Методы визуальной аналитики вариативности речевого поведения пользователей социальных сетей в зависимости от психологических черт личности

К.В. Рябинин[1], С.И. Чуприна[1], К.И. Белоусов[1], С.С. Пермяков[1]
kostya.ryabinin@gmail.com|chuprinas@inbox.ru|belousovki@gmail.com|rewmad@gmail.com
[1]Пермский государственный национальный исследовательский университет, Пермь, Россия

Работа посвящена вопросам создания методов и средств визуального анализа, направленного на выявление зависимостей между параметрами речевого поведения пользователей социальных сетей и психологическими характеристиками личности. Данные, подлежащие анализу, обладают высокой связностью, поэтому в качестве основного визуального средства отображения предлагается использовать различные типы графов. В работе представлены новые способы визуализации кругового графа с кольцевой иерархической шкалой и шкалой срезов данных, а также различные варианты укладки на плоскости графа свободной структуры. Оба средства визуализации имеют унифицированный интерфейс доступа к данным, что позволяет подбирать наиболее подходящий для анализа способ отображения данных и тем самым легко адаптировать аналитическую систему к специфике решаемых задач. Для увеличения когнитивной мощности разработанных средств предусмотрены возможности интерактивного взаимодействия с пользователем и механизмы настраиваемой семантической фильтрации данных. Разработанные средства включены в состав управляемой онтологиями адаптивной мультиплатформенной системы научной визуализации SciVi, которая интегрирована в систему лингвистического анализа Семограф в качестве основного инструмента визуальной аналитики.

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

Visual analytics methods of the verbal behavior variability of social networks users depending on their individual psychological features

K.V. Ryabinin[1], S.I. Chuprina[1], K.I. Belousov[1], S.S. Permyakov[1]
kostya.ryabinin@gmail.com|chuprinas@inbox.ru|belousovki@gmail.com|rewmad@gmail.com
[1]Perm State University, Perm, Russia

The paper is devoted to the creation of visual analytics methods and means for identifying the dependencies between the parameters of verbal behavior of social networks users and their psychological characteristics. The data, which are to be analyzed, have a high connectivity, therefore we suggest to display them using different types of graphs. The paper presents new ways to visualize a circular graph with a hierarchical ring scale and a data states scale, as well as various options for laying out a free-structure graph on the plane. Both visualization tools have a unified data access interface, which allows to choose the most suitable depicting way for particular data analysis and thereby to adapt the analytical system to the specifics of the tasks being solved. To increase the cognitive power of the developed tools advanced user interactions and custom semantic data filtering mechanisms are supported. The developed tools are included in the SciVi ontology driven adaptive multiplatform scientific visualization system. This system is integrated into the Semograph linguistic analysis system as the main visual analytics engine.

Keywords: *visual analytics, ontology engineering, graphs, language parameters, psychological characteristics, BFI, social networks users.*

1. Введение

В современной науке существует запрос на создание фундаментальной концепции личности, которая бы позволила описывать, объяснять и прогнозировать речевое и неречевое поведение человека и социальных групп, включая группы пользователей социальных интернет-сервисов (англ. Social Network Services, SNS). Несмотря на широкий спектр задач, решаемых в области исследования SNS, в открытых источниках пока не встречаются концепции комплексного анализа типов их пользователей, взаимосвязей между ними и моделей их поведения.

Комплексное описание пользователей SNS должно основываться на моделях интеграции социального, поведенческого, психологического и языкового (и, более широко, мультимодального) профилей цифровых проекций личностей. В качестве социальных параметров рассматривается информация 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 психологических опросов, а языковые – на основе анализа комментариев пользователей и тегов к размещаемым материалам.

Для автоматизации анализа и интерпретации данных из социальных сетей предлагается использовать аппарат визуальной аналитики, основанный на когни-

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

тивной графике, а также на методах и средствах научной визуализации. Данные из социальных сетей зачастую обладают высокой связностью, и именно наличие и характер связей обычно выступают предметом анализа. Соответственно, в качестве базового способа визуализации предлагается использовать интерактивные графы различной структуры, расширенные настраиваемыми средствами предобработки (в частности, семантической фильтрации) отображаемых данных.

В качестве аналитической платформы предлагается использовать систему лингвистического анализа Семограф [2]. Целью данной работы является внедрение в Семограф новых средств визуального анализа вариативности речевого поведения пользователей социальных сетей в зависимости от психологических черт их личности. За визуализацию данных, обрабатываемых системой Семограф, отвечает интегрированная с ней система научной визуализации SciVi [10]. Для достижения поставленной цели разработаны и включены в состав SciVi два модуля визуализации графов, обладающие унифицированным интерфейсом доступа к данным: модуль визуализации круговых графов с настраиваемой кольцевой иерархической шкалой и шкалой срезов данных, а также модуль визуализации графов со свободной структурой.

