семантическое расстояние между графами равно (3):

$$r(G_s G_k) = \sum r(SG_{si}\ SG_{kj}) \quad (3)$$

При вычислении семантических расстояний между подграфами двух графов необходимо также определить порог фильтрации. В настоящее время такой порог установлен по значению 0,5, однако в дальнейшем требуется проводить тестирование модели для определения оптимальных диапазонов семантического порога учета близости двух подграфов.

ТЕСТИРОВАНИЕ ИССЛЕДОВАТЕЛЬСКОЙ МОДЕЛИ

Описанная математическая модель сопоставления графов на основе использования метода модулярности была реализована на платформе визуальной аналитики SciVi в виде диаграммы потока данных (рис. 5), представляющих следующий набор действий:

1) загрузка матрицы расстояний между концептами в моделях Word2Vec или психосемантической локализации и построение графа;

2) фильтрация по показателям семантического расстояния между концептами (целевая плотность графа составляла около 16% от всех возможных связей между концептами);

3) применение метода модулярности к графу и выделение в нем кластеров (подграфов) – количество подграфов варьировалось от 8 до 11;

4) выполнение действий пунктов 1 – 3 со всеми анализируемыми графами;

5) вычисление семантических расстояний между всеми графами на основе формул (2) и (3) с последующей визуализацией результатов на материале десяти случайно отобранных ответов информантов и двух графов СМИ и ВКонтакте (рис. 6). На рис. 6 видно, что концептуальные представления информантов распределились по двум кластерам, выстроенным вокруг концептосфер СМИ и ВКонтакте, при этом наиболее близкий (из рассмотренных) для информантов концептуальный профиль ожидаемо оказался ВКонтакте. Концептосфера обыденного сознания (представленная в текстах соцсети) демонстрирует взгляд "как есть", в то время как концептосфера массмедиа формируется под влиянием акторов и отражает желательную (с теми или иными отклонениями) реальность. Сопоставление фрагментов двух концептосфер позволяет более осмысленно подходить к отбору наиболее значимых социальных концептов, учитывать их окружение (в первую очередь, эмоциональное и аксиологическое), а значит, и возможную реакцию на данные концепты.

В то же время некоторые концептуальные представления демонстрируют слабые связи с рассмотренными концептосферами: это позволяет предполагать, что при включении в графовую модель i-й концептосферы, некоторые вершины графа (концептуальные представления) переместились бы в i-й кластер. Отсюда следует необходимость расширения списка концептосфер (и их векторных моделей) для более адекватного профилирования пользователей. При этом цели профилирования могут рассматриваться в рамках данной задачи многоаспектно:

- как определение интересов аудитории для лучшей персонализации сервисов, "чтобы соответствовать требованиям, предпочтениям и потребностям пользователей при предоставлении услуг" [27]. В этом случае речь может идти о фрагментации контента ресурса и создании текстовых массивов, объединенных темой/темами, с последующим их векторным представлением. Таким образом пользователю будет предлагаться контент, наиболее близкий его концептуальному профилю;
- как обнаружение более точных соответствий концептуальному представлению пользователя в виде четко определенных концептосфер, что позволяет на микроуровне маркировать самого пользователя, имеющего определенные интересы, поведение и мировосприятие, а на макроуровне – выделять сообщества таких пользователей.

Заметим, что принципиальное отличие от традиционного рекомендательного подхода, основанного на коллаборативной фильтрации, состоит в возможности пользователя самому задавать (и изменять) свой концептуальный профиль, меняя набор и композицию используемых концептов интересующей его предметной области (например, литературы, кино, спорта, политики, хобби и др.).

