Применение метода главных компонент для выявления семантических различий и анализа изменения положения в пространстве при анализе информационного контента сетевых сообществ

И.А. Рыцарев[1,2], Р.А. Парингер[1,2], А.В. Куприянов[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. В работе мы предлагаем подход к анализу социальных групп и их положения относительно друг друга на основе выявления семантических различий в текстах, представленных в виде частотных словарей. Исходные текстовые данные мы получили путём сбора записей тематических интернет-сообществ. Для сбора записей мы реализовали специализированный программный модуль, позволяющий анализировать и загружать как посты, так и комментарии из интересующих открытых сообществ социальной сети ВКонтакте. Для составления частотного словаря, мы разработали, алгоритм, который учитывает особенности данных, собираемых из социальных сетей. В статье мы предлагаем подход, основанный на использовании методов снижения размерности пространств признаков, для выявления ключевых слов на основе анализа частоты их употребления. Алгоритм, который мы представили, использует метод главных компонент. В результате работы мы показали, что, используя коэффициенты полученного линейного преобразования можно оценить значимость слов. С использованием полученных оценок, мы не только смогли выявить не только ключевые слова, но и составить семантические различия в сообществах социальных сетей, а так же построить графики изменения положения этих групп в пространстве относительно друг друга.

1. Введение

Разработка новейших информационных и производственных технологий, скорость и динамичность их широкомасштабного внедрения оказывает влияние на все сферы личной и общественной жизни людей. Можно видеть, как быстро, на всех уровнях, меняются формы активного взаимодействия между субъектами социальных отношений.

Социальные сети представляют собой модель реальности, в которой строится взаимодействие людей. Они, с одной стороны, повторяют все существующие закономерности коммуникации людей, а с другой, являются иной, не похожей на реальность.

При этом двойственность отношений, которая характерна современным людям, актуализирует необходимость переключаться из одного контекста в другой либо же находиться в них одновременно. Например, обсуждая в обыденном (оффлайн) общении какую-либо новость из

сети, люди переносят отношения из одного контекста в другой, а затем продолжают транслировать эти отношения из оффлайн в онлайн контекст. Сложность взаимодействия создаёт ряд противоречивых ситуаций, где перед субъектами встают задачи, связанные с установлением (или даже обновлением) доверительного общения, проверкой информации на точность либо безоговорочным принятием контента.

Сетевые сообщества, с одной стороны, являются социальными группами, которые можно описать с помощью существующего понятийного аппарата социальной психологии (в этих группах есть общие интересы и деятельность, возможность контакта каждого с каждым и эффект «мы» как результат членства этой в этой группе), а с другой стороны, являются самобытными единицами в силу специфики условий существования.

Само по себе явление интернет-коммуникации столь масштабное, что изучать его традиционными методами социологической науки уже не представляется возможным. Интернет-сообщества и способы их взаимодействия лишь отчасти могут быть описаны социально-психологическими характеристиками, принятыми в науках, изучающих явления общественной жизни.

Количество виртуальных сообществ постоянно возрастает. Вместе с тем возрастает и их тематическое разнообразие. Виртуальная среда всё больше начинает выполнять функции «третьей сигнальной системы». Участие в большинстве сообществ не ограниченно ни возрастными, ни социальными, ни территориальными рамками. Свобода участия (отсутствие членских билетов, взносов, обязательств), вступление в различные социальные группы и выход из них в любой момент, по собственному желанию, позволяет людям выбирать контент в соответствии со своими интересами и влечениями. К тому же, для того, чтобы видеть контент, совсем не обязательно быть членом группы.

Благоприятным фактором, способствующим решению исследовательских задач в сфере интернет, оказалась сама цифровая поисковая «природа» интернет функционирования. Возможность исчисления и классификации «цифровых следов» неограниченного объёма, высокая скорость (оперативность) и проверяемость получаемых данных, прозрачность и относительно невысокая финансовая стоимость проведения исследований, определили её несомненные преимущества [1].

