

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования

**Омский государственный университет
им. Ф. М. Достоевского**

Исторический факультет
Кафедра социологии

Нагорный Олег Станиславович

**Методы text mining
в социологии**

КУРСОВАЯ РАБОТА
СТУДЕНТА 4 КУРСА

Научный руководитель
кандидат социологических наук
К.В. Павленко

Омск 2014 г.

Содержание

Введение	2
1 Теоретическая часть. Text mining как метод анализа данных	4
1.1 Место text mining в структуре исследовательских методов	4
1.1.1 Дуальность статистики	4
1.1.2 Data mining как объединение подходов	9
1.2 Методология text mining	10
1.3 Область применения и примеры использования методов text mining	13
1.4 Отличие text mining от контент анализа	15
2 Практическая часть. Исследование образа губернатора омской области в местных Интернет-СМИ	19
2.1 Определение целей исследования	19
2.2 Оценка доступности и характера данных. Сбор данных.	19
2.3 Предварительная обработка данных	22
2.4 Анализ данных	25
Литература	26

Введение

Последние несколько десятилетий наука анализа данных претерпевает самые существенные изменения.

С одной стороны, появление глобальной сети Интернет и распространение персональных компьютеров привело к тому, что информации стало больше и производится она намного быстрее, до его возникновения. Значительная часть человеческой коммуникации переместилась в виртуальную сферу. Практически у каждой газеты или журнала имеются или электронная версия номера, или веб-сайт, где постоянно появляются новые материалы, происходит коммуникация пользователей между собой и с редакцией, проводятся голосования и прямые трансляции. Некоторые СМИ и вовсе отказываются от бумаги и полностью перебираются в электронный формат. Предоставляя более удобные средства потребления, хранения и поиска информации, чем традиционные печатные СМИ, Интернет становится новым центром притяжения как для издателей, так и для их аудитории.

С другой стороны, благодаря развитию технических средств и совершенствованию алгоритмов оперировать информацией стало проще. Обычный персональный компьютер теперь способен обрабатывать миллионы строк текста за считанные секунды.

Эти изменения открывают перед исследователями невиданные ранее перспективы. На основе наработок в области искусственного интеллекта, машинного обучения, статистики и проектировании баз данных в 80-х гг. XX века сформировалась новая междисциплинарная область знания — Data Mining или интеллектуальный анализ данных. Особенность методов, объединяемых данным понятием, заключается в их способности извлекать из «сырых» данных ранее неизвестные нетривиальные знания. Системы Data Mining сейчас находятся на острие исследований и разработок в области анализа, моделирования и практического использования информации и знаний, создавая новую культуру анализа данных.

Сфера применения данных методов практически ничем не ограничена — их можно применять везде, где имеются какие-либо данные [1, стр. 81]. Одной из таких сфер применения является интеллектуальный анализ данных — прежде всего текста — в социальных науках. Группа методов Data Mining, предназначенная для интеллектуального анализа неструктурированного текста объединяется под названием Text Mining.

В социологии анализ текстов обычно осуществляется следующими традиционными методами: дискурс-анализ, контент-анализ, когнитивное картирование и т.п. Однако, как уже говорилось, виртуальное пространство является хранилищем огромного количества текстов. Поэтому обрабатывать и анализировать их обычными, привычными для социологов методами не представляется возможным. Здесь на помощь социальному исследователю могут прийти методы text mining. С помощью Text Mining можно получить результаты, недоступные классическим методам анализа данных, например, с высокой точностью спрогнозировать результаты выборов¹ или предсказать популярность фильма до выхода в прокат на основе его обсуждения в сети².

¹Прогноз выборов в Венесуэле. URL: <http://vox-populi.ru/venezuela.phtml>

²Predicting the Future With Social Media. URL: <http://www.hpl.hp.com/research/sci/papers/socialmedia/socialmedia.pdf>

Однако по оценке некоторых учёных, многие российские социологи не знакомы с данными методами, что нельзя признать нормальным, поскольку «отбрасывает» отечественную социологию на 20-30 лет назад. Отсутствие соответствующей подготовки в области анализа данных приводит к поверхностному анализу эмпирических данных, в то время как важные и полезные неочевидные закономерности в данных «ускользают» от внимания исследователя [2]. Такое игнорирование современных методов анализа данных вполне может стать «фатальной ошибкой»³ и привести к возникновению «чёрной дыры»⁴ в российской социологии. Сказанное позволяет считать, что работа, показывающая перспективы применения методов Data Mining в социологических исследованиях, является **актуальной**.

В данном исследовании мы ставим цель рассказать о таком методе анализа данных как text mining и на практическом примере показать его актуальность для социологического анализа.

Проблема исследования заключается в недостаточности наработок в области применения методов Data Mining в социологии.

Объект исследования — методы Text Mining в социологическом исследовании.

Предмет исследования — возможности применения Text Mining для задач обработки естественного языка, моделирования тем, анализа настроений и неструктурированного текста в социологическом исследовании.

³Давыдов А. А. Фатальная ошибка социологии. URL: <http://ecsocman.hse.ru/text/28973359/>

⁴Орлов А. И. Черная дыра отечественной социологии. URL: http://www.ssa-rss.ru/index.php?page_id=19&id=456

Глава 1

Теоретическая часть. Text mining как метод анализа данных

1.1 Место text mining в структуре исследовательских методов

Если данные говорят с вами,
значит вы — байсовец.

Филип А. Шродт [3, стр. 11]

Bayes' theorem is nominally a
mathematical formula. But it is
really much more than that. It
implies that we must think
differently about our ideas.

Нэт Сильвер¹. The Signal and the
Noise.

В грамм добыча, в годы труды.
Изводишь единого слова ради
Тысячи тонн словесной руды.

В. В. Маяковский

1.1.1 Дуальность статистики

Дуальность статистики берёт своё начало из философского спора Аристотеля и Платона [4, стр. 7]. Аристотель считал, что реальность может быть познана только эмпирически и что исследователь должен тщательно изучать вещественный мир вокруг себя. Он пришёл к убеждению, что можно разложить сложную систему на элементы, детально описать эти элементы, соединить их вместе и, затем, понять целое. Именно таким механистичным

¹Американский статистик, давший самые точные прогнозы президентских выборов в США в 2008 и 2012 гг. Входит в 100 самых влиятельных людей в мире по версии журнала Times.

путём долгое время следовала наука. Однако в дальнейшем стало понятно, что не всегда целое можно представить как простую сумму частей, его составляющих. Часто, будучи соединёнными вместе, совокупность этих частей приобретает новое качество.

В отличие от своего ученика, Платон считал что свойством подлинного бытия обладают только идеи, а человек может лишь воспринимать и воплощать в вещах их смутные очертания. Для Платона идея (целое) была большим, чем сумма её материальных проявлений.

Эта дихотомия восприятия реальности проявляется во многих аспектах человеческой мысли, в том числе и в сфере статистического знания, в котором с XVIII в. существует две основных философских позиции относительно того, как применять вероятностные модели. Первая определяет вероятность как нечто, заданное внешним миром. Вторая утверждает, что вероятность существует в головах людей. [5, стр. 18]. В русле первого подхода возникли вначале классическая и затем развивающая её частотная концепции вероятности. Вторым подходом нашёл выражение в концепции байесовской вероятности.

Сторонники классического подхода исходят из того, что истинные параметры модели не случайны, а аппроксимирующие их оценки случайны, поскольку они являются функциями наблюдений, содержащих случайный элемент. [6, стр. 5-6] Параметры модели считаются не случайными из-за того, что классическое определение вероятности исходит из предположения равновозможности как объективного свойства изучаемых явлений, основанного на их реальной симметрии [7, стр. 24]. На такое представление о вероятности повлияло то, что в начале своего развития теория вероятности применялась прежде всего для анализа азартных игр. Суждение вида «Вероятность выпадения шестёрки при бросании игрального кубика равняется $1/6$ » основывается на том, что любая из шести граней при подбрасывании на удачу не имеет реальных преимуществ перед другими, и это не подлежит формальному определению. Таким образом, вероятностью случайного события A в её классическом понимании будет называться отношение числа несовместимых (не могущих произойти одновременно) и равновозможных элементарных событий m к числу всех возможных элементарных событий n :

$$P(A) = \frac{m}{n} \quad (1.1)$$

Однако такое определение наталкивается на некоторые непреодолимые препятствия, связанные с тем, что не все явления подчиняются принципу симметрии. Например, из соображений симметрии невозможно определить вероятность наступления дождливой погоды. Для преодоления подобных трудностей был предложен статистический или частотный способ приближённой оценки неизвестной вероятности случайного события, основанный на длительном наблюдении над проявлением или не проявлением события A при большом числе независимых испытаний и поиске устойчивых закономерностей числа проявлений этого события. Если в результате достаточно многочисленных наблюдений замечено, что частота события A колеблется около некоторой постоянной, то мы скажем, что это событие имеет вероятность. Данный тип вероятности был выражен Р. Мизесом в следующей математической формуле:

$$p = \lim_{x \rightarrow \infty} \frac{\mu}{n}, \quad (1.2)$$

где μ — количество успешных испытаний, n — количество всех испытаний [7, стр. 46-47]. Вероятность здесь понимается как частота успешных исходов и является чисто объективной мерой, поскольку зависит лишь от точного подсчёта отношения количества успешных и неуспешных событий.

