Научная визуализация, 2018, том 10, номер 3, страницы 1 - 10, DOI: 10.26583/sv.10.3.01

Средства визуальной аналитики для комплексного исследования результатов многопараметрического описания пользователей социальных интернет-сервисов

К.В. Рябинин[1*], К.И. Белоусов[2*], С.И. Чуприна[3*], С.А. Щебетенко[4+*], С.С. Пермяков[5*]

[*] Пермский государственный национальный исследовательский университет, Пермь, Россия
[+] Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

[1] ORCID: 0000-0002-8353-7641, kostya.ryabinin@gmail.com
[2] ORCID: 0000-0003-4447-1288, belousovki@gmail.com
[3] ORCID: 0000-0002-2103-3771, chuprinas@inbox.ru
[4] ORCID: 0000-0001-5790-9731, shebetenko@rambler.ru
[5] ORCID: 0000-0002-8486-779X, rewmad@gmail.com

Аннотация

Статья посвящена вопросам применения методов и средств визуальной аналитики для систематизации результатов многопараметрического описания пользователей социальных интернет-сервисов. В данной статье под многопараметрическим описанием понимается описание языковых характеристик письменной речи пользователей социальных интернет-сервисов, извлекаемых из написанных ими комментариев и сообщений, а также психологические и социальные характеристики, извлекаемые из указанных в их профилях сведений, и из результатов пройденных ими социологических и психологических опросов. Предлагаются методы и средства визуальной аналитики, которые позволяют, во-первых, наглядно представить выявленные взаимосвязи между разными параметрами пользователей, а во-вторых, без повторения исходного эксперимента выдвинуть и проверить гипотезы с использованием исключительно средств визуального интерактивного анализа и, как следствие, обнаружить новые закономерности в данных. В качестве системы сбора и предварительной обработки исследуемых данных предлагается использовать информационную систему для лингвистических исследований Семограф, а в качестве программного средства визуальной аналитики – построенную на методах онтологического инжиниринга адаптивную мультиплатформенную систему научной визуализации SciVi.

В основе описываемых средств визуальной аналитики лежит использование графовой модели представления данных, ввиду того, что исследуемые данные обладают высокой степенью связности. Реализованы два варианта графов – круговой и граф свободной структуры. Для обоих видов графов разработаны соответствующие модули визуализации в составе системы SciVi, обеспечивающие необходимую функциональность наглядного представления и анализа данных. В статье описаны основные интерактивные возможности взаимодействия с визуальным представлением указанных графов и особенности их реализации. В частности, описаны различные способы фильтрации вершин и связей графов, а также возможность переключения между различными срезами данных посредством дискретных шкал состояний.

Разработанные средства визуальной аналитики протестированы на реальных данных – результатах психологического опроса пользователей социальной сети ВКонтакте и результатах анализа их речевого поведения.

Ключевые слова: визуальная аналитика, онтологический инжиниринг, графы, языковые параметры, психологические характеристики, BFI, параметры пользователей социальных сетей.

1. Введение

Бурное развитие сервисов интернета, социальных сетей, мессенджеров, игровой среды, технологий виртуальной и дополненной реальности, интернета вещей ставит задачу комплексного многопараметрического описания и типологизации пользователей социальных интернет-сервисов (англ. *Social Network Services*, SNS) с целью выявления закономерностей в их поведении. Комплексный многопараметрический подход, на наш взгляд, должен опираться на принципы:

– системного характера когниции, предполагающего взаимосвязь и взаимодействие всех компонентов когнитивной системы человека и проявляющегося как в языковой/речевой, так и в неязыковой деятельности человека;
– обусловленности когнитивных структур социальным опытом индивида.

Такой подход, в свою очередь, позволяет рассматривать результаты комплексного описания пользователей SNS в качестве модели интеграции социального, поведенческого, психологического и языкового компонентов личности.

В качестве социальных параметров нами рассматриваются данные из профиля пользователя (пол, возраст, образование, сфера интересов, социальное окружение и др.); в качестве поведенческих – предпочтения (например, отмеченные как понравившиеся публикации и др. материалы, размещаемые в сети) и т. п. Психологические параметры выявляются в результате психологического опроса, а языковые – на основе анализа комментариев пользователей. В роли психологического опросника использовалась русская версия «Вопросника Большой Пятёрки» (BFI – Big Five Inventory) [1, 2]. Адаптированная русскоязычная версия (автор С.А. Щебетенко) была успешно апробирована [3]. Каждая из психологических характеристик личности описывается с помощью пятибалльной шкалы проявления двух противопоставленных признаков: «++» – максимальное проявление признака, «+» – значимое проявление признака, «0» – признак не выражен. Например, экстраверсия/интроверсия информанта может описываться как сильно выраженная экстраверсия («экстраверсия++»), или выраженная экстраверсия («экстраверсия+»), или невыраженная экстраверсия/интроверсия («0»), или выраженная интроверсия («интроверсия+»), или сильно выраженная интроверсия («интроверсия++»).

Материал исследования представляет собой данные профилей участвовавших в психологическом опросе пользователей и их тексты в социальной сети ВКонтакте (https://vk.com). Для сбора информации из социальной сети ВКонтакте был использован API этой социальной сети (программный интерфейс, который позволяет получать информацию из базы данных с помощью HTTP-запросов к соответствующему серверу). Стандартные средства API ВКонтакте при определённых условиях позволяют собирать данные из профилей пользователей, однако не предоставляют возможности получить все его комментарии при помощи одного запроса. Эта проблема решается путём автоматического перебора комментариев к записям на личных страницах пользователя и его друзей, а также проверкой их авторства. Все полученные сведения были автоматически обезличены и собраны в одной базе. Общий объем материала – 19161 автоматизировано собранных реплик 340 пользователей, прошедших психологический опрос.

На следующем этапе после загрузки этих данных в информационную систему (ИС) лингвистических исследований Семограф, осуществлялась их экспертная классификация. Для лингвистиче-

ского анализа материала был разработан многоуровневый (иерархический) классификатор, учитывающий такие языковые параметры, как дейктические показатели, модальность, субъективно-оценочные значения, использование эмотиконов, бранной лексики и др. [4]. Процедура классификации состояла в приписывании каждой реплики к определённым ячейкам классификатора на основании представленности в данной реплике определённого языкового параметра.

При проведении классификации компонентов в проекте соблюдаются следующие принципы:

1) классификация проводится несколькими экспертами (в процессе классификации вырабатывается согласованная позиция всех экспертов по спорным вопросам);
2) каждое поле (класс) формируется рядом лингвистических единиц, которые обладают общим признаком (данный признак может иметь любую природу, как лингвистическую – грамматическую, семантическую, синтаксическую, стилистическую, текстовую и т. д., так и экстралингвистическую);
3) одна реплика может быть отнесена к нескольким полям (если она включает несколько лингвистических единиц, например, одновременно бранную лексику и эмотиконы).

Таким образом, один и тот же материал рассматривается экспертами с различных точек зрения и в процессе анализа создаётся его многопараметрическая лингвистическая классификация.

На следующем этапе результаты полевого анализа реакций информантов обрабатываются с помощью инструментария ИС Семограф. При этом автоматически вычисляются объёмы выделенных полей; семантические карты, отражающие совместную встречаемость полей в репликах; таблицы сопряжённости, передающие распределение выделенных полей в зависимости от социальных и психологических параметров информантов. Полученные результаты экспортируются в модуль визуального анализа, созданный на базе разработанной ранее адаптивной мультиплатформенной системы научной визуализации SciVi [5].

На Рис. 1 представлена общая схема этапов проведённого исследования:

– психологический опрос 821 информанта;
– сбор данных (активность, социальные параметры и комментарии) участвовавших в опросе пользователей социальной сети ВКонтакте;
– экспертный анализ пользовательского контента, осуществлённый в ИС Семограф;
– визуальная аналитика полученных результатов, направленная на выявление взаимосвязей между психологическими характеристиками пользователей социальной сети и речевыми параметрами их поведения.

