О.Ю. Кольцова

ВЫЯВЛЕНИЕ СОЦИАЛЬНЫХ ПРОБЛЕМ И ИЗМЕНЕНИЙ ЧЕРЕЗ АНАЛИЗ БОЛЬШИХ МАССИВОВ ТЕКСТОВ В БЛОГАХ И СОЦИАЛЬНЫХ СЕТЯХ

Вот уже около десятилетия у социальной реальности появился новый «дом» - Интернет, ставший индикатором, предиктором и двигателем социальных процессов и изменений. «Арабские революции» этого года показали его значение даже в тех обществах, где доля пользователей Интернета не очень велика. Казалось бы, перед социологами открылись небывалые возможности для исследования и прогнозирования социальных процессов, однако исследовательское сообщество сталкивается с проблемой отсутствия средств анализа громадных объемов текстовых данных, который невозможен без специализированного программного обеспечения, продвинутых математических алгоритмов и компьютерно-лингвистических подходов.

Этот текст посвящен предварительным результатам работы междисциплинарного исследовательского коллектива в рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач», поддержанного грантом Научного фонда НИУ ВШЭ № 11-04-0006, 2011-2012, участники: Е.Ю.Кольцова (руководитель), А.В.Кинчарова, Л.В.Пивоварова, К.А.Маслинский, Т.Г.Ефимова, Е.А.Терещенко, Ю.В.Павлова; техподдержка: С.Н.Кольцов, Р.М.Бахмудов. Общие задачи проекта – выявить на больших массивах данных русскоязычной блогосферы тематические кластеры постов (о чем говорят?) и сообщества, основанные на комментировании (кто с кем говорит?); выяснить, совпадают ли комментовые сообщества с тематическими кластерами (т.е. основана ли общность комментирования на общности темы?). Тестовой тематикой является тема Ислама. В данном тексте рассматриваются только этапы технологической цепочки текстового анализа и связанные с ними трудности; сетевой анализ, необходимый для выявления сообществ комментирования, и проблемы его сопоставления с текстовым не рассматриваются.

Что такое блог

Блог – это вид Интернет-контента, который требует сайта с определенными техническими характеристиками. Блог представляет собой дневник, в котором автор располагает записи в обратном хронологи-

ческом порядке. Записи («посты») могут сопровождаться картинками, видео- и аудифайлами, ссылками. Отличие блога от новостной ленты – жанровое: блог предполагает индивидуальное авторство и, как правило, носит непрофессиональный или неофициальный характер. Каждый пользователь может самостоятельно создать или заказать сайт для блога, но львиная доля блогов находится на специальных блог-сервисах или блог-хостингах, предоставляющих простые конструкторы для создания блогов. Так, в русскоязычной блогосфере насчитывается около 50 миллионов блогов (что говорит о распространенности и, соответственно, социальной значимости этого явления); из них автономных блогов – чуть больше миллиона. Авторы блогов под своими именами (а люди, не ведущие блогов – анонимно), могут оставлять комментарии к записям других блоггеров; на некоторых блог-сервисах комментарии имеют древовидную структуру (т.е. можно отвечать на конкретный комментарий, а не на сам пост); на других они выстраиваются в линейку. Это определяет и разную структуру дискуссий.

Для русскоязычной блогосфеоры характерно слияние блог-сервисов и социальных сетей. Ярким примером этого является Живой журнал: классический блог-хостинг предоставляет не функцию дружбы, а функцию blog-roll, т.е. ссылок на понравившиеся блоги, зачастую не зависимо от сервиса, на котором они расположены. Поэтому, например, в США связность блогов зависит не от принадлежности блогов к блог-хостингам, а в большей степени от социальных факторов (например, общности тематики или политической позиции). В России функции френдования в гораздо большей степени замыкают коммуникацию между блоггерами внутри одной блог-платформы; чрезвычайно редки комментарии с других платформ (Этлинг, Алексанян и соавт. 2010).