Основываясь на этом подходе, статистика занималась созданием вероятностных моделей, которые включали в себя параметры, которые, как предполагалось, связаны с харак-

теристиками исследуемой выборки. Параметры никогда не могут быть известны с абсолютной точностью до тех пор, пока мы не исследуем всю генеральную совокупность [5, стр. 1]. До тех пор всегда существует вероятность отклонить гипотезу, когда она на самом деле верна, т. е. совершить ошибку первого рода. Для обозначения вероятности такой ошибки частотники используют понятие уровня значимости α . Именно вероятность ошибки первого рода частотники ставят во главу анализа, определяя вероятность события. После каждого своего утверждения они обычно добавляют «... на доверительном уровне в 95%», подразумевая, что исследователь допускает вероятность ошибки в пяти процентах случаев (при $\alpha = 0,05$) [4, стр. 10-11].

Иногда параметры вообще не возможно интерпретировать применительно к реальной жизни, поскольку модели редко бывают абсолютно верными. Модели, как мы надеемся, — это некоторые полезные приближения к истине, на основании которых можно делать прогнозы. Тем не менее прежде всего классическое статистическое исследование сосредоточено на оценке параметров, а не на предсказании [5, стр. 1].

Частотный подход доминировал в XX веке, придя на смену другому пониманию вероятности, связанном с именем английского математика Томаса Байеса [8, стр. 2]. Сущность байесовского подхода составляют три элемента: априорная вероятность, исходные статистические данные, постаприорная вероятность.

Байесовская статистика начинает построение своей модели при помощи понятия априорной вероятности, с помощью которой описывается текущее состояние наших знаний, относительно параметров распределения [5, стр. 18]. Априорная вероятность, таким образом, — это степень нашей уверенности в том, что исследуемый параметр примет то или иное значение ещё до начала сбора исходных статистических данных. На этом основании байесовское понимание вероятности относят к группе субъективистских трактовок вероятности. Чаще всего предполагается, что для оценки степени уверенности необходимо привлечь экспертов, чьё субъективное свидетельство позволит избежать действительной многократной реализации интересующего нас эксперимента² [10, стр. 34].

Следующий элемент — это исходные статистические данные. По мере их поступления статистик пересчитывает распределение вероятностей анализируемого параметра, переходя от априорного распределения к апостериорному, используя для этого формулу Байеса:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1.3)$$

где $P(A)$ — априорная вероятность гипотезы A , $P(A|B)$ — вероятность гипотезы A при наступлении события B (апостериорная вероятность), $P(B|A)$ — вероятность наступления события B при истинности гипотезы A , $P(B)$ — полная вероятность наступления события B . Суть формулы в том, что она позволяет переставить причину и следствие: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Эту формулу также называют формулой обратной вероятности. Процесс

²Не следует путать субъективный характер байесовской вероятности в целом с внутренним разделением сторонников данного подхода на объективистов и субъективистов, основанном на различном отношении к роли рациональных ограничений при определении априорной вероятности. В качестве примера различного подхода к определению априорной вероятности рассмотрим ситуацию, где событием является изъятие мячика из урны, наполненной красными и чёрными мячиками — и это всё, что нам известно об урне. Зададим вопрос: какова априорная вероятность (до изъятия мячика), что изъятый мячик будет чёрного цвета? Субъективисты, считающие роль рациональных ограничений относительно небольшой, ответят, что любая вероятность от 0 до 1 может быть рациональной, так как по их мнению наша оценка априорной вероятности зависит большей частью от нерациональных факторов — социализации, свободного выбора и др. Объективисты же будут настаивать, что априорная вероятность в данном случае равняется 1/2, поскольку именно такая вероятность в соответствии с принципом неопределённости Джейнса инвариантна к к размерам и трансформациям мячиков [9].

пересмотра вероятностей, связанных с высказываниями, по мере поступления новой информации составляет существо **обучения на опыте**³ [6, стр. 21-22] и является одним из возможных способов формализации и операционализации следующего тезиса: *«степень нашей разумной уверенности в некотором утверждении (касающемся, например, неизвестного численного значения интересующего нас параметра) возрастает и корректируется по мере пополнения имеющейся у нас информации относительно исследуемого явления»* [11, стр. 93]. В частотном подходе данный тезис интерпретируется в свойстве состоятельности оценки неизвестного параметра: чем больше объём выборки, на основании которой мы строим свою оценку, тем большей информацией об этом параметре мы располагаем и тем ближе к истине наше заключение. Специфика байесовского подхода к интерпретации этого тезиса основана на том, что вероятность, понимаемая как количественное значение степени разумной уверенности в справедливости некоторого утверждения, пересматривается по мере изменения информации, касающейся этого утверждения. Поэтому в данном подходе вероятность всегда есть условная вероятность, при условии нынешнего состояния информации (в русле классического подхода исследователь скорее склонен рассматривать совместную вероятность [4, стр. 5]).

Дискуссии вокруг того, какой же метод предпочтительней, ведутся уже не одно столетие, породив великое множество книг и статей на эту тему [12], [8], но к однозначному выводу прийти не удалось. Острота дискуссии объясняется тем, что спор сторонников байесовского и частотного подхода к статистическому выводу отражает два различных взгляда на способ добычи научного знания. Именно поэтому от ответа на этот, казалось бы, локальный вопрос математической статистики зависит развитие всей науки.

Так или иначе, в 1980-х годах, стало ясно, что частотный подход к статистическому выводу не достаточно хорошо подходит для анализа нелинейных отношений в больших объёмах данных, производимых сложными системами при моделировании процессов реального мира [4, стр. 10]. Для преодоления этих ограничений частотники создали нелинейные версии параметрических методов, такие как множественный нелинейный регрессионный анализ.

В то время как в частотном подходе происходили изменения, немногочисленные сторонники байесовского подхода упрямо продвигали свою точку зрения на модель статистического вывода. Как оказалось, байесовская модель лучше подходит для поиска ответов на некоторые практические вопросы, поскольку полнее учитывает прошлую информацию и располагает к предсказаниям. Например, намного важнее минимизировать вероятность ложноотрицательного диагностирования некоторой опухоли как раковой, чем вероятность её ложноположительного определения (ошибка первого рода).

Продemonстрируем на примере различия в работе частотных и байесовских методов проверки гипотез. Предположим, некоторый стрелок утверждает, что точность его стрельбы составляет 75%. Когда стрелка попросили продемонстрировать свои навыки, он попал в мишень только 2 раза из 8. Какова вероятность, что стрелок сказал правду о своих навыках.

Решение задачи в частотном подходе. Гипотеза H_0 — стрелок сказал правду. Испытание — стрельба по мишени. Событие A — попадание в мишень. $P(A)$ постоянная и равна 0,75. Для расчёта вероятности того, что событие A наступило не более 2 раз в 8 независимых испытаниях, применим формулу Бернулли для количества успешных испытаний $k = 0, 1, 2$ и получим, что $P(A \leq 2) = 0,0042$. Следовательно, при уровне значимости $\alpha = 0,05$ следует признать невероятным, что точность стрелка составляет 75%, гипотеза H_0 отвергается.

³Понятие «обучение на опыте» ещё не раз встретится в данной работе, поскольку именно оно составляет суть машинного обучения — подраздела науки искусственного интеллекта, методы которого используются в text-mining.