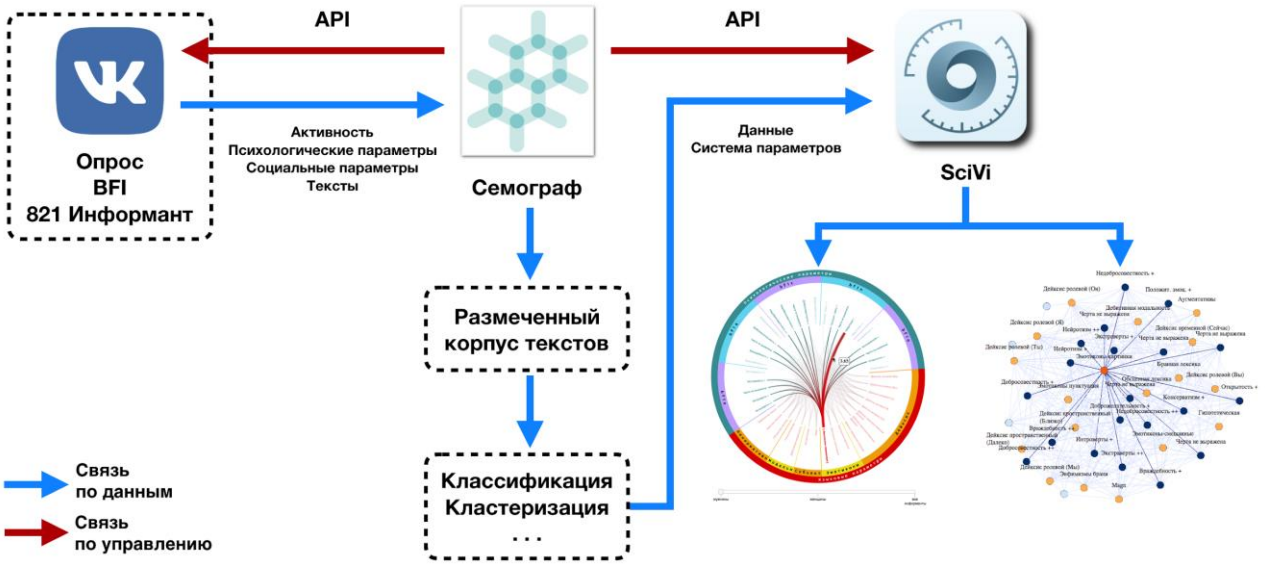

Рис. 1. Этапы проведения исследования на базе интегрированного использования платформ Семограф и SciVi.

На Рис. 1 видно, что решение исследовательских задач (кластеризации, классификации, выявления корреляций между параметрами, установления закономерностей распределения данных, проверки гипотез и др.) возможно как непосредственно в среде ИС Семограф, так и после обработки данных средствами визуальной аналитики системы SciVi. Первая часть результатов является следствием создания классификатора контента и может непосредственно использоваться при решении широкого круга задач, в т. ч. задач машинного обучения. Вторая часть результатов относится к области выявления корреляций параметров и установления закономерностей распределения данных. Инструменты визуальной аналитики в данном случае позволяют систематизировать большой объём взаимосвязанных данных, зависимости между которыми установлены на этапе обработки и анализа в ИС Семограф. Интерактивные возможности средств визуальной аналитики, включая использование семантических фильтров, позволяют выдвигать и проверять гипотезы о зависимостях между данными без их повторной обработки, и, как следствие, появляется возможность обнаружить новые закономерности.

2. Онтологический инжиниринг в визуальной аналитике

В ходе предыдущих исследований были проанализированы наиболее популярные системы научной визуализации (такие как TecPlot, ParaView, Avizo, VizIt и др.) и установлено, что самым серьёзным недостатком большинства из них является отсутствие высокоуровневых средств адаптации к нестандартным задачам [6]. Нестандартность задач визуализации может выражаться как в особенностях источника данных, что требует использования специальных алгоритмов доступа или представления данных в некотором нестандартном формате, так и в специфических требованиях к отображаемым графическим объектам и сценам.

Один из возможных путей для обеспечения высокой гибкости программных средств научной визуализации – использование при их создании модельно-ориентированного подхода. Система, поведение которой полностью или хотя бы частично управляется некоторой декларативной формальной моделью, может быть легко перенастроена для наилучшего соответствия требованиям решаемых задач. В нашем подходе в роли такой модели выступает пред-

метная онтология – формальная модель соответствующей предметной области, включающая в себя тезаурус концептов, множество связей и набор аксиом, определённых на этих концептах и связях [7]. Достоинство онтологий заключается в том, что они интегрируют в себе преимущества логических и графических моделей представления знаний, имеют стандарты описания, одинаково легко читаемы как человеком, так и программными агентами, самодокументируемы и повторно используемы. В состав базы знаний системы визуализации включена онтология визуальных объектов и графических сцен, описывающая поддерживаемые системой средства визуализации, а также онтология семантических фильтров, описывающая допустимые способы трансформации входных данных.

Для решения специализированных задач используются дополнительные онтологии. Например, если источником подлежащих визуализации данных выступает некоторый решатель (англ. *Solver* – расчётная программа или программно-аппаратный комплекс), в состав системы визуализации включается онтология синтаксических конструкций ввода/вывода языка программирования, на котором написан этот решатель. Такая онтология служит для целей автоматической генерации синтаксического анализатора, задачей которого является извлечение из исходного кода решателя структуры его выходных данных и управляющих параметров. Это, в свою очередь, позволяет автоматизировать процесс настройки системы визуализации на взаимодействие с решателем. В случае необходимости встраивания системы визуализации в некоторое аппаратное устройство, например, в какой-либо элемент экосистемы Интернета вещей [8], используется онтология электронных компонентов [9]. Она служит для автоматической генерации прошивки электронного устройства и интеграции в эту прошивку кода визуализации.

Принципы построения систем научной визуализации, описанные авторами в [5], были реализованы при разработке упомянутой выше мультиплатформенной системы SciVi. Методы и средства онтологического инжиниринга, использованные при создании SciVi, хорошо зарекомендовали себя на практике в контексте обеспечения высокой настраиваемости и адаптивности системы визуализации к специфике решаемых задач и индивидуальным предпочтениям пользователей, а также унификации процесса пополнения функциональности системы [9]. Так, например, добавление поддержки новых механизмов рендеринга сводится к пополнению онтологии визуальных объектов и графических сцен описанием новой функциональности и ссылками на внешние модули, реализующие эту функциональность. При этом нет необходимости в модификации исходного кода ранее отлаженных функций и ядра системы.

Зачастую, помимо наглядного представления научных данных, требуется также их глубинный анализ, что влечёт за собой необходимость выполнения некоторых преобразований над данными. Например, фильтрацию данных в соответствии с заданными критериями (пороговыми функциями и т. п.), математические преобразования (масштабирование, нормализация и т. п.), классификацию и кластеризацию, статистический анализ и т. д. Для поддержки и унификации такого рода трансформаций предлагается использовать механизм т. н. *семантических фильтров*, которые представляют собой операторы преобразования подлежащих визуализации данных. Описание алгоритма работы этих операторов, соответствующих входов, выходов и настроечных параметров хранится в онтологии семантических фильтров. Это позволяет пополнять набор поддерживаемых семантических фильтров путём внесения изменений только в базу знаний системы научной визуализации с помощью средств высокоуровневого пользовательского интерфейса.