Первые исследования поведения людей в социальных сетях принадлежат психологу Михаилу Косинскому и его коллегам из Кембриджского университета [2]. С помощью специального приложения они собирали данные пользователей Facebook, обобщали и строили психологические портреты на основе теста Р. МакКрае и П. Коста «Большая пятёрка». Результаты использовались для генерации индивидуально ориентированных рекламных предложений. Эти исследования стали предпосылкой появления работ, связанных с изучением поведения пользователей в социальных сетях, с предъявлением им целенаправленно выбранного содержания (рекламы, рекомендаций друзей, социально/политически значимых проектов и т.д.).

Авторами программного комплекса «Социальный эхолот», разработанного в Самарском университете в 2017 году (А.А. Гришин, К.С. Лисецкий, Н.Ю. Самыкина, П.В. Шиверов), были описаны социально-психологические эффекты, работающие в социальных сетях (эффект социальной фасилитации, конформизма, фаворитизма, маятника, Рингельмана и т.д.). На основе этого ими были предложены варианты управления групповой динамикой в сетевых сообществах. Под динамикой в данном случае понимается совокупность внутригрупповых социально-психологических процессов и явлений, характеризующих весь цикл жизнедеятельности группы, и его этапы: образование, функционирование, развитие, стагнацию, регресс, распад.

2. Сбор и первичная обработка данных

Для анализа сообществ была выбрана социальная сеть ВКонтакте. Данная сеть является одно из самых популярных в русскоязычном сегменте интернета. Особенностью социальной сети является то, что она доступна для всех и не имеет строго определённой тематики. Кроме того, данная социальная сеть свободно предоставляет программный интерфейс приложения для написания внешних приложений.

Получение данных из социальной сети ВКонтакте возможно только при использовании стандартных инструментов, предоставленных разработчиками сети, поэтому в рамках первого этапа для начала работы с социальной сетью необходимо получить доступ к серверам социальной сети ВКонтакте. Для этого нужно создать приложение на сервере ВКонтакте и получить ключи доступа, выполнив следующие шаги:

1. В разделе для разработчиков нужно создать приложение следуя инструкциям мастера создания приложений.
2. Получить ID приложения и секретный ключ для быстрого подключения приложения к серверам.
3. В настройках приложения дать приложению разрешения для работы с сообществами и записями.

Следующим этапом работы являлась разработка программного модуля сбора данных. Реализация велась на языке программирования Python с использованием библиотеки для написания скриптов социальной сети ВКонтакте. Данная библиотека разработана сторонними разработчиками и обладает множеством методов, построенных на базе официального программного интерфейса, написанного на языке Python от разработчиков ВКонтакте. После установки модуля в систему он импортируется и создаётся объект авторизации который будет хранить в себе логин, пароль, ID приложения и ключ. Именно через него происходит все взаимодействие с социальной сетью.

Для дальнейшего описания алгоритма действий нужно ввести несколько терминов. Запись – это любое текстовое сообщение в сообществе. Пост – это запись, несущая в себе информацию о некотором или некоторых событиях, чаще всего является побуждающей для начала обсуждения какой-либо темы. Комментарий – запись под постом, отражающая реакцию конкретного пользователя на данный пост или комментарий другого участника.

Этап сбора записей со стены состоит из двух процедур: процедуры сбора постов и процедуры сбора комментариев к постам.

Процедура сбора постов заключается в инициализация специального метода программного интерфейса официального приложения для сбора данных, возвращающего список постов пользователя или сообщества и обработке полученного ответа с постами. Для инициализации метода необходимо указать уникальный идентификатор пользователь или уникальный адрес сообщества, информацию из которой нужно загрузить.

Процедура сбора комментариев к посту осуществляется аналогичным образом: через инициализацию метода программного интерфейса и обработке полученного ответа. Отличие заключается в используемом методе и необходимом наборе параметров: для получения комментариев необходимо указать уникальный идентификатор не только сообщества, но и поста.

Применение двух процедур, описанных выше позволяет производить сбор информации из любых открытых сообществ социальной сети ВКонтакте, но существует программное ограничение на количество запросов в сутки, что существенным образом затрудняет сбор информации. Так сторонние приложения имеют ежедневное ограничение в 2500 запросов и получать не более 100 записей за один запрос.