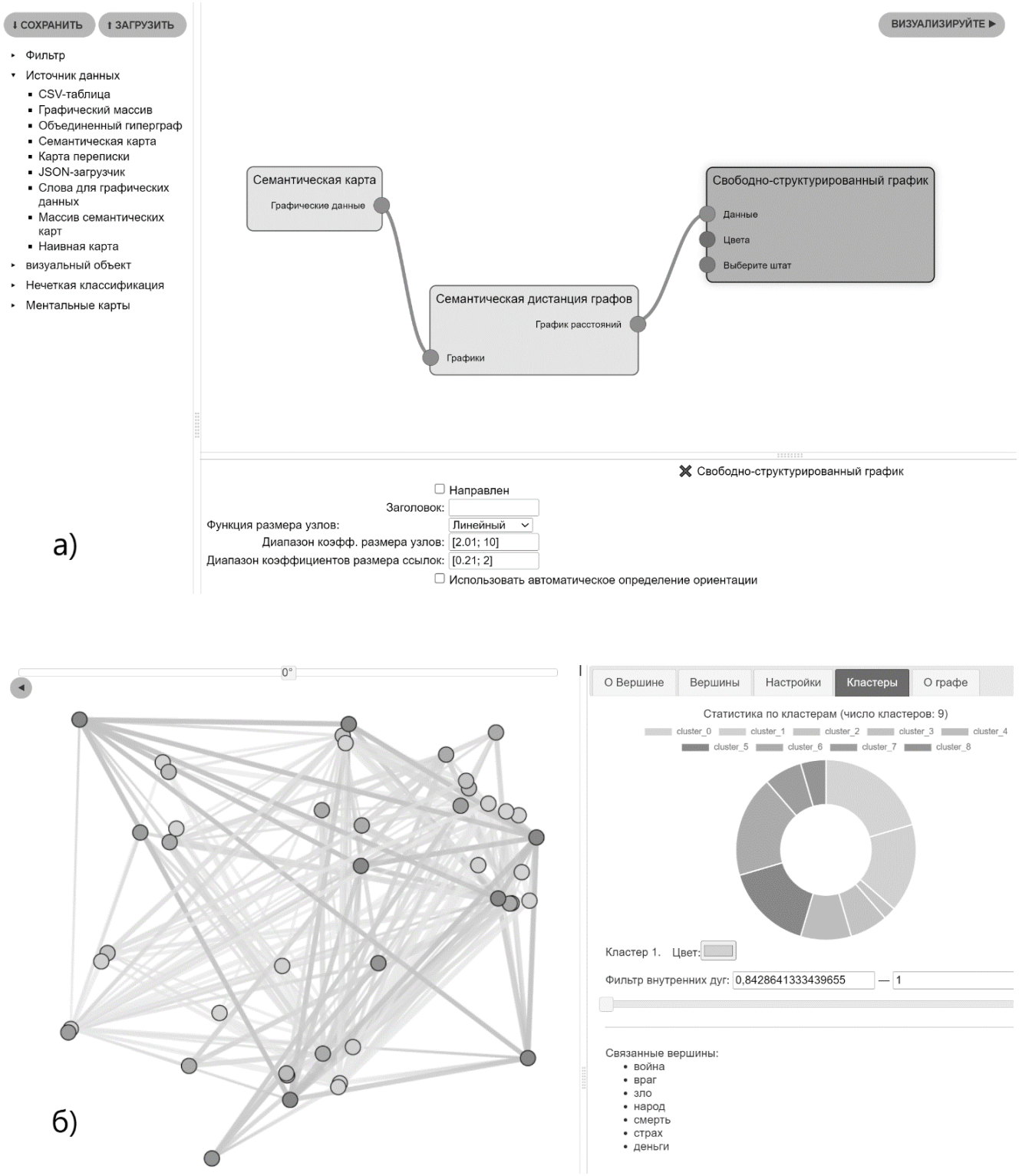

Рис. 5. Вычисление на платформе SciVi подграфов в одном из графов:
а) диаграмма потока данных, состоящая из двух операторов: загрузки данных (Семантическая карта) и визуализации в виде графа свободной структуры; б) результаты редактирования графа: фильтрации ребер (до 16% плотности графа) и модулярности (9 кластеров с описанием включенных вершин). Каждый выделенный подграф (кластер 1, состоящий из вершин *война, враг, зло, народ, смерть, страх, деньги*) передается далее для вычисления семантической близости подграфов.