Семантические фильтры допускают суперпозицию, задаваемую диаграммой потока данных [10]. В системе SciVi при-

сутствует специальный высокоуровневый графический редактор для составления таких диаграмм. Вершинами в них выступают источники данных, семантические фильтры, визуальные объекты и графические сцены, а связи представляют пути передачи данных. Палитра допустимых вершин формируется автоматически на основе соответствующих онтологий. Используя доступные в палитре фильтры пользователь может самостоятельно создать диаграмму, описывающую процесс предобработки (фильтрации) и визуализации данных, и тем самым в динамике настроить систему SciVi на конкретную задачу визуального анализа.

Наличие расширяемого набора семантических фильтров, а также удобных средств для их комбинирования, наряду с другими описанными в данной статье возможностями, превращает систему научной визуализации SciVi в полноценное средство визуальной аналитики [6].

3. Онтологический профиль анализируемых данных

Принципиальная схема взаимодействия систем SciVi и Семограф представлена на Рис. 2.

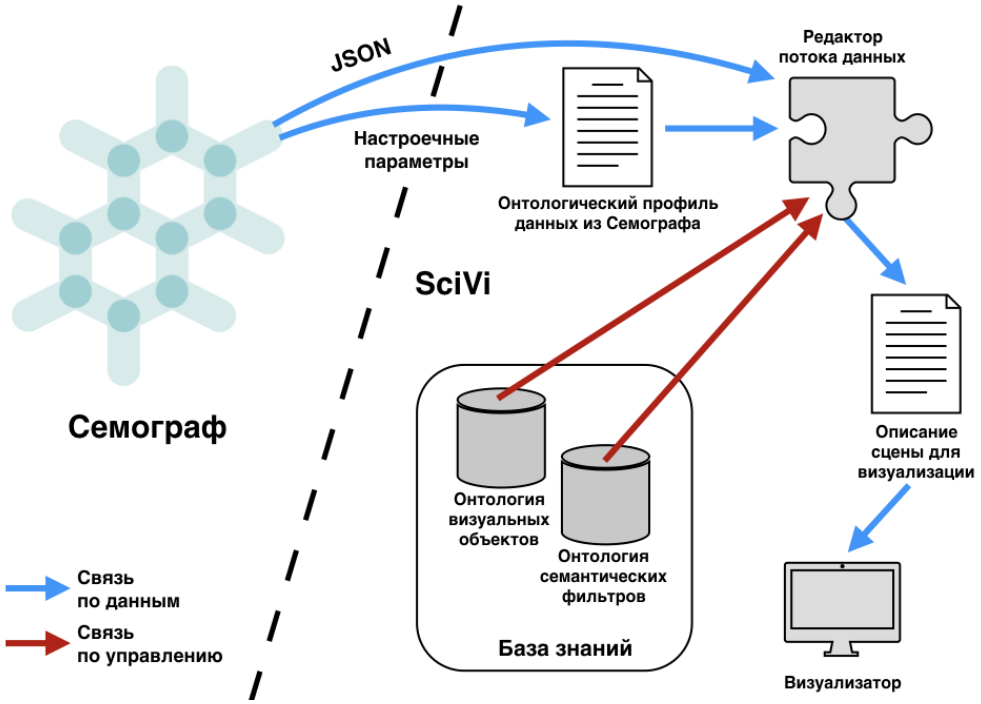

Рис. 2. Принципиальная схема взаимодействия систем SciVi и Семограф.

Ключевым связующим звеном в интеграции этих систем является т. н. *онтологический профиль* данных, получаемых от системы Семограф. Под онтологическим профилем здесь понимается онтологическое описание данных при помощи тех же концептов, которые используются для описания семантических фильтров и визуальных объектов (выходные данные с указанием их типов, настроечные параметры и т. п.). Цель создания онтологического профиля – управление процессом автоматической интерпретации передаваемых в SciVi данных, предварительно сериализованных в JSON-представление.

Онтологический профиль для каждого конкретного набора данных создаётся в среде Семограф на базе единого шаблона, который затем конкретизируется путём записи в него необходимых настроечных параметров. Обобщённый вид такого шаблона представлен на Рис. 3. Как видно из рисунка, онтологический профиль учитывает специфику визуализации данных в виде графов различной структуры.

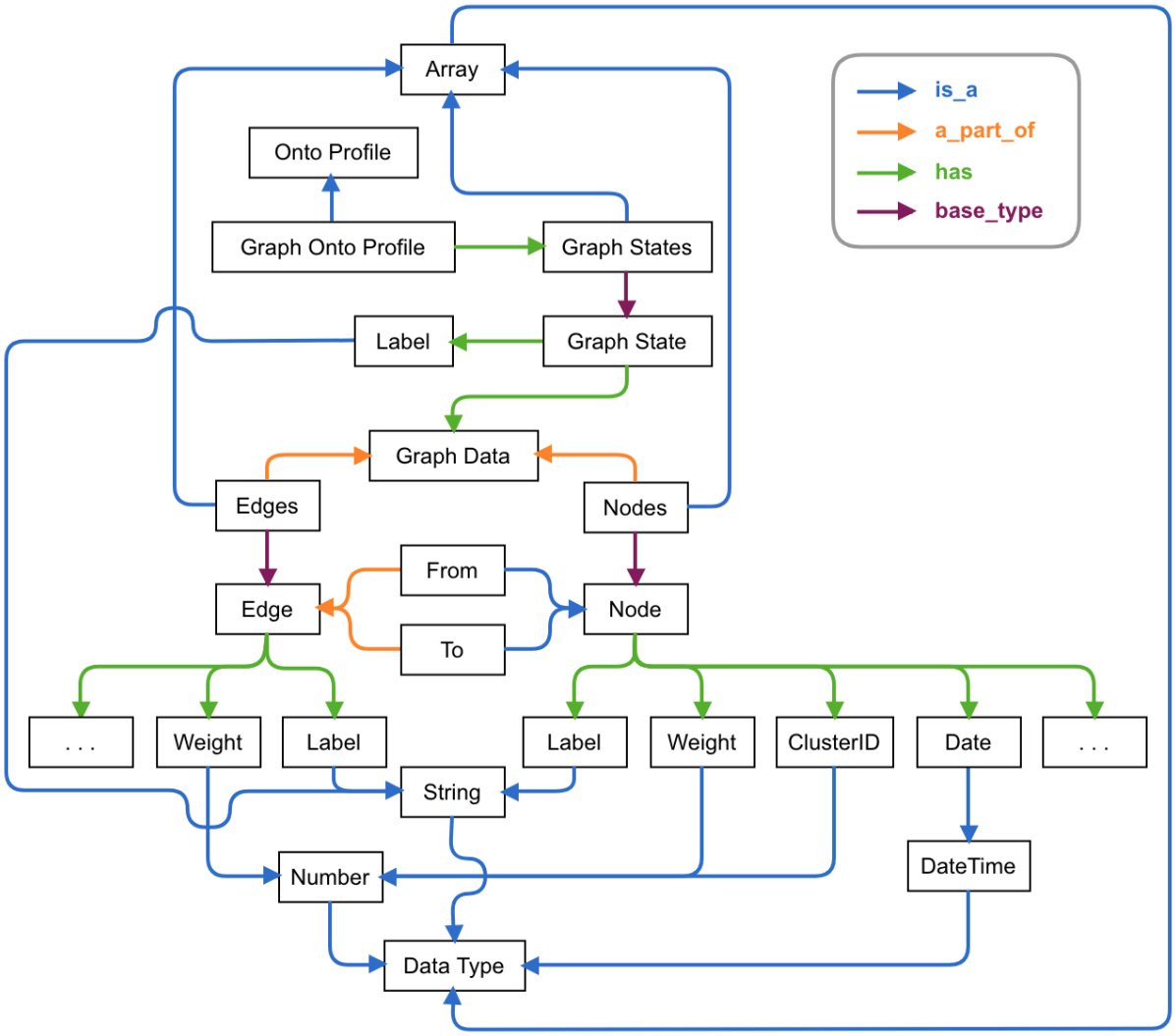

Рис. 3. Шаблон онтологического профиля данных для их представления в виде графов различной структуры.