Получение исходных данных

Поскольку наша задача «считывания» тематической и сетевой структуры всей блогосферы или большой ее части, нам необходима не просто разработка методологии, а построение полной технолого-методологической цепочки, от получения сырых данных до их анализа, что и стало основной деятельностью нашего исследовательского коллектива. Проблема первого этапа – получения данных из блогов – на первый взгляд, решается просто: в поисковую систему вводится запрос и закачиваются результаты. Однако исследователь сразу же сталкивается с рядом сложностей:

- недоступность части Интернет-контента поисковым системам;

- невозможность протестировать правильность критериев поиска в виду недоступности генеральной совокупности текстов;
- невозможность получить все результаты поиска (большинство поисковиков выдает только первую тысячу);
- недоступность критериев рейтингования страниц поисковой системой и, соответственно, неясность принципов попадания страниц в выдаваемую тысячу, т.е. несоответствие выдаваемых данных критериям формирования социологических выборок;
- невозможность автоматически выгрузить все полученные результаты, что критично при больших объемах.

Вот почему большинство социологических исследований Интернета на данный момент представляют собой локальную работу с выборками, поддающимися ручной закачке и обработке (Papacharissi 2007), или выгрузке с помощью очень простых, рассчитанных на небольшие объемы и часто алгоритмически непрозрачных программ (Bruns 2009; Rogers 2010). Перечисленные проблемы, в виду своей непривычности, кажутся техническими, а не социологическими. Однако, если вдуматься, они соответствуют проблемам доступа в поле в качественных исследованиях или проблемам организации доступа к респондентам в массовых опросах. Разница в том, что решение этих проблем невозможно без IT-специалистов, в связи с чем в нашей группе и было разработано собственное ПО для закачки, хранения и подготовки данных к анализу. Учитывая замкнутость дискуссий внутри блог-хостингов, было принято решение ограничиться одним из них (LiveJournal), известным повышенной концентрацией общественно-политических дискуссий по предыдущим исследованиям (Alexanyan & Koltsova 2009; Gorny 2004).

Подготовка данных для текстового анализа

Все подходы к анализу больших массивов текстов основаны на автоматизированном частотном анализе слов, словосочетаний, их совместной встречаемости, а отнесение текста к той или иной группе / категории происходит на основании сравнения его лексического состава с лексическим составом других текстов или образцов. Для того, чтобы частотный анализ проводился правильно, во флективных языках, к которым относится и русский, алгоритм должен «понимать», что разные словоформы одного и того же слова должны быть посчитаны вместе. Для этого специальные программные продукты – лемматизаторы – на основе имеющихся словарей и алгоритмов сводят слова к их начальным формам. У процесса лемматизации есть ряд недостатков, связанных с неправильной лемматизацией в некоторых случаях, и с тем, что лемма-

тизация превращает текст в линейную последовательность лемм, отчего такой подход к анализу текстов получил название “bag of words” (мешок слов). Поэтому с помощью таких автоматизированных методов анализа текстов невозможно:

- находить и группировать вместе сходные по смыслу отдельные фразы;
- отличать противоположные по смыслу суждения («люблю ислам» - «не люблю ислам»);
- улавливать иронию, сарказм и иносказание;
- деконструировать идеологические приемы, применяемые в текстах;
- получать представление о социальном контексте суждений и текстов.

Однако с помощью них возможно:

- с определенным уровнем погрешности разделять большие массивы относительно длинных текстов на тематические группы;
- с определенным уровнем погрешности различать разные по идеологии тексты на одну тему, если идеологическая направленность проявляется в частотно-лексических характеристиках;
- определять состав атрибутов, приписываемых понятиям, если лексические маркеры этих понятий заранее выделены экспертами;
- определять общую эмоциональную окраску текста при наличии словарей эмоционально-окрашенной лексики и т.п.

Наши задачи относятся ко второй группе, поэтому мы обратились к лемматизации и, учитывая большие объемы, остановились на единственно возможном подходе «мешка слов», отказавшись от работы со словосочетаниями. Кроме лемматизации, существуют и другие этапы подготовки текстов к анализу, например, построение матрицы различий между текстами, на которых мы не будем здесь останавливаться.

Алгоритмы разделения текстов на группы

Общими проблемами всех алгоритмов анализа – будь то сетевой или текстовый – является соотношение качества и вычислительной сложности как конкурирующими параметрами. Вычислительная сложность алгоритма (O) - оценивается приблизительно, как функция от количества данных. O – не просто вопрос того, сколько времени будет работать компьютер, но также и того, сколько потребуется оперативной памяти и других ресурсов. Особенно критичным это оказалось для кластерного анализа, который, как правило, требует помещения в оперативную память данных сразу обо всех анализируемых текстах. Тестирование показало, что даже ПО, заявлявшее способность работы с большими мас-

сивами текстов, на поверку не приняло более нескольких тысяч постов (напр., R).