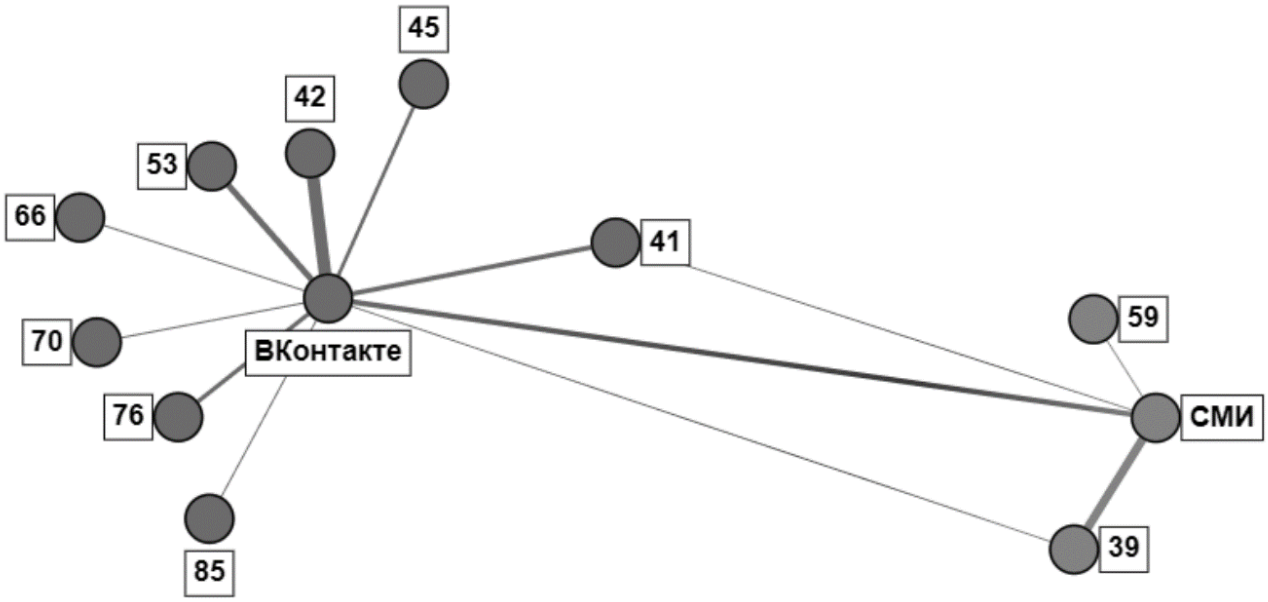

Рис. 6. Кластеризация графов на основе вычисления семантического расстояния между ними
Примечание. Цифрами обозначены графы информантов.

ЗАКЛЮЧЕНИЕ И ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЯ

Представленная в настоящей работе модель профилирования концептуальных систем пользователей интернет-сервисов основывается на следующих базовых положениях.

1. Возможность пользователя изменять свой концептуальный профиль (а вместе с ним и рекомендации сервисных систем).

2. Выдача в качестве результатов запроса таких медиа- или культурных продуктов, которые в наибольшей степени соответствуют поисковому запросу, игнорируя при этом сформированный цифровой след данного пользователя.

3. Ориентация на психосемантические гипотезы пользователя о существовании контента с определенными концептуальными параметрами, что должно, например, побуждать читателей находить неожиданную для себя литературу, выходить за пределы своей уже сформированной читательской "раковины" в поисках новых жанров, смыслов и художественной эстетики в пространстве того, что мы вслед за Х.Л. Борхесом образно называем "Вавилонской библиотекой".

4. Комбинирование методов психосемантики, связанных с репрезентациями концептуальных систем пользователей, и методов машинного обучения, генерирующих модели концептосфер, репрезентированных в массивах текстового контента.

Положение 4 нуждается в комментарии, поскольку с ним в большей мере связано дальнейшее развитие представленной концепции. Так, следует отметить, что необходимая для профилирования пользователя сопоставляемая с его концептуальным представлением модель концептосферы может быть получена разными способами. Это может быть не только рассмотренное здесь векторное представление концептов, но и векторное представление документов и тем (результат тематического анализа документов). Так, в целевом понимании концепции пользователь, чье концептуальное представление наиболее близко определенному жанру художественной литературы (векторная модель жанра строится на основе массива текстов данного жанра), может получить не только жанровую рекомендацию, но и книги, представляющие сами жанры. При этом тексты могут быть представлены и в виде векторной модели документа, и как комбинация выделенных тем в их векторном представлении. Отсюда следует, что успешная реализация концепции требует создания объемных библиотек векторных моделей концептов, документов и тем, воплощенных в текстах разных сфер деятельности человека.