Основными составляющими частями указанного профиля являются вершины (англ. *Nodes*) и связи (англ. *Edges*), имеющие расширяемый набор атрибутов, таких, как название (англ. *Label*), вес (англ. *Weight*) и др. В зависимости от конкретных решаемых задач, набор атрибутов может пополняться. Влияние значений тех или иных атрибутов на результат визуализации задаётся пользователем посредством диаграммы потока данных.

Набор данных для визуализации может включать несколько срезов по какому-либо показателю, например, по времени, месту или персоне. В этом случае отображаемый граф имеет несколько состояний (англ. *Graph States*), переключение между которыми осуществляется при помощи специальной шкалы, что соответствует срезам входных данных.

Для целей анализа зачастую может потребоваться кластеризация данных. Она может быть выполнена заранее (на стороне системы Семограф, различные подсистемы которой выступают здесь в качестве решателей), и в этом случае в число атрибутов вершины будет входить идентификатор кластера (англ. *ClusterID*). Кроме этого, можно задать своего рода «кластеризацию на лету», если того требуют цели визуального анализа данных. В этом случае данные разбиваются на кластеры не решателем, а самой системой визуализации, благодаря наличию в её составе соответствующих семантических фильтров. На дан-

ный момент для кластеризации используется Лёвенский алгоритм [11].

4. Круговой граф с настраиваемой иерархической кольцевой шкалой

Для структурированного отображения данных высокой степени связности используется круговой граф [12] (в системе SciVi соответствующий модуль визуализации носит название CGraph, от англ. *Circular Graph*; доступ к нему реализован посредством онтологии визуальных объектов). Его вершины расположены по окружности на равном расстоянии друг от друга. Вес вершин отображается гистограммой, столбцы которой рисуются как фон для названий вершин. Принадлежность вершины к тому или иному кластеру отображается при помощи цвета. Вершины допускают выделение по клику и перегруппировку с помощью механизма Drag-and-Drop.

Для выделенной вершины в специальной информационной панели отображается контекстная информация: её название (которое можно отредактировать), набор атрибутов, идентификатор и цвет кластера (цвет можно изменить) и список вершин, связанных с ней.

Дуги графа представляют собой квадратичные параболы, построенные по трём контрольным точкам. Первая и третья контрольные точки находятся в соединяемых вершинах, а вторая лежит в центре окружности. Толщина дуг отражает их вес. При наведении на дугу, инцидентную выделенной вершине, возникает всплывающая подсказка, содержащая контекстную информацию о дуге (чаще всего – это числовое значение веса дуги, но также поддерживается отображение произвольного текста из поля Label, см. Рис. 3).

Для представления разных срезов данных по какому-либо показателю (то есть для переключения между разными элементами массива Graph States, представленного на Рис. 3), на графе присутствует специальная шкала, отображаемая в виде слайдера с дискретным шагом. Кроме того, в процессе визуализации среза данных по одному показателю можно применять фильтрацию данных по другим показателям. Описание возможностей фильтрации приведено ниже.

В зависимости от настроек, сделанных пользователем, к вершинам может быть применена многоуровневая группировка по ряду связанных с ними показателей. Принадлежность вершин к группам, сформированным по этим показателям, отображается при помощи иерархической кольцевой шкалы. Количество и порядок следования уровней иерархии кольцевой шкалы может быть в любой момент изменён пользователем (с соответствующей перегруппировкой вершин).

Пример кругового графа с иерархической кольцевой шкалой и дискретной шкалой срезов данных представлен на Рис. 4. Этот граф построен по результатам исследования, связанного с лингвистическим анализом реплик пользователей социальной сети ВКонтакте, участвовавших в психологическом опросе BFI (см. Введение). Часть выделенных языковых параметров представлена в нижнем полукружии, психологические параметры даны в верхнем полукружии. Языковые параметры выделялись при анализе отдельных реплик, психологические параметры информанта приписывались всем его репликам. Таким образом, «психологические» вершины графа, которые представляют определённое количество информантов-носителей данных черт личности, соединялись с «языковыми» вершинами графа, отражающими языковые параметры реплик, принадлежащих информантам данного типа.

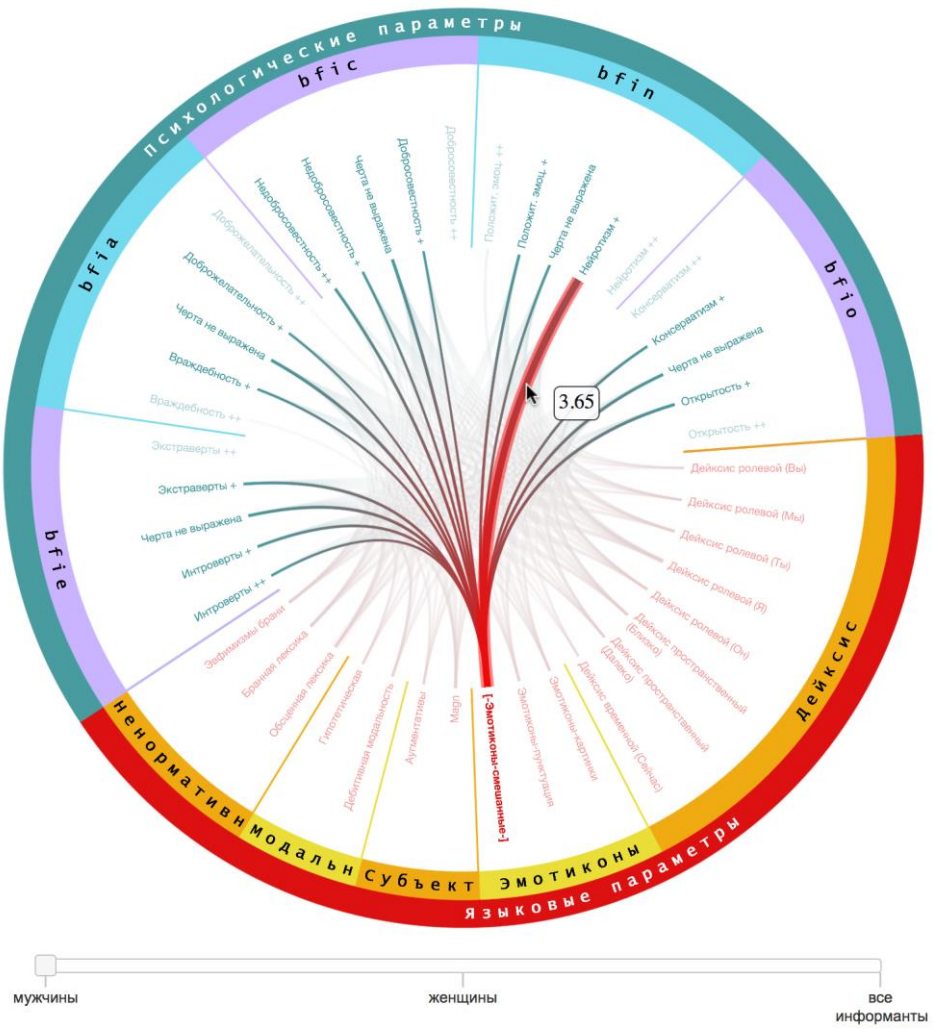

Рис. 4. Круговой граф с иерархической кольцевой шкалой и шкалой срезов данных.

Для обеспечения необходимой аналитической функциональности в круговом графе реализована поддержка различных механизмов фильтрации данных. Например, представленные в виде графа на Рис. 4 результаты, можно отфильтровать по гендерной принадлежности информантов.