Отметим некоторые особенности данного решения. Во-первых, для решения задачи мы фактически использовали только умение рассчитывать совместную вероятность, ведь формула Бернулли является сокращённым видом расчёта совместной вероятности успешных комбинаций. Во-вторых, мы решили, что если гипотеза верна, то вероятность отклонить гипотезу, когда она на самом деле верна должна быть не менее 5%, т. е. нам важно, чтобы вероятность ложноположительного ответа была ниже определённой границы. Вероятность ложноотрацательного ответа не рассматривается.

Решение задачи в байесовском подходе. В данном подходе мы не проверяем гипотезу, а рассчитываем условную вероятность события A (точность стрелка составляет 75%) при условии события B (стрелок попал в мишень не более 2 раз из 8). Прежде всего нам нужно оценить априорную вероятность события A . Это можно сделать посмотрев статистику стрельбы остальных стрелков. Предположим, мы выяснили, что 70% стрелков имеют точность в 75%. Следовательно, $P(A) = 0,7$. $P(B|A)$ мы уже рассчитали в частотном подходе. $P(B)$ легко рассчитывается по формуле полной вероятности. По формуле Байеса $P(A|B) = 0,0301$.

Как видно из этого примера, в байесовском подходе другая логика расчёта вероятности: на основании данных рассчитывается вероятность того, что H_0 верна, в то время как раньше мы рассчитывали вероятность того, что стрелок поразил мишень не более 2 раз в 8 независимых испытаниях. Данные, полученные с помощью данного метода, данные можно использовать более продуктивно. Предположим, что мы рассчитываем не вероятность того, что стрелок с определёнными умениями поразил мишень какое-то количество раз, а вероятность наличия тяжёлого заболевания у человека с каким-то количеством положительных тестов. В случае частотного подхода мы узнаем, какова вероятность того, что больной человек получит n -ое количество положительных тестов. Байесовский же подход позволяет узнать именно то, что нам надо — вероятность того, что человек, получивший n -ое количество положительных тестов, болен. Другой плюс данных методов — они работают даже если размер выборки равен нулю. В таком случае байесовская вероятность равна априорной.

Проведение тестирования на статистическую значимость оценивает лишь вероятность получения похожего результата с другим набором данных при сохранении тех же самых условий. Однако оно предоставляет ограниченную картину такой вероятности, поскольку в расчет принимается ограниченное количество информации относительно исследуемых данных. И оно само по себе не способно вам сказать, являются ли основные положения исследования верными и будут ли подтверждены полученные результаты в различных условиях⁴. Уровень p говорит только о вероятности получения результата при (обычно) совершенно нереалистичных условиях нулевой гипотезы. А это совсем не то, что мы хотим узнать, — обычно мы хотим знать величину эффекта независимой переменной с учетом имеющихся данных. Это байесовский вопрос, а не частотный. Вместо этого значение p часто интерпретируется так, будто бы оно показывало силу ассоциации [3, стр. 11].

С другой стороны у и байесовского метода имеются несколько недостатков. Одним из них является необходимость привлекать для расчёта априорные данные, которые могут быть недоступны. А если они и доступны, то, как отмечалось выше, часто носят субъективный характер. Другой недостаток — сложность вычислений. В вышеописанном примере для вычисления байесовской вероятности нам необходимо было вычислить частотную вероятность, полную вероятность, и, наконец, собственно байесовскую вероятность. Сложность байесовских вычислений частично объясняет тот факт, что байесовские методы вновь обрели популярность с развитием вычислительной техники. Следующий недостаток байесовского метода — неинтуитивность, непонятность его результатов для обыден-

⁴Роль статистической значимости в неудачах науки. URL: <http://inosmi.ru/world/20131114/214743342.html>

ного сознания. Именно на этой неинтуитивности построен знаменитый парадокс Монти Холла, который легко решает с помощью формулы байеса.

1.1.2 Data mining как объединение подходов

Дальнейшее развитие статистических методов, особенно в их байесовском варианте, привело к возникновению следующего поколения методов статистического анализа, а именно методов машинного обучения. Первоначально эти методы развивались в двух направлениях, первое из которых представлено искусственными нейронными сетями, а второе — деревьями принятия решений [4, стр. 11-12].

Развитие методов машинного обучения в свою очередь привело к созданию статистической теории обучения (Statistical Learning Theory), которая направлена на решения проблемы предсказания на основе имеющихся данных [4, стр. 12-13].

Какое место занимают методы Data Mining в описанной структуре? DM — это междисциплинарная область знания, находящаяся на пересечении традиционного статистического анализа, искусственного интеллекта, машинного обучения и развития больших баз данных [4, стр. 5]. Можно даже сказать, что DM — это новая философия, новый взгляд на анализ данных.

Хотя как самостоятельная дисциплина DM окончательно оформился в 1990-х гг. [4, стр. 15], о важности ухода от чистой математической статистики в пользу анализа реальных данных говорил ещё Джон Тьюки, который в 1962 году написал статью под названием «Будущее анализа данных» (The future of data analysis), в которой изложил основные идеи новой тенденции. Тьюки говорил о том, что излишняя сосредоточенность на математических теориях в статистике не помогает в решении реальных жизненных проблем. Он был убеждён, что анализ данных — это работа, схожая с работой следователя и что надо дать данным говорить самим за себя. Однако эти идеи тогда не были восприняты приверженцами чистой математической статистики, которые утверждали, что правильная процедура статистического анализа прежде всего предполагает выдвижение научных гипотез, а затем уже их проверку, на основе полученных данных. Попытка анализа данных до выдвижения гипотезы категорически отвергалась, поскольку считалось, что это приведёт к смещению гипотезы в сторону того, что показали данные. Такая позиция привела к тому, что термин «DM» стали использовать в уничижительном значении [13, стр. 788].

Развитие информационных технологий и вычислительной техники с одной стороны привело к появлению огромного количества данных, а с другой — предоставило инструменты для их удобного сбора, хранения и обработки. Эти процессы также изменили течение академических споров, поскольку учёные осознали перспективы новой парадигмы анализа данных. Почему же DM стал популярен в сложившихся условиях?

Суть философии DM частично выражена в названии этой области знания, которое состоит из двух понятий: поиск ценной информации в большой базе данных (data) и добыча горной руды (mining). Именно в просеивании через сито своих инструментов огромного количества «сырых», часто неструктурированных данных в поисках самородков, т. е. осмысленной, нетривиальной информации — знаний. Более верным названием для этого процесса было бы «knowledge mining from data» (добыча знаний из данных) [14, стр. 5].

Исходное определение термина, которое дал наш бывший соотечественник Григорий Пятнецкий-Шапито, звучит следующим образом: «Data mining — это процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1, стр. 78].

В статистике Data Mining часто иногда отождествляют с таким процессом как Knowledge Discovery in Databases, в то время как компьютерщики (computer scientists) предпочитают рассматривать первое определённую как часть второго.

1.2 Методология text mining

По аналогии с термином Data mining термину Text mining можно дать следующее определение – это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных [15, стр. 211].

Главная цель text mining состоит в обработке неструктурированного текста и, если это требует решаемая с помощью данного метода проблема, слабоструктурированных и структурированных данных, с тем, чтобы извлечь новое, значимое и применимое знание для лучшего принятия решений [16, стр. 78].

Так как по сравнению с остальными устоявшимися статистическими методами text mining является относительно новой и неустоявшейся областью знания, сложно говорить, о наличии единой и общепринятой совокупности методов, направленных на получение устойчивого результата, т. е. о методологии. Во многом, исследователи, использующие методы text mining, руководствуются собственным опытом, приобретённым методом проб и ошибок, и создают собственную методологию. Наиболее значимые причины такого волюнтаризма включают следующее [16, стр. 74]:

- Разные исследователи вкладывают в понятие text mining разные значения. Данное определение ещё только формируется.
- Неструктурированный характер данных открывает широкие возможности для действий исследователя.
- Существует несколько форматов неструктурированных данных, некоторые из которых могут быть классифицированы как полуструктурированные (HTML, XML, JSON и другие).
- Огромные объёмы данных часто требуют сокращения и упрощения.

Самым популярным вариантом методологии Data-Mining является CRISP-DM (CRoss Industry Standard Process for Data Mining) – Стандартный межотраслевой процесс Data Mining. Так как главное отличие Text Mining от Data Mining заключается в том, что первый специализируется на определённом типе данных, с небольшими изменениями CRISP-DM можно применить и для него. Весь цикл обработки данных этой методологии представлен шестью последовательными этапами [16, стр. 74].