СПИСОК ЛИТЕРАТУРЫ

1. Knoke D. Emerging Trends in Social Network Analysis of Terrorism and Counterterrorism // The Individual and Society Social Networks. – 2015. – P. 1-15. DOI: https://doi.org/10.1002/9781118900772. etrds0106.
2. Teutsch D., Niemann Ju. Social Network Sites As A Threat To Users' Self-Determination And Security: A Framing Analysis Of German Newspapers // The Journal of International Communication. – 2016. – Vol. 22, № 1. – P. 22-41. DOI: 10.1080/ 13216597. 2015.1111841
3. Borah P. Political Facebook Use: Campaign Strategies Used in 2008 and 2012 Presidential Elections // Journal of Information Technology and Politics. – 2016. – Vol. 13, № 4. – P. 326-338. DOI: 10.1080/ 19331681.2016.1163519
4. Fernandez J., Llopis F., Gutierrez Y., Martınez-Barco Patricio, Dıez A. Opinion Mining in

Social Networks versus Electoral Polls // Proceedings of Recent Advances in Natural Language Processing. – 2017. – P. 231–237. DOI: https://doi.org/10.26615/978-954-452-049-6_032

5. Gustafsson N. The Subtle Nature of Facebook Politics: Swedish Social Network Site Users and Political Participation // New Media and Society. – 2012. – Vol. 14, № 7. – P. 1111-1127. DOI: 10.1177/1461444812439551.
6. Heiss R., Matthes Jörg. Who 'likes' Populists? Characteristics of Adolescents Following Right-Wing Populist Actors on Facebook // Information, Communication and Society. – 2017. – Vol. 20, № 9. – P. 1408-1424. DOI: 10.1080/1369118X.2017.1328524
7. Ko E., Chun E., Song S., Mattila P. Exploring SNS as a consumer tool for retail therapy: explicating semantic networks of "shopping makes me happy (unhappy)" as a new product development method // Journal of Global Scholars of Marketing Science. – 2015. – Vol. 25, № 1. – P. 37-48. DOI: 10.1080/21639159.2014.984891
8. Gil de Zuniga H., Diehl T., Huber B., Liu J. Personality Traits and Social Media Use in 20 Countries: How Personality Relates to Frequency of Social Media Use, Social Media News Use, and Social Media Use for Social Interaction // Cyberpsychology, Behavior, And Social Networking. – 2017. – Vol. 20, № 9. – P. 540-552. DOI: 10.1089/cyber.2017.0295
9. Gena C. Methods and techniques for the evaluation of user-adaptive systems // The Knowledge Engineering. – 2005. – Vol. 20, № 1. – P. 1-37.
10. Khosrow-Pour M. Encyclopaedia of information science and technology, Electron. –Hershey: PA Idea Group Reference, 2005. – P. 2063-2067.
11. Elizarov A.M., Lipachev E.K., Zhizhchenko A.B., Zhil'tsov N.G., Kirillovich A.V. Doklady Mathematics. – 2016. – Vol. 93, № 2. – P. 231-233.
12. Filipyev A.V. Item-based recommender system with statistical learning for unauthorized customers // Software & Systems. – 2019. – Vol. 32, № 2. – P. 221–226. DOI: 10.15827/0236-235X.126.221-226
13. Langer R. Towards a constructivist communication theory? Report from Germany // Nordicom information. – 1999. – № 1-2. – P. 75-86.
14. Павиленис Р.И. Проблема смысла: современный логико-философский анализ языка. – Москва: Мысль. – 1983. – 286 с.
15. Ryabinin K., Chumakov R., Belousov K., Kolesnik M. Ontology-Driven Visual Analytics Platform for Semantic Data Mining and Fuzzy Classification // Frontiers in Artificial Intelligence and Applications. – 2022. – Vol. 358: Fuzzy Systems and Data Mining VIII. – P. 1–7. DOI: 10.3233/FAIA220363
16. Sousa T.B. Dataflow Programming Concept, Languages and Applications // Doctoral Symposium on Informatics Engineering. – 2012. – Vol. 130.
17. Pinho D., Aguiar A., Amaral V. What about the usability in low-code platforms? A systematic literature review // Journal of Computer Languages. – 2023. – Vol. 74. – 101185. DOI: 10.1016/j.cola.2022.101185