Вершины и дуги могут быть отфильтрованы по весу. Однако в ряде случаев введение общего весового порога отображения вершин и/или дуг для всего графа оказывается нецелесообразным. Разные группы вершин (объединённые в сектора при помощи кольцевой иерархической шкалы) могут обладать разной средней частотностью или использовать разные шкалы (например, порядковые и интервальные), в связи с чем применение общего фильтра для всех групп сразу приводит к неадекватным с точки зрения анализа данных результатам. С целью решения этой проблемы реализован механизм иерархических фильтров, добавляемых (и удаляемых) пользователем для выделенных групп, как показано на Рис. 5.

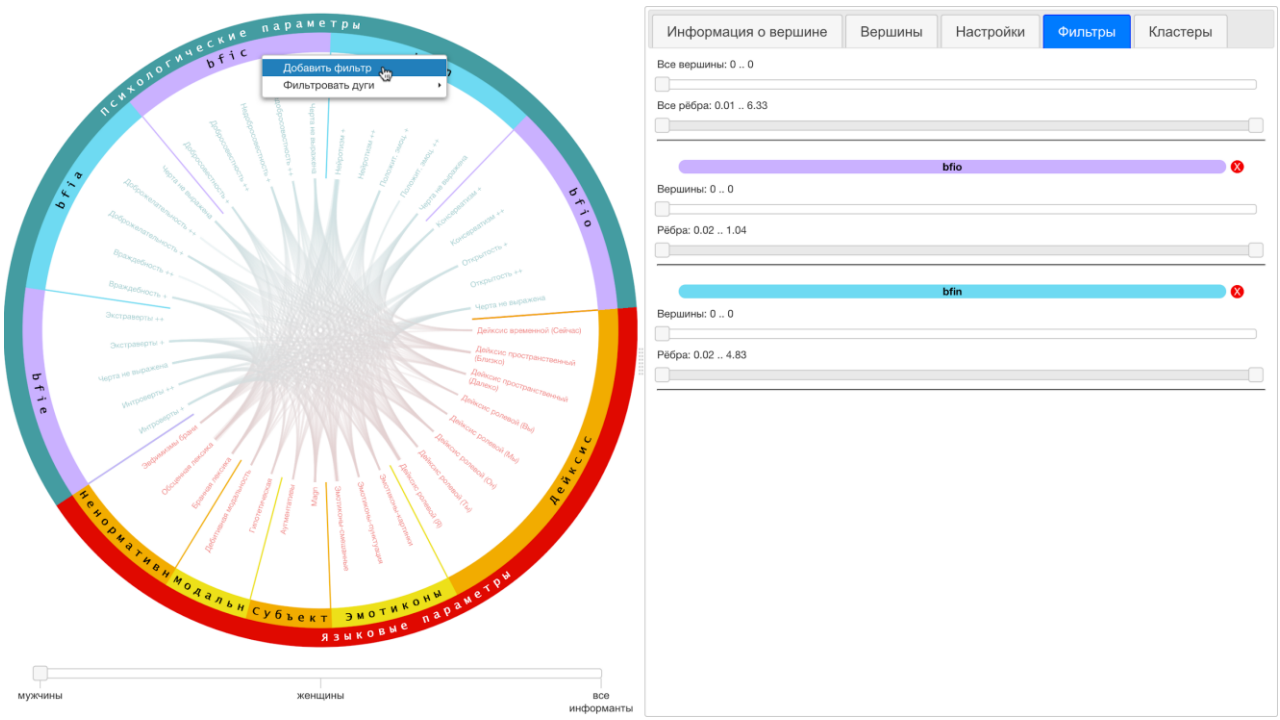

Рис. 5. Процесс формирования иерархического фильтра для групп (в правой панели добавлены фильтры для групп «bfio» и «bfin», в левой панели открыт диалог добавления фильтра для группы «bfic»).

Фильтры, добавленные для вложенных групп, а также общий фильтр для всех вершин и дуг графа работают по принципу «максимальной строгости»: для определения итогового отображения конкретной вершины и конкретной дуги из всех воздействующих на неё фильтров выбирается минимальный по длине диапазон (ищется наибольший нижний порог и наименьший верхний) и проверяется попадание веса вершины/дуги в этот диапазон.

Кроме этого, для дуг поддерживается фильтрация по принадлежности к группе. Она реализована в двух вариантах. Первый вариант предполагает отображение дуг, соединяющих между собой только те вершины, которые принадлежат выбранной группе. Второй вариант предполагает отображение дуг, соединяющих вершины из выбранной группы с вершинами из других групп. Результат такой фильтрации показан на Рис. 6.

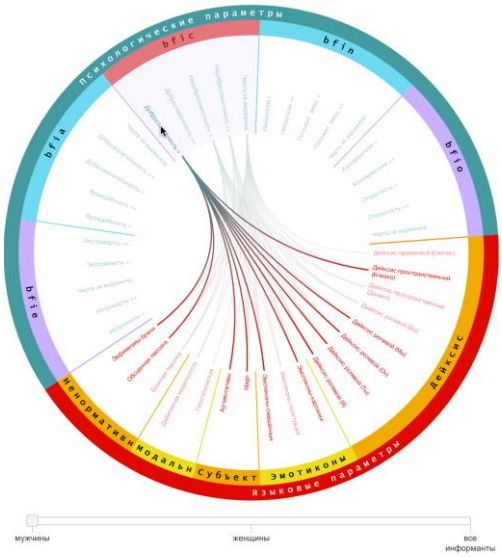

Рис. 6. Фильтрация дуг, одна из вершин, соединяемых которыми, принадлежит к группе «bfic».

Для организации удобной навигации по данным реализован поиск вершин по подстроке, по строке целиком и по регулярному выражению, а также возможность масштабирования и перехода к отображению одиночных кластеров вершин (т. н. квазизум [13]). На Рис. 7 представлены результаты применения регулярного выражения с целью фильтрации вершин, содержащих в названии знак «+» (выраженность психологической черты) и знак «(» (уточнение параметра).

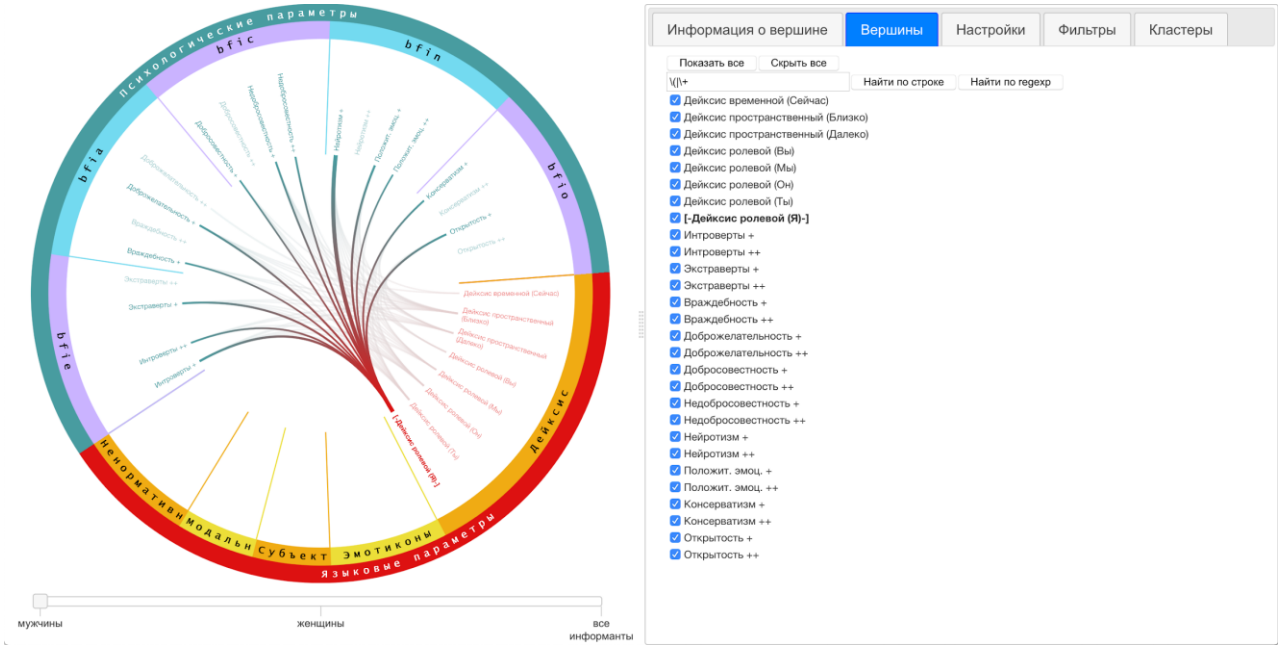

Рис. 7. Применение регулярных выражений с целью фильтрации визуализируемых данных.