Этап 1. Определение целей исследования. С этого начинается практически любая осмысленная деятельность. Грамотная постановка цели требует глубокого понимания всех аспектов ситуации, в которой проводится исследование и чёткого определения результата, который мы хотим получить. Для этого необходимо изучить проблему, на решение которой направлено исследование.

Этап 2. Оценка доступности и характера данных. Данный этап включает в себя следующие задачи:

- Определение источников текста. Текст может иметь цифровую форму или написан на бумаге, находится внутри или за пределами исследуемой организации.

- Оценка доступности и применимости данных.
- Сбор первичных данных.
- Оценка содержательности данных (содержится ли в них необходимая для исследования информация).
- Оценка количества и качества данных.

После того, как разведывательная часть исследования успешно завершена, можно приступить к сбору данных из различных источников.

Этап 3. Подготовка данных. Подготовка данных – необходимый для text mining этап. По мнению многих, специфика данного метода по сравнению с data mining заключается в более трудоёмких стадиях сбора и обработки данных [16, стр. 77]. Данный этап состоит из следующих фаз:

Создание корпуса. В лингвистике корпус – это большой структурированный набор текстов. На данном этапе необходимо собрать все текстовые документы, относящиеся к исследуемой проблеме. Исследователю предстоит решить, какие данные и в каких объемах необходимо собрать и проанализировать, чтобы решить поставленную задачу. Следует помнить, что методы data mining сильно зависят от точности полученных результатов от их количества.

После того, как документы будут собраны, их необходимо трансформировать таким образом, чтобы они были представлены в единой форме (например в базе данных или текстовом файле) для компьютерной обработки.

Предварительная обработка данных. На данном этапе мы должны решить одну из главных проблем анализа текстов, а именно большое количество лишних слов в документе [15, стр. 213], которые только создают помехи при включении их в анализ. Таким образом целью данного этапа будет удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду. При подготовке данных обычно используют следующие приёмы:

- Удаление стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа. Обычно заранее составляются списки таких слов, и в процессе предварительной обработки они удаляются из текста. Типичным примером таких слов являются вспомогательные слова и артикли, например: "так как "кроме того" и т. п.
- Стёмминг или лемматизация терминов, т. е. приведение их к простейшей форме, чаще всего к корню или к начальной форме слова (1-е лицо, единственное число, именительный падеж). Например, слова "социолог "социологический" и "социология" различны, но относятся к одной и той же теме. Вследствие процедуры стёмминга, основанного на приведении к корню, всё они будут приведены к одному термину "социолог". Это позволит сократить количество терминов и увеличить их частоту.
- Вышеописанные приёмы значительно уменьшают количество терминов в корпусе. Обычно после этого на основе корпуса создаётся матрица терминов (document-term matrix), строками которой являются отдельные документы корпуса, а колонками – уникальные термины. Соответственно в ячейках матрицы записывается число повторений терминов в документах. Представленные в таком виде текстовые данные удобно использовать для дальнейшего анализа.

Этап 4. Разработка и калибровка модели. На этом этапе происходит применение методов извлечения знаний. В text mining используется четыре основных метода: классификация, кластеризация, ассоциация, анализ трендов.

Классификация. Вероятно, наиболее распространённым методом, использующимся в интеллектуальном анализе данных является распределение объектов по классам согласно каким-либо важным признакам. В отношении к text mining эта задача известна как *категоризация текста* и заключается в нахождении верной темы или понятия для каждого документа из корпуса. Сегодня автоматическая категоризация текста применяется в контекста различных задач, включая фильтрацию от спама, определение жанра, категориацию веб-страниц в иерархических каталогах и многое другое.

Существует два основных подхода к классификации текста. В первом подходе знания экспертов о категориях кодируются в правила, на основе которых объект относится к тому или иному классу. Второй подход, пришедший из машинного обучения, построен на работе определённого алгоритма, который обучившись на уже классифицированном наборе данных, способен в дальнейшем с некоторой вероятностью определять класс остальных объектов.

Кластеризация. Кластеризация – это упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к классу задач обучения без учителя. Это означает, что в процессе кластеризации не используется какая-либо предварительная информация о характеристиках групп, которые должны получиться в итоге. В этом отличие кластеризации от классификации, где для определения класса объекта используется обучающая выборка или знания экспертов (происходит обучение с учителем).

Создание правил ассоциации. Ассоциация – это процесс поиска повторяющихся образцов в группе объектов. Этот метод используется в интернет магазинах, чтобы на основании выбранных пользователем товаров предложить ему другие варианты. Главная идея этого метода в том, чтобы определить, правила, на основании которых определённые и часто непохожие между собой объекты объединяются в единый набор.

В text mining данный метод используется чтобы измерить отношения между понятиями или группами понятий. В правиле ассоциации $X \Rightarrow Y$

Этап 5. Проверка результатов. После того, как модель создана и проверена, мы должны произвести общую проверку всех действий. Например, необходимо убедиться, что выборка произведена правильно. Также случается, что в процессе построения исследования теряется основная цель, для достижения которой оно начиналось. На данном этапе следует проверить, решает ли модель сформулированную проблему и служит ли, таким образом, достижению цели. Если что-то упущено, необходимо вернуться назад к этапу, породившему рассогласованность между целью и результатом.

Этап 6. Внедрение. В случае, если по итогам проверок было решено, что модель решает поставленную проблему, её можно применять. В самом простом случае внедрение может принимать форму написания отчета о результатах исследования. В сложном – построение интеллектуальной системы на основе построенной модели с тем, чтобы она могла быть повторно использована для принятия решений.

1.3 Область применения и примеры использования методов text mining

Интеллектуальный анализ текста находит своё применение во многих областях. В экономике с его помощью можно установить, как настроения в СМИ влияют на котировки фондового рынка [17], имеется ли связь между отзывами о продукте в Интернет-магазине и его продажами [18], как макроэкономические показатели могут быть измерены поисковыми запросами [19] и текстами из социальных медиа.

В психологии этот метод позволяет узнать, как психическое состояние человека выражается в его языке [20] и правда ли, что суточные и сезонные циклы настроения носят надкультурный характер [21].

Одним из самых известных и ранних примеров применения методов text mining в исторических исследованиях является установление авторства сборника статей «Федералист» [22]. Здесь text mining принял форму стилометрии. Другое исследование в области text mining продемонстрировало, что в XVIII понятием «литература» объединялся более широкий класс явлений, чем сегодня [23].

Социолингвисты использовали text mining для идентификации географически зависимых лингвистических переменных и, на основании этого, предсказания местоположения пользователя на основе написанного им текста [24].

Text-mining также можно использовать в качестве вспомогательного метода, уточняющего результаты традиционных опросов [25].

Рассматриваемый метод активно используется в политологических и социологических исследованиях. Так как в данной работе будет представлено исследование именно такого вида, рассмотрим из подробней.

В 2012 году было опубликовано работа, посвящённая выявлению политических предпочтений бельгийских Интернет-СМИ в ситуации политического кризиса [26]. Суть кризиса состояла в том, что на протяжении более чем полутора лет ведущие валлонские и фламандские партии не могли договориться о составе федерального правительства. Корпус документов, используемых в исследовании, составили 68 000 статей, опубликованные в онлайн версиях восьми крупнейших фламандских газет в период с начала 2011 года до завершения политического кризиса в октябре того же года. Помимо даты публикации, критерием выбора статьи для анализа служило наличие в ней ключевых слов. Такими ключевыми словами считались названия фламандских политических партий, имеющих по крайней мере одно место в парламенте, и имена их важнейших представителей.

Первичная обработка данных включала удаление дубликатов. Затем на основе тонального словаря из более чем 3000 прилагательных, которые чаще всего встречались в отзывах на товары и которые вручную были проранжированы по шкале полярности (1 – позитивное, -1 – негативное) и субъективности (0 – объективное, 1 – субъективное), в каждой статье был произведён анализ тональности упоминаний выбранных политических партий и политиках. Для этого подсчитывалась полярность каждого прилагательного в пределах двух предложений до и двух предложений после упоминания партии. Для уменьшения шума исключались прилагательные набравшие меньше 0,1 и больше -0,1 очка по шкале полярности. В результате было выделено 360 613 оценок.