Для тестирования и отладки созданных средств визуализации в работе использовались данные 821 пользователя социальной сети ВКонтакте, участвовавших в психологическом опросе (Вопросник Большой Пятерки – BFI [8]). Языковые параметры выделялись на материале многоуровневого лингвистического анализа 18000 автоматизировано собранных реплик информантов. Анализ выполнялся тремя экспертами в системе Семограф. В данной работе рассмотрены некоторые из анализируемых речевых параметров, относящиеся к семантике (дейксис, модальность), стилистические параметры (бранная лексика, аугментативы и др.) и параметры, касающиеся использования в диалогах пользователей графических средств (эмотикон) в качестве полноценных реплик.

2. Онтологический инжиниринг в визуальной аналитике

В ходе предыдущих исследований были проанализированы наиболее популярные системы научной визуализации (например, 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 либо в особенностях источника данных (что требует использования специальных алгоритмов доступа или представления данных в некотором нестандартном формате), либо 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 быстро перенастроена для наилучшего соответствия требованиям решаемых задач. В роли такой модели предлагается использовать онтологии, так как они обеспечивают человекочитаемость и самодокументируемость. В состав базы знаний системы визуализации предлагается включить онтологию визуальных объектов и графических сцен, описывающую поддерживаемые системой средства визуализации, а также онтологию семантических фильтров, описывающую допустимые способы трансформации входных данных.

Для решения специализированных задач предлагается использовать дополнительные онтологии. Например, если источником подлежащих визуализации данных выступает некоторый решатель (англ. Solver – расчётная программа), в состав системы визуализации предлагается включить онтологию 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, предлагается использовать онтологию электронных компонентов [7]. Она служит для автоматической генерации прошивки электронного устройства и интеграции в эту прошивку кода визуализации.

Предложенные принципы построения систем научной визуализации 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 [7]. 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.

В тех случаях, когда помимо наглядного представления научных данных требуется также их глубинный анализ, зачастую необходимо выполнить над ними некоторые преобразования. Например, фильтрацию данных в соответствии с заданными критериями (поро-

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

говыми функциями и т. п.), математические преобразования (масштабирование, нормализация и т. п.), классификацию и кластеризацию, статистический анализ и т. д. Для поддержки и унификации механизма такого рода трансформаций предлагается использовать т. н. семантические фильтры – операторы преобразования подлежащих визуализации данных. Описание алгоритма работы этих операторов, их входов, выходов и настроечных параметров хранится в онтологии семантических фильтров. Это позволяет пополнять их набор путём внесения изменений только в базу знаний системы научной визуализации с помощью средств высокоуровневого пользовательского интерфейса.

Семантические фильтры допускают суперпозицию, задаваемую диаграммой потока данных [5]. В системе SciVi присутствует специальный высокоуровневый графический редактор для составления таких диаграмм. Вершинами в них выступают источники данных, семантические фильтры, визуальные объекты и графические сцены, а связи выражают пути передачи данных. Палитра допустимых вершин формируется автоматически на основе соответствующих онтологий.

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

Для того чтобы представить алгоритм обработки и визуализации данных в виде диаграммы потока данных, необходимо в качестве начальной вершины задать их источник. Для этого источник данных должен быть описан онтологически 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.). Такое описание составляет т. н. онтологический профиль.

На рис. 1 в обобщенном виде приведен онтологический профиль данных, извлекаемых из системы Семограф, которые предназначены для визуализации в виде графов различной структуры.

Основными составляющими частями такого 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.

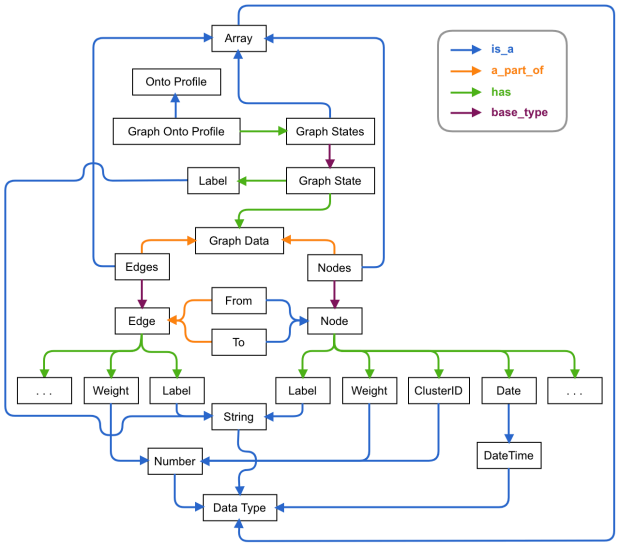

Рис. 1. Обобщённый вид онтологического профиля данных, предназначенных для представления в виде графа.