Также в свойствах записей содержится дата публикации, поэтому для снижения числа запросов, существует возможность ограничения количества получаемых записей в соответствии с интересующим временным интервалом. Сравнивая время в записи с интересующим интервалом, можно получить только записи, соответствующие интересующему временному интервалу.

Описанные процедуры составляют основу разработанного алгоритма сбора записей из социальной сети ВКонтакте.

3. Алгоритм составления частотного словаря

Записи представляют собой особую структуру хранения информации, содержащую как текстовые поля (непосредственное содержание записи), так и метаданные, содержащие в себе

информацию об уникальных идентификаторах однозначно определяющие данную записи и её автора, данные о времени публикации записи и т.д.

Поскольку основой исследования, проводимого в данной работе является составление частотного словаря, то разрабатываемые алгоритмы относятся к прежде всего к полю, содержащему текст рассматриваемых записей, но остальная информация записей также потенциально может быть учтена и использована для обработки.

После обращения к соответствующему полю записи и получению данных, полученный фрагмент текста разделяется на слова. Словом, текста в данном случае называется последовательность буквенных символов, разделённая пробелами, цифрами или знаками препинания. Поскольку записи в социальных сетях изобилуют опечатками и ошибками, то для корректного учёта количества слов при составлении частотного словаря необходимым этапом алгоритма являлась орфографическая проверка полученных данных. Для выполнения процедуры проверки орфографии в рамках данной работы использовалась сторонняя библиотека. Параллельно производится учёт количества постов и комментариев.

После выполнения описанных этапов, подсчёт количества вхождений каждого уникального слова w во всём множестве S текстовых данных осуществлялся по формуле:

$$count(w) = \sum_i (w_i \in S)$$

Полученное количество слов делится на число записей и таким образом оценивается частота употребления слов в записях – формируется частотный словарь. Схематически разработанный алгоритм представлен на рисунке 1.

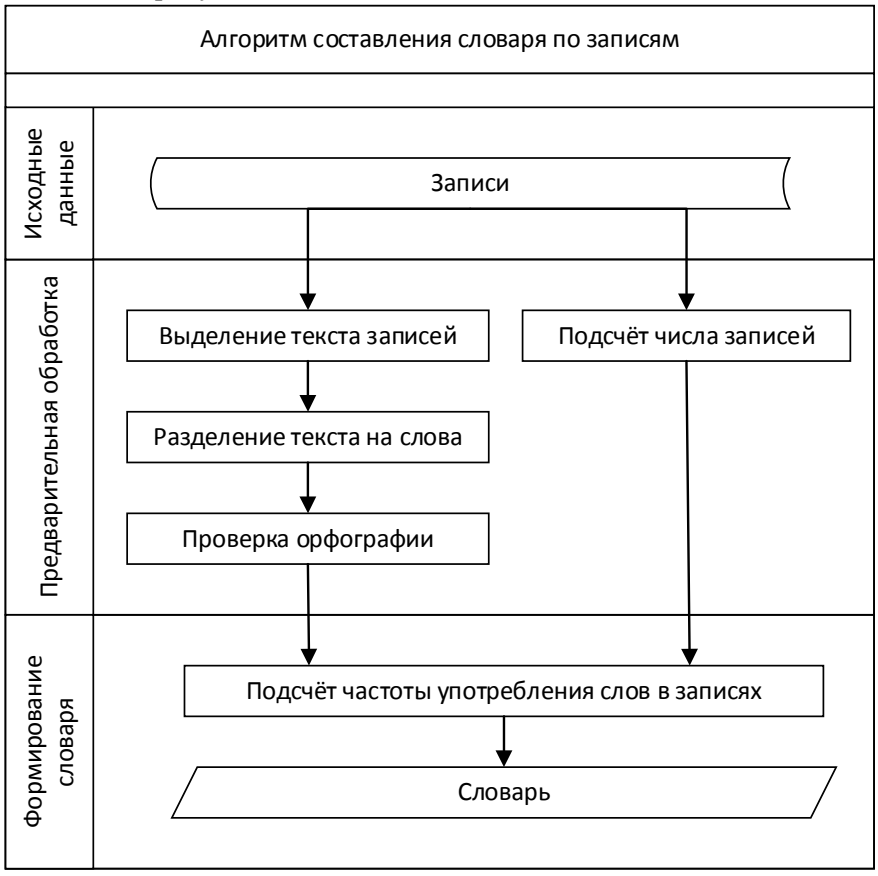

Рисунок 1. Схема алгоритма составления частотного словаря.