Следующий шаг в данном исследовании – определение степени представленности и популярности политической партии. Степень представленности $coverage(e, s)$ политического субъекта e в газете s определялась как отношение количества статей газеты, где упоминалась данная партия, к количеству всех статей данной газеты A_s :

$$coverage(e, s) = \frac{\#\{a | a \in A_s \wedge e \in a\}}{\#A_s} \quad (1.4)$$

Популярность $popularity(e)$ политического субъекта e определялась через относительное количество голосов, отданных за неё в результате голосования в 2010 году $v(e)$:

$$popularity(e) = \frac{v(e)}{\sum_{e' \in \varepsilon} v(e')} \quad (1.5)$$

Популярность использовалась в качестве априорного распределения для расчёта степени склонности газеты к освещению определённой политической партии. Данная склонность определялась как разность между представленностью партии в газете и её реальной популярностью, определённой в результате выборов:

$$bias(e, s) = coverage(e, s) - popularity(e) \quad (1.6)$$

В результате данных манипуляций были выявлено, какие политические субъекты пользуются популярностью электронных СМИ в большей или меньшей степени, чем среди населения в целом.

Следующий шаг – выявление тональности упоминания политических партий и их представителей. Для каждого субъекта было подсчитано количество положительных и отрицательных отзывов, составлен график изменения тональности во времени.

В результате исследования при помощи методов анализа текстов были выявлены политические предпочтения главных фламандских новостных сайтов во время политического кризиса.

Более интеллектуальные методы были применены для выявления различий в освещении событий, приведших к восстанию 2011 года в Египте, египетскими государственными и негосударственными СМИ [27]. Материал для анализа составили более 29 000 новостных статей, вышедших в 2010–2011 годах. В методологической части работы был использован такой метод тематического моделирования как латентное размещение Дирихле (LDA), с помощью которого можно выполнить задачу категоризации документов. Алгоритм сам определяет оптимальное количество категорий (тем) и распределяет документы между ними.

Было показано, что правительственные СМИ при освещении таких событий акцентировали внимание на угрозе дестабилизации и терроризма и старались рассказывать о проведении реформ в стране. Независимые же СМИ наоборот были нацелены на мобилизацию в целях противостояния режиму и фактически игнорировали действия правительства. Таким образом, было доказано, что режим Хосни Мумбарака потерял контроль над медиадискурсом ещё до начала активной фазы протестов.

Существуют примеры использования методов text-mining и в отечественных исследованиях. Дальше всего в этой сфере продвинулись сотрудники НИУ-ВШЭ, в частности заведующая Лабораторией интернет-исследований Кольцова Елена Юрьевна. Исследовательский коллектив под её руководством в рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач» поставил перед собой задачу выявления на больших массивах данных русскоязычной блогосферы тематические кластеры постов (о чем говорят?) и сообществ, основанные на комментировании (кто с кем говорит?), а также выяснения того, совпадают ли комментовые сообщества с тематическими кластерами (т.е. основана ли общность комментирования на общности темы?).

Тестовой тематикой являлась тема Ислама. Эмпирический материал исследования составили 7941 статей топовых блогеров Живого Журнала за период 21-23 и 24-26 декабря 2011 года и комментарии к ним, собранные с помощью паука краулера «Blogminer». Выбор записей с таким временем написания был обусловлен тем, что именно в это время ожидалась реакция со стороны "населения" российской блогосферы на выборы в Государственную Думу, состоявшиеся 4 декабря.

Для анализа данных использовалась программа NodeXL. Сообщества выявлялись путём применения алгоритмов Вакита-Цуруми и Клозэ-Ньюмана-Мура в качестве контрольного.

После операции по выявлению сообществ, которая разделила полную сеть постов на отдельные подмножества исследователи отобрали несколько групп постов для качественного анализа. Его целью было установить, связаны ли посты, входящие в одну группу по смыслу (тематически) или каким-либо другим образом (принадлежат перу одного или нескольких авторов).

По результатам качественного изучения постов из автоматически составленных групп был сделан вывод, что гипотеза исследования не подтвердилась: не были найдены доказательства того, что комментовые сообщества интегрированные общими темами в Живом Журнале.

Несмотря на неподтверждение гипотезы исследования, участие в проекте дало исследователям богатый опыт в организации Интернет-исследований, в результате чего была написана статья «К методологии сбора Интернет-данных для социологического анализа» [28].

1.4 Отличие text mining от контент анализа

После поверхностного взгляда на методы text mining может сложиться впечатление, что они повторяют уже хорошо известный социологом контент-анализ. И правда, из-за своего широкого распространения термин «контент-анализ» иногда используют как обобщающий для всех методов систематического и претендующего на объективность анализа политических текстов и текстов, циркулирующих в каналах массовой коммуникации. Однако такое расширительное понимание контент-анализа неправомерно, поскольку существует ряд исследовательских методов которые не могут быть сведены к стандартному контент-анализу даже при максимально широком его понимании [29]. Обладая большим количеством общих черт, эти методы тем не менее имеют существенные отличия, что оправдывает их выделение в отдельные группы.

Что интересно, хотя методы контент-анализа и text mining имеют много общего, исследователи, работающие в каждой из этих двух сфер, редко ссылаются друг на друга. В литературе о text mining почти никогда не упоминаются методы контент-анализа и наоборот. Ещё сложнее найти источники, где сравниваются два данных метода. Как нам видится, причина такой ситуации прежде всего кроется 1) в недостаточной осведомлённости исследователей о методах интеллектуального анализа текста, 2) отсутствии однозначного определения как контент-анализа [30, стр. 156], так и (в ещё большей степени) интеллектуального анализа текста, 3) и, о чём уже говорилось выше, привычкой называть любой метод анализа текстов контент-анализом.

Итак, как соотносятся такие понятия как контент-анализ и text mining?

Если рассматривать временной критерий, то контент-анализ возник раньше text mining. В советской социологической литературе происхождение контент-анализа связывалось с именами У. Томаса и Ф. Знанецкого. Сейчас же многие зарубежные [31] и отечественные [32] исследователи отмечают, что он возник сто и более лет тому назад, упоминая опыт использования метода, очень близкого к этому, когда в XIII веке в Швеции был осуществлен анализ сборника из 90 церковных гимнов, прошедших государственную цензуру и приобретших большую популярность, но обвиненных в несоответствии религиозным догматам. Наличие или отсутствие такого соответствия и определялось подсчетом в текстах этих гимнов религиозных символов и сравнения их с другими религиозными текстами, в том числе тех, которые считались еретическими. Частота использования определённых заранее собранных слов и тем позволяла судить о том, насколько корректен

текст с точки зрения официального учения церкви. Как сформировавшийся метод анализа контент-анализ впервые был использован Максом Вебером в 1910 году для анализа освещаемости прессой политических акций в Германии [30, стр. 155].

История методов text mining насчитывает гораздо меньше времени, поскольку тесно связана с развитием вычислительной техники и сопутствующих ей дисциплин, таких как искусственный интеллект, обработка естественного языка. Развитие информационных технологий привело к взрывообразному росту количества информации. Этот рост иллюстрирует тот факт, что количество веб-страниц в Интернете возросло с 10 миллионов в 2001 г. до 150 миллиардов в 2009 [16, стр. 4]. Для описания нового характера данных, которые отличаются большим объёмом, высокой скоростью роста и значительным многообразием своих форм, в специальном номере журнала «Nature» от 3 сентября 2008 года был введён термин «большие данные» (big data). Феномен «больших данных» создал потребность в новых методах обработки и анализа, способных извлекать полезное знание из ранее невиданных объёмов неструктурированной текстовой информации. Можно сказать, что методы text mining предназначены в первую очередь для анализа «больших данных».

Появление таких данных поставило проблему доступности необходимой информации. Эта проблема решалась двумя способами: с помощью информационного поиска (information retrieval) и извлечения информации (information extraction). Суть первого процесса состояла в выявлении в некотором множестве документов всех тех, которые удовлетворяют поисковому запросу. Второй процесс заключался в извлечении необходимых данных из набора документов [16, стр. 5].

Решение этой проблемы заключалось в выполнении двух основных задач. Вначале надо произвести суммаризацию текстов (создание для каждого текста краткого содержания), которая бы уменьшила их размер и, таким образом, облегчила обработку. Затем необходимо классифицировать полученные содержания по некоторым категориям [16, стр. 5].

Подходы к доступу к текстовой информации развивались под влиянием трёх дисциплин: библиотекведения, информационной науки (information science) и обработке естественного языка [16, стр. 5].