Модуль визуализации кругового графа функционирует на основе библиотеки графического расширения PixiJS [14] и оптимизирован для работы в WebGL-совместимом браузере. Онтологическое описание этого модуля, являющееся частью онтологии визуальных объектов в составе базы знаний системы SciVi, приведено на Рис. 8.

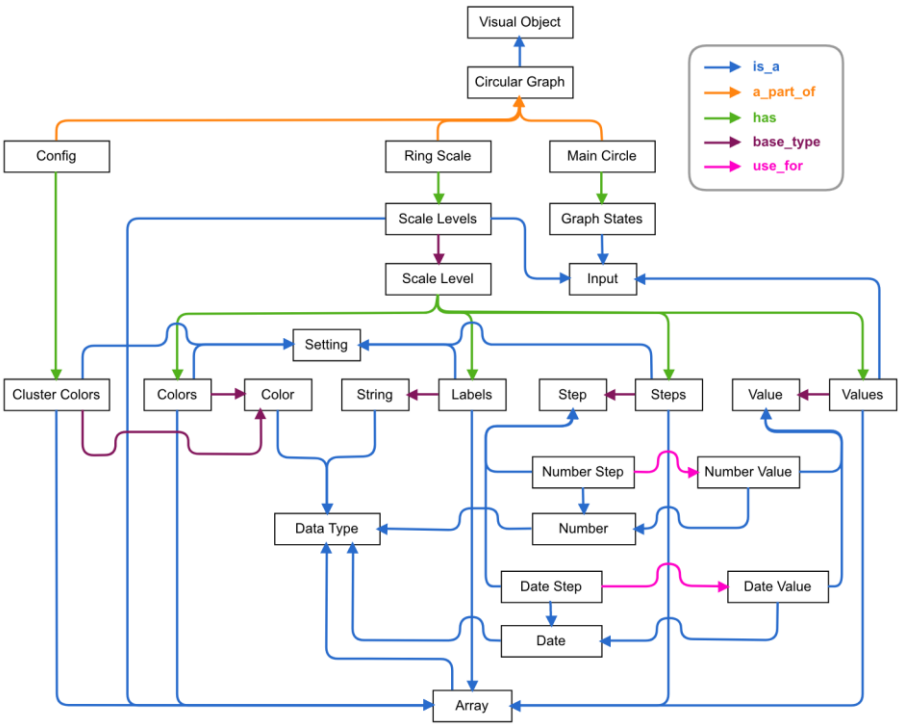

Рис. 8. Онтологическое описание модуля построения круговых графов.

В описании представлены лишь те составляющие модуля, которые соответствуют его публичному API и необходимы для его взаимодействия с ядром системы SciVi (в частности, – с редактором диаграмм потока данных). Так, например, в описании присутствует кольцевая шкала (англ. *Ring Scale*), массив цветов для кластеров вершин (англ. *Cluster Colors*) и соответствующие онтологическому профилю срезы входных данных (англ. *Graph States*, см. Рис. 3). При этом отсутствует описание внутренних фильтров, доступных для пользователя на уровне графического интерфейса, но не доступных посредством API. Такой подход был использован с целью сокращения объёма онтологического описания. В том случае, если какие-то внутренние функции модуля необходимо будет вынести в его публичный API, достаточно будет пополнить его онтологическое описание, но такое расширение не потребует внесения изменений в исходный программный код.

На основе приведённого онтологического описания автоматически создаётся соответствующий элемент палитры редактора диаграмм потока данных в системе SciVi, и пользователь может использовать его при декларировании алгоритмов визуализации. Пример диаграммы потока данных, описывающей граф, показанный на Рис. 4–7, приведён на Рис. 9.

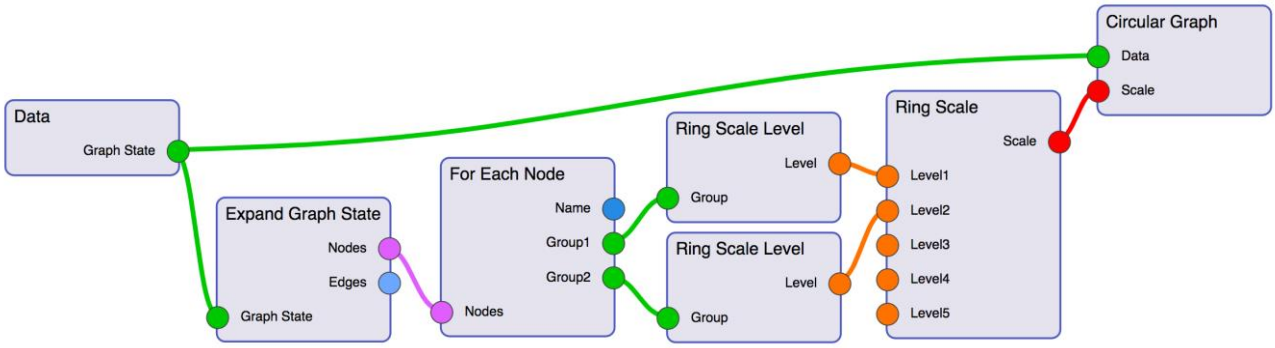

Рис. 9. Диаграмма потока данных для кругового графа с двухуровневой иерархической шкалой.

Вершина «Data» служит для представления источника данных и построена в соответствии с онтологическим профилем, показанным на Рис. 3. Цепочка вершин «Expand Graph State» и «For Each Node» задаёт обход вершин графа с целью их последующей группировки. Вершины «Ring Scale Level» предназначены для формирования уровней иерархической шкалы визуального элемента Ring Scale. Настройка способа группировки осуществляется в специальном окне редактора потока данных, отвечающем за редактирование внутренних параметров семантических фильтров. Вершина «Ring Scale» позволяет объединить уровни иерархической шкалы в единый элемент графа. Вершина «Circular Graph» построена на основе онтологического описания, приведённого на Рис. 8, и определяет вид отображения – круговой граф.

5. Граф свободной структуры

Для отображения произвольных связных данных используется граф свободной структуры (в системе SciVi соответствующий модуль визуализации носит название FSGraph, от англ. *Free Structured Graph*). Его вершины размещаются согласно заданному алгоритму укладки в режиме реального времени. Вес вершин отображается с помощью их размера, принадлежность к тому или иному кластеру обозначается цветом и геометрической формой их пиктограмм. Поддерживается интерактивное взаимодействие, такое как вращение и масштабирование визуального образа графа, а также выделение и перемещение его вершин мышью. При выделении одной вершины автоматически подсвечи-

ваются связанные с ней соседние вершины, и инцидентные ей дуги.

Информационная панель графа свободной структуры аналогична информационной панели кругового графа и также предоставляет контекстную информацию о выделенной вершине, такую как её название (допускается редактирование), набор атрибутов, список связанных вершин. Кроме того, унифицировано представление шкалы срезов данных, описанной в Разделе 4.