4. Алгоритм выявления семантических различий

Для выявление семантических различий на основе анализа частотных словарей был разработан соответствующий алгоритм. Основной идей алгоритма является применение существующих техник снижения размерности пространства признаков для решения задачи ранжирования слов. В рамках данной статьи использовался метод главных компонент [4].

Исходными данными для работы алгоритма являются частотные словари. Каждый словарь может быть представлен вектором в пространстве составляющих его слов. При наличии двух словарей, можно сформировать один новый признак таким образом, что он будет разделять эти словари наилучшим образом. Новый признак при использовании метода главных компонент формируется путём умножения значений исходного пространства на соответствующий вектор коэффициентов. Значения коэффициентов вектора, используемых для формирования нового признака можно использоваться как оценку вклада того или иного слова в формируемый признак. Таким образом можно для сформированного признака составить список слов, внёсших наибольший вклад в его формирования. Составленный подобным образом список слов описывает признак, обеспечивающий разделимость рассматриваемых словарей и таким образом описывает их семантическое различие.

Первым шагом алгоритма является нормировка данных. Значения частоты употребления слов нормировались для получения в частотах словарях приводились к значениям в промежутке [0;1] по следующей формуле:

$$y(x) = \frac{x - x_{min}}{x_{max} - x_{min}},$$

где x_{min} – минимальное значение среди элементов в векторе, x_{max} - максимальное значение среди элементов в векторе.

Следующий шаг алгоритма – применение метода главных компонент.

Метод главных компонент - один из основных способов уменьшить размерность данных, с наименьшей потерей информации. Вычисление главных компонент обычно сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. По определению ковариации двух признаков X_i и X_j вычисляется следующим образом:

$$cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j$$

где μ_i — математическое ожидание i-го признака.

Таким образом матрица ковариации представляет собой симметричную матрицу, где на диагонали лежат дисперсии соответствующих признаков, а вне диагонали - ковариации соответствующих пар признаков. Из отношения Рэлея [5] вытекает, что максимальная вариация набора данных будет достигаться вдоль собственного вектора этой матрицы, соответствующего максимальному собственному значению. Поэтому главные компоненты, на которые нужно спроецировать данные, являются просто собственными векторами соответствующих собственных значений этой матрицы. Таким образом значения элементов собственных векторов являются искомыми оценочными коэффициентами для формирования слов, описывающих семантические различия. Чтобы получить визуальную интерпретацию взаиморасположения рассматриваемых словарей в полученном семантическом пространстве, нужно умножить вектор значений частот употребления слов на соответствующий собственный вектор.

5. Результаты

Для исследования работоспособности разработанных алгоритмов были выбраны сообщества, имеющие схожие тематики. Все выбранные сообщества являются сообществами жителей Самары и Самарской области и три из пяти сообществ являются открытыми площадками для общения студентов и преподавателей двух самых крупных ВУЗов Самары. В связи с существующим ограничением на сбор данных в рамках данного исследования, был введён дополнительный критерий отбора записей на основании времени их публикации (с 1 июня по 30 сентября 2018 года с шагом в 1 месяц). Для сбора и обработки данных были выбраны следующие сообщества и соответствующие им условные обозначения:

I. «Подслушано в Самарском Университете».
II. «Подслушано Самарский Университет» (переименованное сообщество «Подслушано в СамГУ».
III. «Подслушано в СамГТУ».
IV. «Услышано Самара».