Вторая проблема – оценка качества анализа. Как определить, хорошие, правильные ли получились кластеры? Существуют две основные группы методов оценки качества работы различных алгоритмов: (а) внешние: определение доли «правильно» отнесенных единиц через сравнение с образцом и (б) внутренние: вычисление ряда параметров, таких как соотношение внутрикластерной и межкластерной вариации. Для методов анализа текстов ведущими методами являются внешние, основанные на сравнении с образцовым корпусом, разделенным на группы вручную с помощью кодировщиков. Проблемой этого подхода является распространенность некритичного отношения к результатам кодирования и проблематичность экстраполяции результатов, полученных на одних типах образцовых корпусов, на другие типы (напр., другой тематики). Кроме того, методы оценки качества алгоритмов анализа (кластерного, сетевого и др.) только разрабатываются в математическом сообществе; более того, сами алгоритмы также находятся в стадии разработки, и дебатируются даже сами ключевые понятия (кластеров, сообществ). Поэтому социолог сталкивается с проблемой выбора алгоритма из набора средств, надежность которых до конца не установлена, и это приходится принимать как проблему, не имеющую на данный момент решения.

Выбор единицы анализа

Блоги политематичны. Это делает невозможной процедуру кластеризации блогов целиком: слишком много шумов даже для алгоритмов нечеткой кластеризации, которые пока что работают только с малыми данными. Но даже если бы нечеткое разделение на тематические кластеры удалось, не понятно, как сравнивать четкие множества текстов с нечеткими комментовыми сообществами в целях выявления их пересечения. Поэтому в качестве единицы текстового анализа был выбран отдельный пост.

Вычисление меры сходства между текстами

Что такое более или менее похожие тексты? Здесь возможны два основных подхода. Первый подход заключается в том, что экспертами (между которыми достигнута высокая надежность интеркодирования) определяются образцы текстов – скажем, «антиисламский», «шиитский», «суннитский» и т.д. Затем алгоритм анализирует частотно-лексические характеристики этих текстов и экстраполирует получившиеся наборы признаков на новые тексты, раскладывая их по группам, к которым каж-

дый текст находится ближе всего. Эту операцию принято называть классификацией, т.к. она не предполагает поиска латентных групп, а лишь делит корпус на заранее известные. Также как и кластеризация, она может быть полной или неполной, четкой или нечеткой. В нашем исследовании, мы предполагаем, что основной ценностью разрабатываемой методологии может стать возможность находить именно латентные группы – например, группу текстов с такой интерпретацией исламской тематики, которая раньше не предполагалась и которая может иметь потенциал не ожидаемых социальных изменений. Поэтому классификация для нас является менее предпочтительной процедурой.

Второй подход – это формальное вычисление сходства. Наиболее частый его вариант основан на представлении текста в векторной форме (описание см. напр. Andrews & Fox 2007). Выше говорилось, что при обработке больших массивов тексты представляются в виде «мешка» слов, точнее, их лемм, частоты которых подсчитываются. Далее в векторном подходе каждая лемма представляется в виде измерения в N-мерном пространстве, где N – общее число лемм, встречающихся в корпусе. Каждый текст представляется в виде вектора в этом пространстве; частоты лемм в данном тексте соответствуют длине проекции вектора на ось соответствующего данной лемме измерения. Такие вектора становятся математически сравнимыми по длине и по направлению. Существует несколько способов вычисления расстояния между ними.

Одна из проблем такого подхода – т.н. «проклятие многомерности». Нетрудно заметить, что большинство векторов будет иметь нулевые длины по большинству измерений, т.к. большая часть слов корпуса встречается только в небольшой части текстов, а те, что встречаются везде, чаще всего слова с малой дискриминационной силой, типа местоимений. Кластеризация матрицы, построенной на основе таких векторов, будет смазанной из-за большого количества шума, а на больших массивах может и вовсе оказаться невыполнимой. В данный момент мы продолжаем работы по вычленению массива лексики высокой дискриминирующей силой.