Поддерживается несколько видов укладки графа, в т. ч. с использованием алгоритмов на основе квазифизических аналогий [15] в режиме реального времени. Большинство алгоритмов укладки графов являются итеративными. Поэтому, чтобы при запуске приложения сразу получить приемлемый вариант укладки, поддерживается опережающее исполнение заданного числа итераций (настраиваемый параметр). Предусмотрена отдельная панель для задания настроек, связанных с текущими параметрами выбранного алгоритма укладки графа на плоскости. Имеется возможность извлечения из входных данных сведений об исходных позициях вершин графа, что может быть полезным в случае использования сторонних алгоритмов укладки. Для этих целей предусмотрен механизм статической укладки, который допускает интерактивное взаимодействие с вершинами. Имеется возможность фиксации позиции вершины (англ. *Pinning*).

Для обеспечения аналитической функциональности поддерживается механизм фильтрация отображаемых вершин по их весу, при этом дуги, ведущие к скрываемым вершинам, также скрываются автоматически. Скрытые вершины при укладке графа не учитываются. По умолчанию фильтрация вершин осуществляется раздельно для каждого кластера и для каждого среза данных. Последнее может быть изменено пользователем. Кроме того, пользователь имеет возможность в интерактивном режиме показывать/скрывать отдельные вершины.

Поддерживается ранжирование вершин, при этом вершинам разного ранга автоматически назначается различный внешний вид отображения (пиктограмма и цвет). Для каждого ранга пользователь имеет возможность самостоятельно установить нужную ему пиктограмму (круг, квадрат, ромб, треугольник и т. п.) и соответствующий цвет. Возможные виды пиктограмм описаны в онтологии визуальных объектов и доступны для выбора в виде выпадающего списка. Установка цвета для вершин каждого ранга выполняется с помощью стандартного компонента выбора цвета. Поддерживается настройка степени прозрачности цвета вершин. Ранжирование осуществляется на основе предварительной кластеризации данных, которая может быть выполнена как при помощи средств фильтрации в системе SciVi, так и сторонними средствами, например, в системе Семограф.

Для улучшения восприятия визуального образа по умолчанию скрываются названия всех вершин, кроме вершин наивысшего ранга. На Рис. 10 приведён пример графа свободной структуры, построенного по тем же данным, что и круговой граф на Рис. 4.

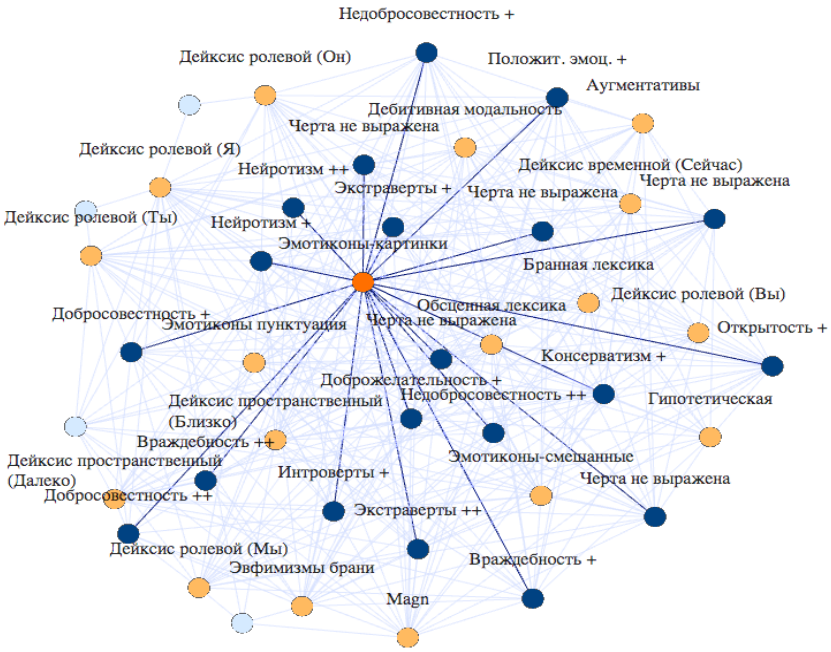

Рис. 10. Двудольный граф свободной структуры.

Модуль визуализации графа свободной структуры разработан на основе библиотеки VivaGraphJS [16] и оптимизирован для работы в WebGL-совместимом браузере. Онтологическое описание этого модуля, являющееся частью онтологии визуальных объектов в составе базы знаний системы SciVi, приведено на Рис. 11.

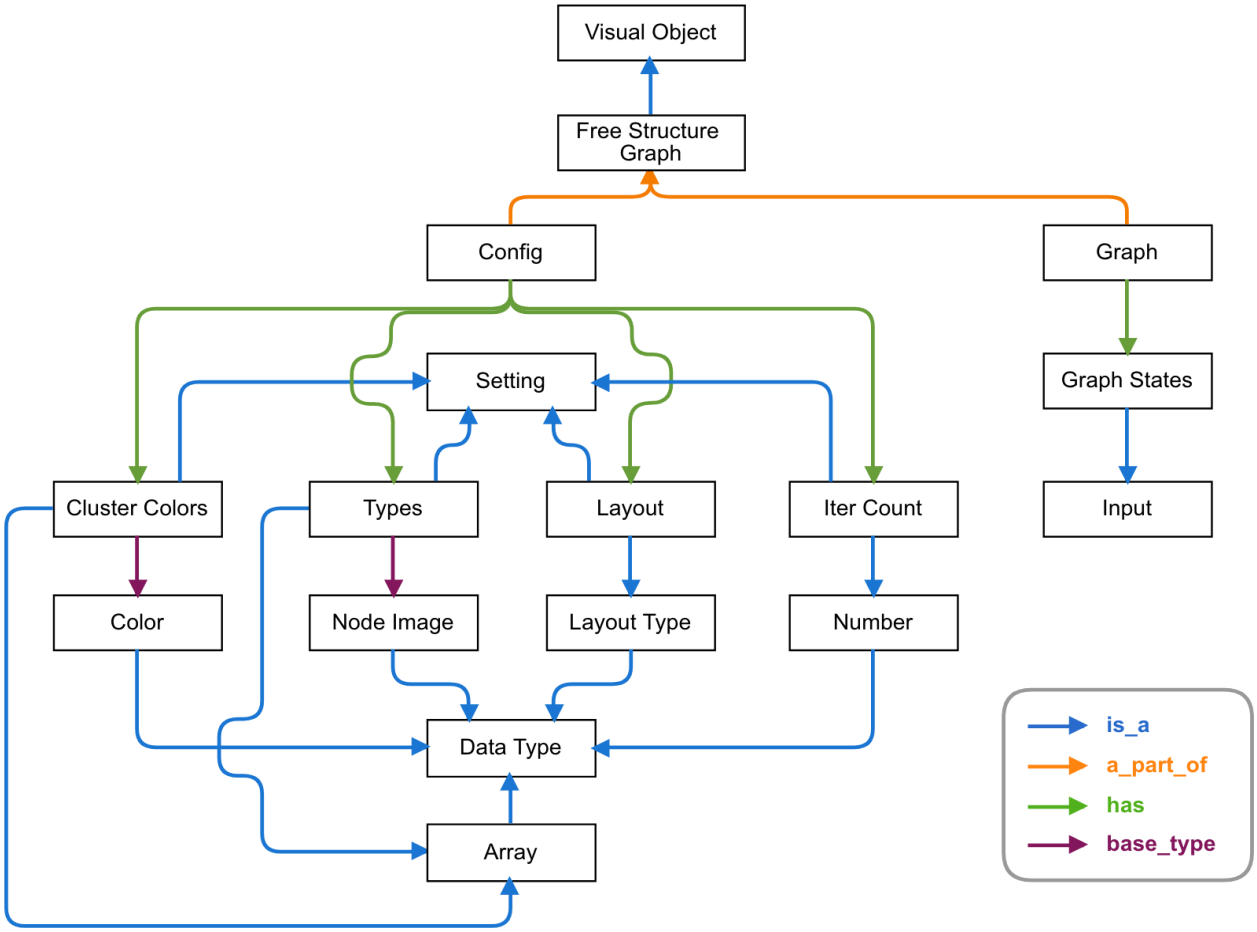

Рис. 11. Онтологическое описание модуля построения графов свободной структуры.