Одним из самых ранних примеров классификации текстов был библиотечный каталог, который представлял собой совокупность расположенных по определенным правилам библиографических записей на документы. Следующим шагом была суммаризация текста для составления аннотаций. В 1958 году Питер Лун (Peter Luhn) приспособил первый коммерческий компьютер для научных вычислений IBM 701 для автоматического составления аннотаций. Для этого с помощью частотного анализа он рассчитал относительную значимость слов, а затем отобрал предложения, в которых было больше всего значимых слов с наименьшим расстоянием между ними. С небольшими изменениями эта методика использовалась на протяжении десятков лет.

Создание в 1948 году Клодом Шенноном теории информации привело к возникновению информационной науки и дальнейшему развитию анализа текстов. Положения этой теории использовались для создания индекса цитирования научных статей, указывающего на значимость статьи и вычисляющегося на основе последующих публикаций, ссылающихся на данную работу. Дальнейшее использование теории информации применительно к анализу текстов повлияло на развитие поисковых систем.

Соединение информационной науки с лингвистикой породило такую гибридную дисциплину, как обработка естественного языка. Ещё задача состояла в понимании и моделировании естественного языка и, таким образом, разработки систем компьютерного перевода текста. С этой дисциплиной частично пересекается компьютерная лингвистика. Являясь ответвлением от науки искусственного интеллекта, она ставит своей целью использование математических моделей для описания естественных языков. Именно в дисциплине обра-

ботки естественного языка были разработаны такие широко используемые сейчас приёмы как токенизация и стемминг и метод кластеризации документов.

Другое различие между контент-анализом и интеллектуальным анализом текста заключается в различии их методологии. Из «Рабочей книги социолога» мы знаем, что контент-анализ относится к формализованным методам анализа документов, суть которых «сводится к тому, чтобы найти такие легко подсчитываемые признаки, черты, свойства документа (например, частота употребления определенных терминов), которые с необходимостью отражали бы определенные существенные стороны содержания с тем, чтобы сделать его доступным точным вычислительным операциям. В настоящее время, программы, ориентированные на контент-анализ используют основанные на классической статистике алгоритмы [33, стр. 735]. Контент-анализ не занимается собственно смыслом, а исключительно частотным распределением смысловых единиц в тексте, или по другому - анализом статических закономерностей частотного распределения смысловых единиц в тексте, и не более того [34, стр. 15].

В процедуре контент-анализа можно выделить несколько этапов [35, стр. 12-13]. Основа контент-анализа – это подсчет встречаемости некоторых компонентов в анализируемом информационном массиве, дополняемый выявлением статистических взаимосвязей и анализом структурных связей между ними, а также снабжением их теми или иными количественными или качественными характеристиками. Таким образом, на первом этапе контент-анализа необходимо выбрать то, что необходимо считать, т. е. определить единицы текста. Этими единицами могут быть слова, абзацы, статьи или квадратные сантиметры площади, которую занимает текст. Следующий этап – составление кодировочной инструкции и кодирование, т. е. трансформация и агрегация исходных данных в категории, которые позволяют точно описать характеристики текста, релевантные для исследования. На этом этапе исследователь подсчитывает количество появлений слова в тексте или решает, относится ли текст к определённой категории (например, определяет наличие в его содержании эротики). Контент-анализ заканчивается статистической обработкой полученных количественных данных (обычно используются процентные и частотные распределения, разнообразные коэффициенты корреляций) и их интерпретацией.

С другой стороны, для text mining используются методы анализа основанные главным образом на байесовском подходе к статистике. Цель интеллектуального анализа текста состоит не в подсчёте частоты некоторых выделенных единиц, а в получении нового, ранее неизвестного знания. Его результатом может быть модель, которая автоматически будет распределять документы по заданным категориям или объединит документы с похожим содержанием в кластеры.

Рассматриваемые методы различаются также в источниках входных данных [33, стр. 735]. Исторически контент-анализ применялся главным образом для анализа социологических, политологических или психологических данных, в то время как с помощью text-mining сейчас анализируют любые текстовые данные (см. Области применения). Однако, что касается типа данных, то контент-анализ является более универсальным методом, поскольку используется для анализа документов самого разнообразного типа – это могут быть визуальные изображения, устная речь, невербальное поведение и т. д.⁵ Text mining же по определению является сферой data mining, ограниченной анализом текстовых данных.

Между этими методами существует ещё одно различие, которое заключается в интенсивности использования компьютеров. Контент-анализ возник ещё задолго до изобретения первых ЭВМ и до сих пор предполагает ограниченное использования компьютера: если

⁵ Например, во время второй мировой войны произошёл один из самых известных случаев применения контент-анализа. На основе анализа нацистской пропаганды на радио, сотрудники ВВС предсказали ракетную атаку на Британские острова.

подсчёт слов легко можно автоматизировать, то процедура кодирования требует непосредственного участия исследователя. Text mining же с самого начала развивался вместе с развитием электронно-вычислительной техники. К тому же его связь с наукой искусственного интеллекта отражается в том, что после программирования модели вмешательство человека часто почти не требуется. Впрочем, это различие постепенно сходит на нет, поскольку сейчас появляются компьютерные системы автоматического контент-анализа, а для успешного интеллектуального анализа текста необходимо пристальное внимание исследователя на всех его этапах.

Однако, следует заметить, что изложенная нами позиция по определению терминов контент-анализа и text-mining далеко не единственная. Как говорилось ранее, существует большая путаница по разграничению объёмов этих понятий. Некоторые отечественные исследователи вообще не оставляют места методам text mining в классификации методов анализа текста (<http://pug.hse.ru/2012/02/20/n980>).

Глава 2

Практическая часть. Исследование образа губернатора омской области в местных Интернет-СМИ

2.1 Определение целей исследования

В данной части работы мы разработаем и проведём исследование, на примере которого будут показаны возможности метода интеллектуального анализа текста в социологии. Цель исследования на данном этапе состоит в том, чтобы с помощью интеллектуального анализа текста выявить некоторые характеристики дискурса о мэре г. Омска в местных Интернет-СМИ. Такими характеристиками являются:

1. Распределение статей с упоминанием мэра во времени. Нас будет интересовать, в какие месяцы или дни недели активизируется соответствующий дискурс. В контексте этой задачи мы сравним распределение статей во времени из генеральной совокупности и распределение статей из выборочной совокупности с тем, чтобы определить, значимо ли они различаются. В зависимости от полученных результатов можно выдвинуть гипотезы, объясняющие полученное распределение.
2. Количество комментариев. Для определения заинтересованности читателей в данной теме, сравним количество комментариев к статьям о мэре со средним количеством комментариев.
3. Тема документа, в котором упомянут мэр. С помощью алгоритмов тематического моделирования мы определим тематический контекст статей из выборки, изучим его распределение во времени и сравним его темами в генеральной совокупности на предмет наличия общих и особенных тем. Попутно мы оценим эффективность работы алгоритма тематического моделирования.
4. Попытка анализа настроений. Нам будет произведено сравнение эмоций по выборке, где упомянут мэр и остальной, а так же в различных темах.

2.2 Оценка доступности и характера данных. Сбор данных.

Прежде чем начать сбор данных, нам придётся поставить перед собой несколько вопросов, представляющих особенную трудность в данного типа исследованиях. А именно,

необходимо определить, что будет являться носителем знаний по исследуемой проблеме (т. е. эмпирическим объектом исследования), каковы границы генеральной совокупности, какой метод будет являться адекватным для построения выборочной совокупности, как определить качественные и количественные характеристики выборки, каковы критерии репрезентативности выборки [28].

Определение эмпирического объекта. В самом общем виде можно сказать, что источником знаний о проблемах, затронутых в данном исследовании является статьи в Интернет-СМИ Омской области. Углубляясь дальше, мы должны решить какие аспекты статьи нас интересуют. Статья в Интернет-СМИ – не просто один лишь неструктурированный текст. Это документ, который имеет свою структуру. В этой структуре нас будут интересовать такие элементы как собственно текст, название, дата публикации, количество комментариев, сами комментарии, принадлежность к тому или иному СМИ. Первая причина, по которой они были выбраны состоит в представленности этих элементов в статьях каждого из рассматриваемых нами Интернет-СМИ. Количество просмотров и ключевые слова (тэги), например, на некоторых ресурсах бывают не указаны. Другая причина – достаточность данных элементов для решения исследовательских задач.

Определение генеральной совокупности. Генеральную совокупность в данном исследовании составляют статьи Интернет-СМИ г. Омска. Интернет-СМИ – веб-сайт, ставящий своей задачей выполнять функцию средства массовой информации (СМИ) в сети Интернет. По данным Агентства Региональных Исследований за июнь 2014 года в Омске работает около 18 Интернет-СМИ с месячным количеством уникальных посетителей в месяц более 10000 [36].