Классический кластерный анализ

Кластеризация, в отличие от классификации – это процесс разбиения массива данных на группы в отсутствие образцов, который предполагает самостоятельный поиск наиболее схожих элементов алгоритмом на основе оценки измеренных между ними расстояний. Два основных классических метода кластерного анализа – плоская и иерархическая – имеют ряд проблем. Плоская кластеризация (k-means, k-center, k-median и производные) начинается со случайного определения k элементов, ко-

торые условно назначаются центроидами, т.е. центральными, или наиболее типичными элементами будущих кластеров. Зависимость алгоритма от случайно выбранных начальных элементов приводит его к неспособности давать стабильные результаты. Для нас изменчивость семантических кластеров означает невозможность сравнения совпадения их с комментовыми сообществами. Кроме этого, алгоритм дает приемлемое качество только, если кластеры массива имеют явно выраженные центры с равноудаленными от них предельными элементами (шарообразные кластеры) и не выявляет, скажем, цепеобразных кластеров, которых вполне можно ожидать в блогосфере.

Иерархическая кластеризация имеет некоторые схожие и некоторые собственные проблемы. Восходящая (агломеративная) кластеризация начинается с определения каждого из n элементов как отдельного кластера и последовательно объединяет наиболее схожие элементы до тех пор, пока не получится один кластер. Нисходящая (дивизимная) кластеризация построена ровно наоборот; оба класса алгоритмов выдают в составе результата все промежуточные разбиения, от 1 до n, образуя полную дендрограмму. Это и создает проблему выбора между получившимися разбиениями единственно верного, особенно, если исследователь имеет дело с сотнями тысяч элементов и с количеством разбиений такого же порядка. Кроме того, алгоритмы агломеративной кластеризации очень вычислительно сложны и не работают на больших данных. У нисходящей кластеризации другая проблема: на каждом шаге массив должен единовременно разбиваться на два максимально различных кластера, следовательно, такому алгоритму на каждом шаге требуется субалгоритм плоской кластеризации – например, k-means. Отсюда нисходящая кластеризация приобретает и проблемы плоской.

В современных ПО в чистом виде эти алгоритмы почти не используются и входят в состав более сложных либо усовершенствованных (см. напр. Carpineto et al 2009), но в целом все они имеют серьезные ограничения по объему данных.

Альтернативы классической кластеризации

Одним из новейших способов является кластеризация, основанная на графах (graph-based clustering). В ней набор объектов (напр., текстов) представляется в виде полного графа, где ребра репрезентируют сходство между объектами, а веса ребер обратно пропорциональны дистанции (мере несходства) между парой объектов. После преобразования матрицы различий в граф, к нему могут быть применены все алгоритмы community detection, в т.ч. Louvain (Blondel 2008) – единственный, рабо-

тающий с графами в 10^9 вершин, и другими алгоритмами оптимизации функции качества под названием «modularity», которая позволяет находить разбиения, оптимальные по количеству и составу субграфов. Однако проблема заключается в том, что эти алгоритмы пока слабо внедрены в стандартное ПО, доступное не-мататематикам.

Другая альтернатива – кластеризация, основанная на моделях или на распределениях (model-based clustering или distribution based clustering) (см. напр. Ahlquist & Breunig 2011), в основе которой лежит следующая цепочка рассуждений. В каждом массиве объектов (напр., людей или текстов) их параметры (напр., возраста или частоты слов) распределены по определенной функции. Если массив состоит из подмассивов, внутри которых объекты более схожи, чем между массивами, то распределение параметров в каждом из массивов будет несколько иным. Например, распределение возрастов в субмассивах «мужчины» и «женщины» будет несколько отличаться. Т.о. можно представить распределение параметров в общем массиве как сумму распределений этих параметров в субмассивах. Отсюда можно попытаться восстановить число и состав этих распределений из общего распределения и таким образом смоделировать статистичесую структуру массива с его подмассивами. Для этого разработано множество подходов, но все они, к сожалению, сохраняют проблему классической кластеризации: неясность критериев определения количества субмассивов; также, нам не удалось найти доступного ПО, использующего эти методы. К достоинствам можно отнести гораздо меньшую вычислительную сложность некоторых из этих алгоритмов.