18. Ryabinin K., Chuprina S., Labutin I. Tackling IoT Interoperability Problems with Ontology-Driven Smart Approach // Lecture Notes in Networks and Systems. – 2022. – Vol. 342. – P. 77-91. DOI: https://doi.org/10.1007/978-3-030-89477-1_9
19. Chumakov R., Ryabinin K., Belousov K., Duan J. Creative Map Studio: A Platform for Visual Analytics of Mental Maps // Scientific Visualization. – 2021. – Vol 13, No. 2. – P. 79 – 93. DOI: 10.26583/sv.13.2.06.
20. Белоусов К.И., Баранов Д.А., Зелянская Н.Л., Пономарев Н.Ф., Рябинин К.В. Когнитивно-информационное моделирование социальной реальности: концепты, события, приоритеты // Вестник Томского государственного ун-та. Филология. – 2021. – No 72. – С. 5-26. DOI: 10.17223/ 19986645/72/1
21. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. – 2015. – P. 320-332.
22. Řehůřek R., Sojka P. Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010. – P. 46-50. – URL: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf
23. Mikolov T., Sutskever I., Chen K., et al. Distributed Representations of Words and Phrases and Their Compositionality // Proceedings of the 26th International Conference on Neural Information Processing Systems. – 2013. – P. 3111-3119. – URL: https://proceedings.neurips.cc/paper_files/paper/ 2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
24. Белоусов К.И., Обухова И.А., Лабутин И.А. WORD2VEC и графосемантические модели использования эмотиконов и эмодзи в текстах интернет-коммуникации // Вопросы когнитивной лингвистики. – 2023. – No 2. – С. 47-62. DOI: 10.20916/1812-3228-2023-2-47-62
25. Lambiotte R., Delvenne J.-C., Barahona M. Laplacian Dynamics and Multiscale Modular Structure in Networks. – 2009. – URL: https://arxiv.org/pdf/0812.1770.pdf (дата обращения: 15.02.2018).
26. Белоусов К.И. Стратегии структурирования тематического пространства текста // Вестник Пермского университета. Российская и зарубежная филология. – 2014. – No 4. – С. 15-25.
27. Cufoglu A. User Profiling - A Short Review // International Journal of Computer Applications. – 2014. – Vol. 108, No 3. – P. 1-9.

Материал поступил в редакцию 18.05.23.

Сведения об авторах

БЕЛОУСОВ Константин Игоревич – доктор филологических наук , профессор кафедры теоретического и прикладного языкознания, Пермский государственный национальный исследовательский университет
e-mail: belousovki@gmail.com

БАШИРОВ Рустам Каримович – аспирант, инженер научно-исследовательской лаборатории социокогнитивной и компьютерной лингвистики, Пермский государственный национальный исследовательский университет
e-mail: bashirov_rustam@bk.ru

ЗЕЛЯНСКАЯ Наталья Львовна – кандидат филологических наук, доцент кафедры журналистики и массовых коммуникаций, Пермский государственный национальный исследовательский университет
e-mail: zelyanskaya@gmail.com

ЛАБУТИН Иван Александрович – аспирант, ассистент кафедры математического обеспечения вычислительных систем, Пермский государственный национальный исследовательский университет
e-mail: linux.rf@gmail.com

РЯБИНИН Константин Валентинович – кандидат физико-математических наук, доцент кафедры математического обеспечения вычислительных систем, Пермский государственный национальный исследовательский университет
e-mail: kostya.ryabinin@gmail.com

ЧУМАКОВ Роман Владимирович – аспирант, ассистент кафедры математического обеспечения вычислительных систем, Пермский государственный национальный исследовательский университет
e-mail: chumakoff.r.v@gmail.com