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 данный момент для кластеризации используется Лёвенский алгоритм [3].

4. Круговой граф с настраиваемой иерархической кольцевой шкалой

Для структурированного отображения данных высокой связности используется круговой граф [1]. 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.

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.

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.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

Для обеспечения необходимой аналитической функциональности в круговом графе реализована поддержка интерактивности, включающая возможности масштабирования, фильтрации вершин и дуг по весу, выделения, переноса и переименования отдельных вершин, изменения цвета, выделения дуг, перехода к отображению одиночных кластеров вершин (т. н. квази-зум [9]), а также изменения порядка следования уровней иерархии кольцевой шкалы (с соответствующей перегруппировкой вершин). Модуль визуализации кругового графа функционирует на основе библиотеки графического расширения PixiJS[1] и оптимизирован для работы в WebGL-совместимом браузере.

На рис. 2 приведён круговой граф, построенный по результатам лингвистического анализа 18000 реплик информантов – пользователей социальной сети ВКонтакте, участвовавших в психологическом опросе BFI [8]. Часть из 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 определенное количество информантов-носителей данных черт личности, соединялись с «языковыми» вершинами графа, отражающими языковые параметры реплик, принадлежащих информантам данного типа. Кроме того, результаты, представленные на графе, можно дополнительно отфильтровать по гендерной принадлежности информантов.

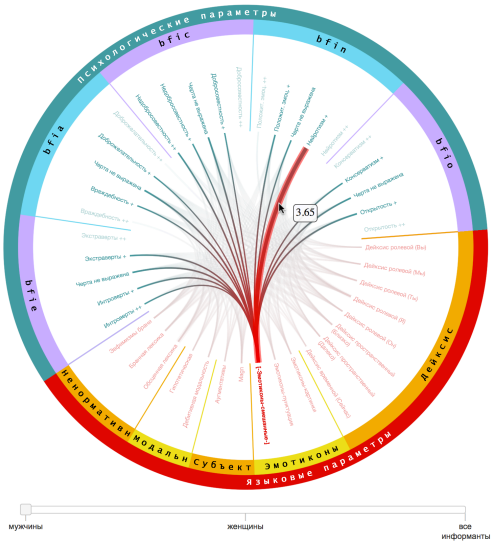

Рис. 2. Круговой граф с иерархической кольцевой шкалой и шкалой срезов данных.

5. Граф свободной структуры

Для отображения произвольных связных данных предлагается использовать граф свободной структуры. Вершины в нём располагаются согласно алгоритму укладки на основе квази-физических аналогий [4] в режиме реального времени.

Вершины представляют собой различные геометрические примитивы, такие как квадрат или окружность. Поддерживается ранжирование вершин, при этом вершинам разного ранга назначается различный внешний вид. Ранжирование осуществляется на основе кластеризации данных по различным критериям. Размеры вершин отражают их веса. Дуги графа отображаются при помощи отрезков прямых, их веса отражаются толщиной отрезков. Модуль визуализации графа свободной структуры разработан на основе библиотеки VivaGraphJS[2] и оптимизирован для работы в WebGL-совместимом браузере.

На рис. 3 приведён пример графа свободной структуры, построенного по тем же данным, что и граф на рис. 2.

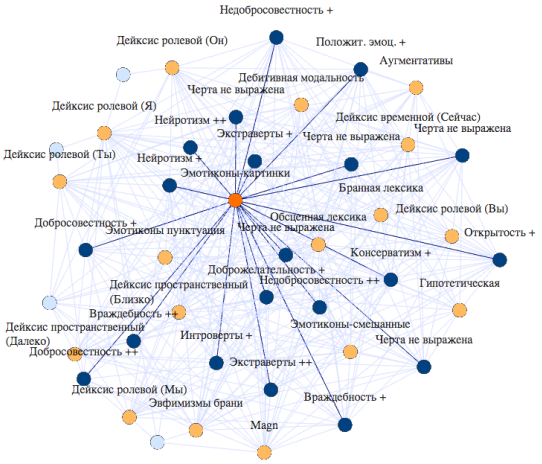

Рис. 3. Двудольный граф свободной структуры.