Этим же достоинством обладает и алгоритм латентной Дирихле-аллокации (Latent Dirichlet allocation; Blei et al 2003)). Он реализован в ПО Stanford Topic Modelling Toolbox, разработанном специально как основа для тематической кластеризации текстов (Ramage 2009, 2010), и успешно опробован нами для работы с большими массивами. Этот алгоритм предполагает, что корпус текстов содержит N скрытых тем. На основе анализа совместной встречаемости слов и моделирования их распределений в предполагаемых темах-кластерах, алгоритм приписывает каждому слову вероятность принадлежности к каждой теме, а затем, на основе анализа частотно-лексических распределений каждого текста, приписывает вероятность принадлежности каждого текста к каждому кластеру. Таким образом, каждая тема оказывается представлена списком слов с убывающей вероятностью принадлежности, которые затем могут быть «раскластеризованы» по темам к четким, так и нечетким образом, в зависимости от выбранного алгоритма. Однако поскольку список топ-слов уже дает исследователю представление о содержании темы, эта процедура не обязательна. Также, каждый текст оказывается представлен в

виде набора вероятностей принадлежности к теме, после чего тексты могут быть «раскластеризованы» по темам, как четким, так и нечетким образом. Достоинством такого подхода, во-первых, является его малая вычислительная сложность, которая достигается именно за счет экономичного представления текстов в виде набора вероятностей, количество которых равно количеству искомых тем. Во-вторых – это возможность нечеткой кластеризации итоговой матрицы вероятностей, недоступная для более тяжелых алгоритмов. Однако и этот алгоритм несвободен от проблемы поиска правильного числа кластеров.

Как видно, большинство современных алгоритмов сохраняет проблему определения количества кластеров. Попытки внедрения критериев остановки для иерархической кластеризации и апробирования таких алгоритмов со сравнительной оценкой их качества предпринимались достаточно давно (Milligan & Cooper 1985). На данный момент алгоритмы со встроенными критериями остановки описаны и апробированы Дж. Кариписом (Zhao & Karypis 2002) и реализованы в его ПО gCLUTO (Rasmussen & Karypis 2004), которое сейчас также опробуется нами. Однако это единственное ПО такого рода, которое сейчас с трудом настраивается нами на русский язык; все остальные ПО требуют самостоятельного определения числа кластеров. В связи с этим возникает особенно острая потребность в мерах оценки качества кластеризации.

Способы оценки качества алгоритмов

Наиболее распространенными внешними мерами оценки качества кластеризации текстов являются меры, заимствованные из информационного поиска: Precision (точность), Recall (полнота) и интегральный индекс F-мера. Как уже говорилось, в нашем случае очень трудно создать образцовый корпус текстов, во-первых, из-за больших объемов. Вручную классифицировать можно лишь небольшую выборку, но нет никаких гарантий, что в ней будут представлены все темы, поскольку структура генеральной совокупности не известна. Во-вторых, состав тем также заранее не известен, а привнесение представлений исследователя о возможных темах исключает обнаружение новых тем.

Что касается внутренних мер оценки качества, то здесь налицо отсутствие не только доминирующего способа, но и вообще хоть какого-либо единства в исследовательском сообществе. Отчасти это связано с тем, что кластерный анализ имеет дело с очень разными объектами, разделяемыми на группы для очень разных целей. Для одних задач кластеризации важна внутренняя связность кластеров, для других – их разделимость и т.д. Множество мер было разработано уже для классических

алгоритмов кластеризации; уже достаточно давно сравнительный анализ около 30 из них был проведен тем же Миллиганом (Milligan 1981); наилучшем был признан индекс gamma (Baker & Hubert 1972). Для современных алгоритмов разработаны и продолжают разрабатываться другие меры, но пока они слабо внедрены в ПО, доступное не-математику.

Заключительные замечания

В этой работе рано ставить точку. На данный момент нам удалось реализовать автоматизированный сбор и препроцессинг больших данных, разобраться в общих достоинствах и недостатках существующих алгоритмов и связанных с ними ПО. Но большая часть работы впереди: нам предстоит оценить пригодность алгоритмов для социологических задач, получить социологические значимые выводы.

Литература:

Этлинг Б., Алексанян К., Келли Дж., Палфри Дж., Гассер У. Публичный дискурс в россифской блогосфере: анализ публичной политики и мобилизации // Исследования центра Беркмана No 2010-11, 19 октября 2010. URL: http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere-RUSSIAN.pdf. URL оригинала на англ.яз .: http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere (дата обращения: 21.11.2011).