V. «Подслушано Самара».

После сбора записей выбранных сообществ за интересующий промежуток времени, в соответствии с разработанным алгоритмом составления частотного словаря: были выделены тексты записей, проведена процедура разделения полученных текстов на слова, выполнена проверка орфографии и подсчитано количество постов и комментариев для каждого из исследуемых сообществ (таблица 1). Полученные в результате работы алгоритма, частотные словари частично представлены в таблице 2

Таблица 1. Количество постов и комментариев.

Месяц	Запись	Номер сообщества				
		I	II	III	IV	V
Июнь 2018	**Пост**	29	141	304	361	1296
	Комментарий	147	333	662	3948	54364
Июль 2018	**Пост**	23	150	292	359	1356
	Комментарий	182	430	978	3860	66429
Август 2018	**Пост**	39	268	313	454	1375
	Комментарий	238	564	991	3526	89785
Сентябрь 2018	**Пост**	99	236	481	382	1228
	Комментарий	415	375	1043	3070	49253
Общая сумма	**Пост**	190	795	1390	1556	5255
	Комментарий	982	1702	3674	14404	259831

Таблица 2. Фрагмент частотных словарей.

Слово	Номер сообщества				
	I	II	III	IV	V
мне	0.032	0.025	0.022	0.024	0.032
меня	0.030	0.023	0.016	0.019	0.026
только	0.024	0.031	0.036	0.033	0.026
просто	0.016	0.020	0.023	0.022	0.019
будет	0.015	0.008	0.006	0.009	0.012
может	0.014	0.009	0.010	0.007	0.013
люди	0.011	0.007	0.006	0.007	0.009
автор	0.011	0.008	0.012	0.011	0.010
такие	0.010	0.005	0.004	0.005	0.008
время	0.010	0.007	0.007	0.007	0.008
детей	0.008	0.007	0.007	0.008	0.007
человек	0.007	0.012	0.006	0.010	0.007
тогда	0.007	0.005	0.003	0.003	0.006
конечно	0.006	0.012	0.010	0.011	0.007
много	0.006	0.006	0.005	0.008	0.006
всем	0.032	0.025	0.022	0.024	0.032

Графическое представление результатов применения метода главных компонент к ежемесячным частотным словарям, представлены на рисунках 2 (а-г) и 3 (а-г).

Для того чтобы все точки были в одной системе координат за основу был взят базисная июня 2018 года.

6. Заключение

На основании анализа полученных результатов можно сделать вывод, что сообщества в социальных сетях –динамично изменяющиеся объекты, анализировать которые необходимо

при помощи разных методов и подходов. Однако, это так же показывает динамику «перетекания» пользователей из одно группы в другу в зависимости от времени года.
Задачей дальнейших исследований является разработка технологии автоматической фильтрации частотных словарей, а так же подходов к анализу сообщений на основе «свежих» часто употребляемых слов.

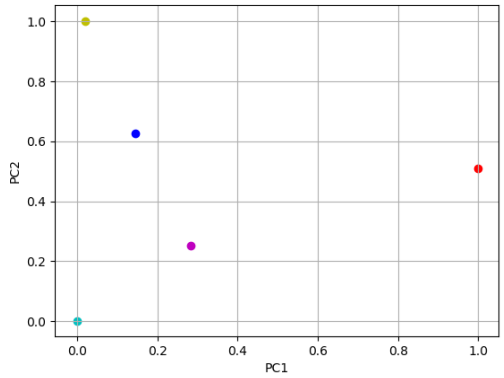

Рисунок 2. Результат обработки данных за июнь 2018.

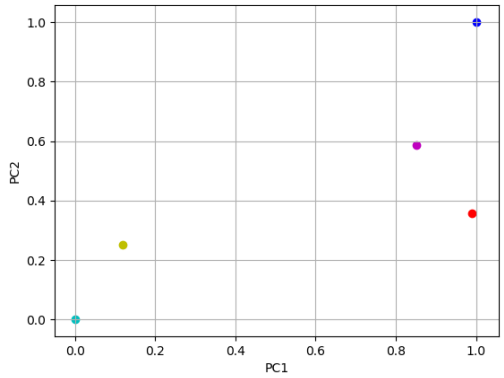

Рисунок 3. Результат обработки данных за июль 2018.