Определение выборочной совокупности. Использование данных со всех возможных ресурсов – очень трудоёмкая задача, поскольку требует практически полного переписывания соответствующей части программы, ответственной за собственно сбор данных, и частичной переработки модуля предварительной обработки. Вполне привычным для социолога решением будет конструирование выборки. Однако как рассчитать выборку, если не известны объёмы генеральной совокупности. А даже если мы знали количество статей каждого из рассматриваемых Интернет-СМИ за любой промежуток времени, разве было бы корректно использовать традиционные методы определения выборочной совокупности в такого типа исследованиях? Эта аналогия видится некорректной по причине кардинального различия эмпирических объектов – человека и текста. При определении людей в качестве эмпирических объектов исследования социолог как правило предполагает, что они в равной степени могут служить источником информации о проблеме. Исключение из этого правила встречается, когда исследователь отдельно изучает мнение экспертов. Но экспертные опросы – это отдельная часть исследования, в которой как правило используются другие методы сбора и анализа информации.

В нашем случае, исходя из цели исследования – определение образа мэра, транслируемого Интернет-СМИ в обществе, – важна не сама статья, а влияние, оказываемое ей на общество. Именно это влияние и определяет степень, с которой статья может служить источником информации о проблеме. Его прямым индикатором служит количество просмотров данной статьи. Но не все Интернет-СМИ предоставляют эту информацию, поэтому можно опереться на косвенный индикатор – количество просмотров всех статей исследуемого ресурса. То есть мы предполагаем, что чем больше совокупное количество просмотров у одного ресурса, тем более сильно влияние каждой его отдельной статьи. Для верности этого тезиса необходимо только чтобы общее количество статей в каждом из ресурсов не сильно отличалось друг от друга. привести данные

Приведённые выше рассуждения позволяют считать, что при выборе Интернет-СМИ, статьи которых будут подвергнуты анализу, стоит опираться на общее количество просмотров. Однако это не даёт ответа на вопрос, сколько и какие статьи должны быть ото-

браны. Способов расчёта этих значений на сегодняшний день нет. Существуют публикации отдельных исследователей из разных отраслей, разрабатывающих свой методологический аспект при исследовании текстов в сети Интернет. Определение выборочной совокупности прежде всего зависит от того, что информация о чём важно для исследования. На основании этого определяет эмпирический объект (это может быть текст, комментарий, отдельное высказывание и др.) и принцип определения его важности (например, степень влияние на общество). В любом случае сохраняет один принцип – из выборки необходимо получить репрезентативное подмножество, – но исследователи пытаются достичь его разными путями.

З. Папачарисси в при исследовании блогосферы определяла в качестве генеральной совокупности все блоги, расположенные на платформе blogger.com. Она объясняет свой выбор тем, что это наиболее популярный и большой по числу блоггеров англоязычный сайт, который предоставляет возможности для персональных публикаций в стиле любительской журналистики. Любой блог, по мнению Папачарисси, размещённый на этом сайте, представлял собой единицу анализа, отвечающую по своим характеристикам признакам принадлежности к генеральной совокупности. Однако нельзя согласиться, что блоги с blogger.com репрезентативны относительно всей блогосферы. Выборочную совокупность блогов Папачарисси составляла используя случайную отправную точку и случайный выборочный интервал. Однако исследователем не было оговорено, каким именно данные вводились в поисковую систему для поиска релевантных блогов и какие именно блоги считались релевантными, сколько блогов входило в генеральную совокупность и почему было отобрано именно 260. К тому же использование поисковых систем для поиска блогов выглядит сомнительно: алгоритмы данных систем неизвестны исследователю и нельзя сказать, почему были отобраны эти блоги, а не иные.

Этот пример исследования был приведён нами, чтобы проиллюстрировать отсутствие единой позиции в способах определения выборочной и генеральной совокупности в Интернет-исследованиях. Каждый исследователь придумывает сам, каким способом наиболее полно реализовать принципы выборки.

В нашем случае необходимо определить несколько Интернет-СМИ, все статьи которых будут отобраны для исследования. Мы выяснили, что при определении значимости, веса статьи определяющей характеристикой является количество просмотров. Хотя Интернет-СМИ в Омске немало, не все из них одинаково крупны. Судя по тем же данным АРИ, в Омск существует всего четыре новостных ресурса, страницы которых просматривают более одного миллиона раз в месяц. В процентном отношении они занимают 65% рынка омских Интернет-СМИ. Представляется, что анализ статей, получивших более половины всех просмотров является достаточным основанием для выделения их в качестве выборочной совокупности, по результатам анализа которой можно будет делать выводы об омских Интернет-СМИ в целом. Таким образом в исследовании будут проанализированы все новостные статьи с сайтов gorod55.ru, bk55.ru, ngs55.ru, omskinform.ru за период с 1 сентября 2013 по 1 сентября 2014. Новостными статьями будут считаться те, которые публикуются на данном ресурсе в разделе «Новости». Статьи из категорий «Работа», «Объявления», «Блоги» и др. в анализе не участвуют.

Определившись с данными, которые необходимо собрать, нужно решить, каким способом это сделать, т. е. с использованием каких инструментов и технологий будет производиться сбор данных. Для этого мы будем использовать язык программирования Python. Основанием для такого выбора является его простота, поддержка многопоточности, что полезно для более быстрого сбора данных, наличие сторонних библиотек, что позволяет избежать написание рутинного кода, а также то, что обработка и анализ данных также будет производиться на этом языке – это обеспечивает некоторую консистентность исследования. Ближайшей альтернативой данному решению видится использование программной

платформы node.js из-за хорошей поддержки асинхронных запросов (и, следовательно, высокой скорости) и наличия множества качественных библиотек для сбора данных или языка R, который традиционно популярен в академической среде для сбора и анализа данных.

Для хранения данных будет использована база данных MongoDB. Как говорилось выше, в БД будут присутствовать следующие поля: название статьи (title), содержимое статьи (content), ссылка на статью (url), дата публикации (date), количество комментариев (commentsCount) и список комментариев к статье (comments). Статьи с каждого источника будут храниться в отдельной коллекции.

Результаты сбора данных следующие:

- С сайта gorod55.ru было собрано 6302 статьи
- Больше всего новостных статей за указанный промежуток времени было опубликовано на bk55.ru – 14078 статей на bk55.ru
- Наименьшее количество статей – 4780 – было найдено на ngs55.ru
- 8727 статей по указанным параметрам было собрано с сайта omskinform.ru

Всего таким образом в анализе участвовало 33887 статей.

На этом этапе крайне важно контролировать корректность и полноту собираемых данных. Сложнее всего было с сайтом bk55.ru, поскольку в нём использовались несколько различных шаблонов для отображения информации, каждый из которых необходимо было отследить и создать под него набор правил для извлечения данных.

2.3 Предварительная обработка данных

Предварительная обработка данных – один из важнейших этапов в анализе текста. Наша цель на этом этапе – удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду.

На самом деле удалять несущественные данные мы начали ещё на этапе сбора данных, поскольку перед записью в базу данных весь текст, если это было необходимо, очищался от html-разметки. Преобразование же данных на том этапе заключалось в конвертации текста, содержащего информацию о дате публикации, в специальный тип данных, позволяющий обращаться к этим данным как к дате, например, производить выборку статей за определённый период.

Следующий шаг в предварительной обработке данных заключается в удалении из каждой статьи признаков, свидетельствующих о её принадлежности к какому-либо источнику. Если посмотреть на полученные тексты, то можно увидеть, что редакция каждого СМИ устанавливает собственные правила оформления документов, касающиеся оформления ссылок на источники данных, фотографий, указание имён авторов. В случае если эти отличительные черты не будут устранены, алгоритмы тематического моделирования, которые мы в дальнейшем собираемся применить к собранному корпусу текстов, будут стремиться образовать темы вокруг источников. Процедура унификации статей из различных источников достаточно трудоёмка и требует ручного анализа множества статей с каждого из них, с тем чтобы выявить в них специфические черты для каждого сайта. Такими чертам могут быть имена журналистов данного издания или правила оформления фото и видео материалов (например, около каждой фотографии может указываться копирайт).