Ahlquist J.S., Breunig C. Model-Based Clustering and Typologies in the Social Sciences. Aug. 15, 2011. URL: https://mywebspace.wisc.edu/jahlquist/web/ModelBasedClusteringPoliticalAnalysisAug2011.pdf (дата обращения: 21.11.2011).

Alexanyan K., Koltsova O. Blogging in Russia is not Russian blogging // Russel A., Echchaibi N. (eds) International Blogging: Identity, Politics and Networked Publics. Peter Lang, 2009.

Andrews N.O, Fox E.A., Recent Developments in Document Clustering, October 16, 2007. URL: http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf (дата обращения: 21.11.2011).

Baker F.B., Hubert L.J. Measuring the Power of Hierarchical Cluster Analysis // Journal of American Statistical Association, 70:31-38, 1972.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John. Latent Dirichlet allocation // Journal of Machine Learning Research 3: pp. 993–1022. doi:10.1162, 2003.

Bruns, A.; Adams, D. Mapping the Australian Political Blogosphere // Russel, A., Echchaibi, N. (eds) International Blogging: Identity, Politics and Networked Publics. NY: Peter Lang Publishin, 2009. P. 85-110.

Carpineto C., Osiński S., Romano G., Weiss D. A survey of Web clustering engines // ACM Computing Surveys (CSUR), Volume 41, Issue 3 (July 2009), Article No. 17.

Gorny E. Russian LiveJournal: National specifics in the development of a virtual community. Version 1.0 of 13 May 2004 // Russian-cyberspace.org. URL http://

www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf (дата обращения: 21.11.2011).
Milligan G.W., Cooper M.C. An examination of procedures of determining the number of clusters in data set // Psychometrika Vol. 50, No. 2, 159-179, June 1985.
Milligan, G.W. A Review of Monte Carlo Tests of Cluster Analysis // Multivariate Behavioral Research, 16, 379–407, 1981.
Papacharissi, Z. Audiences as Media Producers: Content Analysis of 260 blogs // Tremayne, Mark (ed). Blogging, Citizenship and the Future of Media. NY&London: Routledge, 2007.
Ramage D., Dumais S., Liebling D. Characterising Microblogs with Topic Models // ICWSM 2010. URL: http://www.stanford.edu/~dramage/papers/twitter-icwsm10.pdf (дата обращения: 21.11.2011).
Ramage D., Rosen E., Chuang J., Manning C.D., McFarland D.A.. Topic Modeling for the Social Sciences // NIPS 2009 Workshop on Applications for Topic Models. URL: http://www.stanford.edu/~dramage/papers/tmt-nips09.pdf (дата обращения: 21.11.2011).
Rasmussen M., Karypis G. gCLUTO: An Interactive Clustering, Visualization, and Analysis System // UMN-CS TR-04-021, 2004.
Rogers R., Mapping Public Web Space with the Issuecrawler // Brossard C., Reber B. (eds.) Digital Cognitive Technologies: Epistemology and Knowledge Society. London: Wiley, 2010, 115-126.
Zhao Y., Karypis G. Criterion Functions for Document Clustering: Experiments and Analysis // University of Minnesota, Department of Computer Science / Army HPC Research Center. Minneapolis, MN 55455. Technical Report #01-40.

М.А. Захарова

СОЦИАЛЬНО-КУЛЬТУРНОЕ ВЗАИМОДЕЙСТВИЕ В ИНТЕРНЕТ-СРЕДЕ

Рассмотрим определение понятия «виртуальное сообщество». Г.Рейнгольд предлагает следующее: «Виртуальные сообщества – это общественные объединения, которые появляются в сети, когда определенное (достаточно большое) число людей на протяжении длительного времени участвует в публичных дискуссиях, испытывая при этом свойственные человеку эмоции и создавая сеть личных контактов» (Рейнгольд 2006: 416).

Таким образом, виртуальные сообщества определяются как естественные социальные образования, взаимодействие внутри которых протекает преимущественно в глобальной компьютерной сети. М. Кастельс, ссылаясь на мнение Г. Рейнгольда, определяет виртуальные сообщества