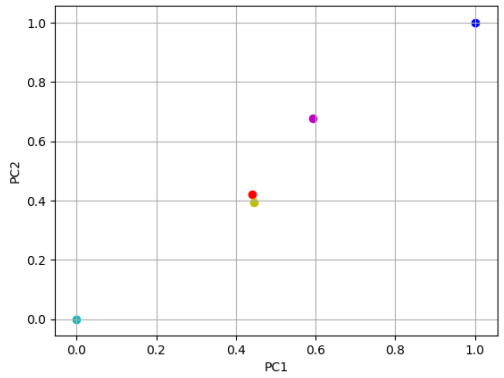

Рисунок 4. Результат обработки данных за август 2018.

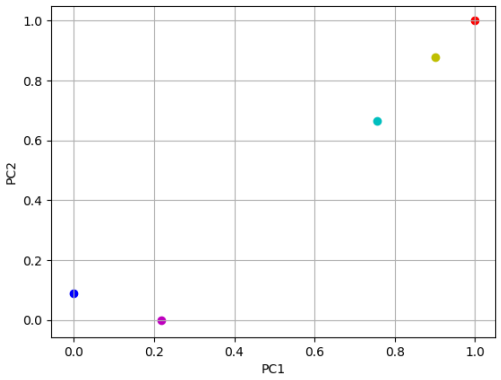

Рисунок 5. Результат обработки данных за сентябрь 2018.

7. Литература

[1] Rytsarev, I.A. Application of the principal component analysis to detect semantic differences during the content analysis of social networks / I.A. Rytsarev, D.D. Kozlov, N.S. Kravtsova, A.V. Kupriyanov, K.S. Liseckiy, A.K. Liseckiy, R.A. Paringer, N.Yu. Samykina // CEUR Workshop Proceedings. – 2018. – Vol. 2212. – P. 262-269.

[2] Kosinski, M. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines / M. Kosinski, S.C. Matz, S.D. Gosling, V. Popov, D. Stillwell // American Psychologist. – 2015. – Vol. 70(6). – P. 543-556. DOI: 10.1037/a0039210.

[3] Schwartz, H.A. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach / H.A. Schwartz, J.C. Eichstaedt, M.L. Kern // PLoS ONE. – 2013. – Vol. 8(9). – P. e73791. DOI:10.1371/journal.pone.0073791.

[4] Jolliffe, I.T. Principal Component Analysis // Principal Component Analysis, Springer, 2002.

[5] Wasserman, L. All of Statistics: A Concise Course in Statistical Inference // Springer, 2005.

[6] Horn, R.A. Matrix Analysis / R.A. Horn, C.A. Johnson // Cambridge University Press, 1985.

Благодарности

Работа выполнена при частичной поддержке Федерального агентства научных организаций (соглашение № 007-ГЗ/Ч3363/26); Министерства образования и науки РФ в рамках реализации мероприятий Программы повышения конкурентоспособности Самарского Университета среди

ведущих мировых научно-образовательных центров на 2013–2020 годы; грантов РФФИ № 16-41-630761, № 17-01-00972, № 18-37-00418; в рамках госзадания по теме № 0026-2018-0102 "Оптоинформационные технологии получения и обработки гиперспектральных данных".

Analysis of components to identify semantic proximity and analyzing changes in position in space in tasks of content analysis of social networks

I.A. Rytsarev[1,2], R.A. Paringer[1,2], A.V. Kupriyanov[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. All articles *must* contain an abstract. The abstract text should be formatted using 10 point Times or Times New Roman and indented 25 mm from the left margin. Leave 10 mm space after the abstract before you begin the main text of your article, starting on the same page as the abstract. The abstract should give readers concise information about the content of the article and indicate the main results obtained and conclusions drawn. The abstract is not part of the text and should be complete in itself; no table numbers, figure numbers, references or displayed mathematical expressions should be included. It should be suitable for direct inclusion in abstracting services and should not normally exceed 200 words in a single paragraph. Since contemporary information-retrieval systems rely heavily on the content of titles and abstracts to identify relevant articles in literature searches, great care should be taken in constructing both.