Представленный на Рис. 11 онтологический профиль служит основной для соответствующего элемента палитры инструментов в редакторе диаграмм потока данных SciVi. Как и в случае кругового графа, для использования графа свободной структуры при отображении конкретных данных при построении диаграммы потока данных достаточно задать вершину «Free Structure Graph» в качестве концевой для соответствующей цепочки обработки данных. На Рис. 12 демонстрируется пример диаграммы потока данных для графа, представленного на Рис. 10. Поскольку большинство настроек указанного графа не зависит напрямую от входных данных, эти настройки задаются в виде редактируемых свойств вершины «Free Structure Graph», а не через связи по данным. Построение диаграммы в этом случае намного проще, чем в случае кругового графа.

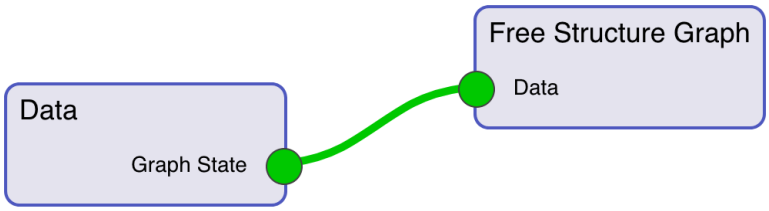

Рис. 12. Диаграмма потока данных, задающая граф свободной структуры.

Как видно из вышеизложенного, предлагаемые средства визуальной аналитики можно легко настроить на решение различных задач многопараметрического анализа данных о пользователях социальных сетей.

6. Заключение

Предложена концепция проведения комплексных гуманитарных исследований для многопараметрического описания и типологизации пользователей социальных интернет-сервисов с использованием средств визуальной аналитики. Сбор и первичную экспертную обработку данных, автоматически собираемых из социальных сетей, предлагается осуществлять в информационной системе для лингвистических исследований Семограф. Для визуализации результатов обработки данных в ИС Семограф интегрированы средства системы научной визуализации SciVi.

Цель визуального анализа состоит в обнаружении скрытых закономерностей данных за счёт их наглядного графического представления и возможностей интерактивного взаимодействия с визуальным представлением графа. Для достижения этой цели система SciVi была пополнена новыми средствами отображения и фильтрации данных высокой степени связности в виде кругового графа с настраиваемой кольцевой иерархической шкалой и шкалой срезов данных, а также графа свободной структуры, поддерживающего различные способы ранжирования вершин и различные варианты укладки графа на плоскости.

Благодаря управляемым онтологиями адаптационным возможностям системы SciVi, разработанные средства могут быть использованы для визуального анализа любых многомерных данных, генерируемых различными решателями. Это позволяет унифицированным образом адаптировать описанное в данной работе решение к широкому кругу задач комплексного многопараметрического анализа данных из различных предметных областей.

Инструменты визуальной аналитики, используемые в исследовании, позволили упорядочить связи между переменными, относящимися к неязыковым и языковым аспектам черт личности и её поведения. Многоаспектная система фильтрации данных позволила выявить психологические и языковые доминанты для каждого сектора многопараметрического классификатора и установить структуры доминантных связей между переменными для разных состояний графа, передающих социальные параметры информантов, например, пол.

Это, в частности, дало представление о влиянии гендерных различий на поведение пользователей, обладающих одними и теми же психологическими чертами.

Непосредственной перспективой разработки средств визуальной аналитики является расширение графа до охвата им всех выделенных в исследовании языковых и неязыковых параметров с передачей контента, относящегося к данным параметрам. Для исследователя в области психологии и лингвистики такой визуальный инструментарий создаст недостижимый ранее эпистемологический контекст для интерпретации изучаемого фрагмента социогуманитарной действительности.

7. Благодарности

Работа выполнена в рамках государственного задания Минобрнауки России (проект 34.1505.2017/4.6).

Список литературы

1. John, O.P., Donahue, E.M., Kentle, R.L. The Big-Five Inventory-Version 4a and 54. // Berkeley, CA: Berkeley Institute of Personality and Social Research; University of California. – 1991.
2. John, O.P., Naumann, L.P., Soto, C.J. Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues // O.P. John, R.W. Robins, L.A. Pervin (Eds.). Handbook of personality: Theory and research. – New York, NY: Guilford Press, 2008. – PP. 114–158.
3. Shchebetenko, S. Reflexive Characteristic Adaptations Explain Sex Differences in the Big Five: but not in Neuroticism // Personality and Individual Differences. – 2017. – Vol. 111. – PP. 153–156. DOI: 10.1016/j.paid.2017.02.013.
4. Belousov, K., Erofeeva, E., Leshchenko, Y., Baranov, D. “Semograph” Information System as a Framework for Network-Based Science and Education // Smart Innovation, Systems and Technologies. Smart Education and e-Learning. – 2017. – PP. 263–272. DOI: 10.1007/978-3-319-59451-4_26.
5. Ryabinin, K., Chuprina, S. Development of Ontology-Based Multiplatform Adaptive Scientific Visualization System // Journal of Computational Science. – Elsevier, 2015. – Vol. 10. – P. 370–381. DOI: 10.1016/j.jocs.2015.03.003.
6. Ryabinin, K., Chuprina, S. High-Level Toolset For Comprehensive Visual Data Analysis and Model Validation // Procedia Computer Science. – Elsevier, 2017. – Vol. 108. – P. 2090–2099. DOI: 10.1016/j.procs.2017.05.050.
7. Noy, N.F., McGuinness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology [Электронный ресурс]. – Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001. – 25 p. – URL: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf (дата обращения 27.10.2018).
8. Rose, K., Eldridge, S., Chapin, L. The Internet of Things: an Overview [Электронный ресурс] // The Internet Society (ISOC). – 2015. – URL: https://www.internetsociety.org/resources/doc/2015/iot-overview (дата обращения 27.10.2018).
9. Ryabinin, K., Chuprina, S., Kolesnik, M. Calibration and Monitoring of IoT Devices by Means of Embedded Scientific Visualization Tools // Lecture Notes in Computer Science. – Springer, 2018. – Vol. 10861. – P. 655–668. DOI: 10.1007/978-3-319-93701-4_52.
10. Lee, B., Hurson, A. Issues in dataflow computing // Adv. Comput. – 1993. – Vol. 37. – PP. 285–333 (1993). DOI: 10.1016/S0065-2458(08)60407-6.
11. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – No. 10. – 12 P. DOI: 10.1088/1742-5468/2008/10/P10008.
12. Ageev, A. A Triangle-free Circle Graph with Chromatic Number 5 // Discrete Mathematics. – 1996. – Vol. 152. – PP. 295–298. DOI: 10.1016/0012-365X(95)00349-2.

13. Бондарев А.Е., Галактионов В.А., Шапиро Л.З. Обработка и визуальный анализ многомерных данных // Научная визуализация. – М.: Национальный исследовательский ядерный университет МИФИ, 2017. – К. 4, Т. 9, №5. – С. 86–104. DOI: 10.26583/sv.9.5.08.
14. Графический движок PixiJS [Электронный ресурс]. URL: http://www.pixijs.com/ (дата обращения 27.10.2018).
15. Jacomy, M., Venturini, T., Heymann, S., Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software // PLoS ONE. – 2014. – No. 9, I. 6. DOI: 10.1371/journal.pone.0098679.
16. Движок построения графов VivaGraphJS [Электронный ресурс]. URL: https://github.com/anvaka/VivaGraphJS (дата обращения 27.10.2018).