Поддерживаются различные средства интерактивности: перемещение, масштабирование и поворот всего изображения; перемещение отдельных вершин; задание/изменение цвета, приписываемого вершин одного ранга; возможность показывать/скрывать отдельные вершины. Реализован механизм подсветки выбранной вершины в зависимости от её ранга: при выборе вершины автоматически подсвечиваются все связанные с ней вершины более низкого ранга.

Для уменьшения загруженности визуального образа по умолчанию скрываются названия всех вершин, кроме вершин наивысшего ранга. Пользователь может изменять настройки, связанные с укладкой графа на плоскости. Поддерживается фильтрация отображаемых вершин и дуг по весу, при этом в случае работы с двудольным графом фильтрацию вершин можно осуществлять отдельно для каждой доли.

6. Заключение

В работе представлены новые возможности системы научной визуализации SciVi: круговой граф с кольцевой иерархической шкалой и шкалой срезов данных,

[1] http://www.pixijs.com/
[2] https://github.com/anvaka/VivaGraphJS

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

а также граф свободной структуры, поддерживающий ранжирование вершин и различные варианты укладки на плоскости. Эти средства были использованы для наглядного представления и визуальной аналитики извлекаемых из социальных сетей данных о речевом поведении пользователей. Первичная обработка данных производилась средствами системы лингвистического анализа Семограф.

Интеграция SciVi в систему Семограф позволила выявить соответствия между установленными в результате психологического опроса, в котором принимали участие более 800 пользователей социальной сети ВКонтакте, чертами личности пользователей и особенностями их речевого поведения в социальной сети, выявленными при помощи средств лингвистического анализа.

Разработанные средства визуализации позволили на основе анализа имеющихся и отсутствующих связей между отдельными языковыми и психологическими параметрами получить значимую для предметной области информацию. Так, в частности, были выявлены различия в использовании средств ролевого (в том числе социально маркированного) и пространственного дейксиса у пользователей-экстравертов и интровертов. Интересны различия в использовании бранной и обсценной лексики в письменной речи пользователей, имеющих ярко выраженные черты консерватизма и открытости и мн. др. Кроме того, речевая вариативность может объясняться не только психологическими различиями, но и гендерными (в частности, публичное использование обсценной лексики имеет связи с разными психологическими характеристиками в мужской и женской группах пользователей). Визуальная модель языковых и психологических соответствий, построенная с учетом гендерных характеристик, позволяет извлекать релевантную информацию об организации данной предметной области; строить интерпретации с опорой на визуальную основу (это особенно удобно, так как модель дает возможность перебора самых разных вариантов); детализировать направления научного поиска, которые дают нетривиальную информацию о связях психологической, языковой и социальной природы человека в процессе осуществления своей жизнедеятельности.

Благодаря управляемым онтологиями адаптационным возможностям системы SciVi разработанные средства могут быть использованы для визуального анализа любых многомерных данных, генерируемых различными решателями, включая устройства в составе экосистемы Интернета вещей.

В ближайших планах предусмотрено использование SciVi для комплексного анализа речевого и неречевого поведения пользователей социальных сетей. В дальнейшем в составе SciVi планируется реализовать модули визуализации данных с картографической привязкой, а также модули построения настраиваемых многоуровневых трёхмерных иерархических графов.

7. Благодарности

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

8. Литература

[1] 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.

[2] 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. – Springer, 2017. – PP. 263–272. DOI: 10.1007/978-3-319-59451-4_26.

[3] 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.

[4] 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.

[5] Lee B., Hurson A. Issues in Dataflow Computing // Advances in Computers. – Elsevier, 1993. – Vol. 37.– PP.285–333.DOI: 10.1016/S0065-2458(08)60407-6.

[6] Ryabinin K., Chuprina S. High-Level Toolset For Comprehensive Visual Data Analysis and Model Validation // Procedia Computer Science. – Elsevier, 2017. – Vol. 108. – PP. 2090–2099. DOI: 10.1016/j.procs.2017.05.050.

[7] 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. – PP. 655–668. DOI: 10.1007/978-3-319-93701-4_52.

[8] Shchebetenko S. Reflexive characteristic adaptations explain sex differences in the Big Five: But not in neuroticism // Personality and Individual Differences. – Elsevier, 2017. – Vol. 111. – PP. 153–156. DOI: 10.1016/j.paid.2017.02.013.

[9] 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.

[10] Рябинин К.В., Баранов Д.А., Белоусов К.И. Интеграция информационной системы Семограф и визуализатора SciVi для решения задач экспертного анализа языкового контента // Научная визуализация. – М.: Национальный исследовательский ядерный университет МИФИ, 2017. – К. 4, Т. 9, №4. – С. 67–77. DOI: 10.26583/sv.9.4.07.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18