Например, чтобы удалить имена журналистов из текстов статей на сайте bk55.ru, необходимо было во-первых, составить их список. Для составления списка, была написана

небольшая программа, выводящая два последних слова каждого документа, если они начинались с заглавной буквы (как правило имена авторов указывались в конце документа, хоть и не всегда). Из полученного списка примерно в пятьсот пар были вручную отсеяны пары, не являющиеся именем и фамилией. Те пары из этого списка, которые встречались больше двух раз, считались нами именем и фамилией журналистов сайта bk55.ru. На последнем этапе фамилии журналистов удалялись из каждого документа. К тому же, так как после имён журналистов часто указывалась другая мета-информация (главным образом ссылки источники информации), то также удалялся весь текст после имён, если по размеру этот текст не превышал определённое количество символов (чтобы предотвратить удаление не мета-информации).

После устранения специфической информации данные из различных источников объединялись в единый корпус и подвергались дальнейшей обработке. Обработка заключалась в следующем:

1. Перевод текста в нижний регистр
2. Токенизация
3. Удаление пунктуации
4. Стемминг
5. Удаление стоп-слов
6. Замена слов

Что касается перевода текста в нижний регистр и удаления пунктуации, то это достаточно тривиальные процедуры, не требующие особых объяснений.

Другим этапом предварительной обработки текста является токенизация. Именно с неё начинается обработка естественного языка как наука и как конкретная деятельность [37]. Под токенизацией понимают процесс сегментации текста на отдельные части, называемые токенами. Именно токены являются теми первичными элементами, которые непосредственно участвуют в процессе анализа.

Выделяют два основных признака токена – лингвистическая значимость и методологическая полезность [37, стр. 1106]. В языках с иероглифической письменностью токенизация является серьёзной проблемой, поскольку один иероглиф может обозначать как морфемы (в таком случае он не удовлетворяет требованиям для того, чтобы считаться токеном), так и целые слова. В английском и русском языках проблема токенизации не стоит так остро и чаще всего токены определяются через пробелы между словами и знаки препинания. Тем не менее, даже в этих языках существуют определённые нюансы.

Нами было протестировано несколько алгоритмов токенизации (токенайзеры TreebankWordTokenizer, WordPunctTokenizer, PunctWordTokenizer и WhitespaceTokenizer из программы NLTK и токенайзер из Pattern). Корректнее всех выделял токены токенайзер из программы [Pattern](#). Например, он единственный интерпретировал url'ы как цельные токены, не выделяя в них отдельные сегменты, на основе знаков препинания.

После токенизации и удаления токенов, являющихся знаками препинания, мы перешли от представления документов как набора символов к документам как списку слов. Дальнейшие наши шаги будут направлены на уменьшение длины этого списка, т. е. на снижение как общего количества токенов, так и количества их уникальных единиц. Необходимость этих шагов обусловлена желанием снизить вычислительную сложность при анализе данных.

Первый шаг направлен на снижение количества уникальных токенов. Для компьютера различные формы одного и того же слова являются совершенно разными словами.

Существует два способа для приведения словоформ к одной лексеме. Первый, самый простой, называется стемминг. Он состоит в отсечении слово- и формообразующих частей – префиксов, суффиксов, окончаний, в результате чего остаётся основа слова – неизменная часть, выражающая его лексическое значение.

Более сложным подходом к решению проблемы унификации словоформ является лемматизация. Лемматизация – это процесс приведения словоформы к лемме – её нормальной (словарной) форме. В русском языке нормальная форма имени существительного имеет именительный падеж и единственно число, для прилагательных добавляется требование мужского рода, а глаголы, деепричастия и причастия в нормальной форме должны стоять в инфинитиве.

Для постановки слова в нормальную форму необходимо иметь словарь, где для каждого слова определены его характеристики, т. е. часть речи, падеж, число, род, форма глагола (если это глагол). Создание такого словаря требует колоссальных трудов. В отличие от этого, стемминг предполагает наличие лишь списка приставок, суффиксов и окончаний, количество которых исчисляется несколькими десятками. К счастью, для русского языка существует так необходимый для лемматизации словарь, созданный в рамках проекта [OpenCorpora](#). Используя этот словарь программа [pymorphy2](#) позволяет приводить слова к нормальной форме.

Между вышеозначенными способами мы выбрали лемматизацию, поскольку получаемые в результате этого процесса леммы легче интерпретировать, чем усечённые основы слов, значение которых не всегда легко восстановить.

Дальнейшие усилия по уменьшению количества токенов связаны с удалением так называемых стоп-слов. Эти слова сами по себе почти не неся полезного смысла, тем не менее, необходимы для нормального восприятия текста. Чаще всего к разряду стоп-слов относятся служебные части речи – предлоги, союзы, частицы. Будучи широко распространёнными в тексте, они мало могут сказать о его теме.

В качестве базы для списка стоп-слов был использован список русских стоп-слов из программы NLTK. Однако его нельзя считать достаточно полным. Включая в себя 151 слово данный список покрывает лишь самые основные случаи. Для его пополнения необходимо обратиться к собранным ранее данным. На их основе был составлен список наиболее часто встречающихся в корпусе токенов. Среди них были выбраны несколько десятков слов, наиболее точно подходящие под описание стоп-слов (это, который, такой, некоторый, другой, тот и др.), которые затем были добавлены в соответствующий список. Представляется, что такой список, дополненный словами, выбранными из числа наиболее распространённых, является достаточно полным, поскольку стоп-слова по своему характеру всегда относятся к наиболее часто встречающимся в тексте. Редкие слова как правило свидетельствуют о принадлежности текста к какой-либо теме, а потому не могут относиться к разряду стоп-слов.

В заключения для удобства анализа была произведена замена некоторых слов. Данная замена включала в себя во-первых, раскрытие аббревиатур (рф → россия, ул → улица и др.), а во-вторых, лемматизацию токенов, которые не были лемматизированы автоматически (расина → расин, парка → парк). Данный шаг не является обязательным и может быть без последствий пропущен.

Как видно, в общих чертах данный набор процедур повторяет составляющие предварительной обработки данных из методологии CRISP-DM.

Необходимо отметить, что после каждой операции с данными на этапе предварительной обработки следует контролировать последствия производимых изменений. Такой контроль поможет выявить проблемы на раннем этапе, что убережёт от лишней работы в

будущем ¹. Легче всего производить контроль через анализ изменений в списке наиболее часто встречающихся слов.

2.4 Анализ данных

¹Например, одной из таких проблем, выявленных на раннем этапе, было наличие в текстах некоторых СМИ неразрывных пробелов. Они мешали токенизации, поскольку сегментация производилась по обычным пробелам. Решением стала замена всех неразрывных пробелов на обычные.

Литература

1. Дюк В. А., Флегонтов А. В., Фомина И. К. Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // Известия Российского государственного педагогического университета им. А.И. Герцена. 2011. № 138. С. 77–84.
2. Давыдов А. А. Knowledge Discovery and Data Mining в системной социологии. 2013. URL: http://www.isras.ru/Davydov_Knowledge.html.
3. Schrodtt Philip A. Seven Deadly Sins of Contemporary Quantitative Political Analysis // APSA 2010 Annual Meeting Paper. 2010. URL: <http://eventdata.psu.edu/7DS/Schrodtt.7Sins.APSA10.pdf>.
4. Nisbet Robert, Elder John, Miner Gary. Handbook of statistical analysis and data mining applications. Academic Press, 2009. с. 864.
5. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians / R. Christensen, W. Johnson, A. Branscum [и др.]. 1 изд. CRC Press, 2010. с. 516.
6. Зельнер А. Байесовские методы в эконометрии. Москва: Статистика, 1980.
7. Гнеденко Б. В. Курс теории вероятностей. 8 изд. Москва: Едиториал УРСС, 2005.
8. Efron Bradley. Modern Science and the Bayesian-Frequentist Controversy. 2005. URL: <http://www-stat.stanford.edu/~ckirby/brad/papers/2005NEWModernScience.pdf>.
9. Talbott William. Bayesian Epistemology // The Stanford Encyclopedia of Philosophy / под ред. Edward N. Zalta. 2013.
10. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики. 2-е, исправленное изд. Москва: Юнити-Дана, 2001. Т. 1. с. 656. URL: <http://ecsocman.hse.ru/text/33442857>.
11. Айвазян С. А. Байесовский подход в эконометрическом анализе // Прикладная эконометрика. 2008. № 1(9). С. 93–130. URL: http://pe.cemi.rssi.ru/pe_2008_1_93-130.pdf.
12. Jeffreys Harold. Theory of Probability. 3 изд. Oxford: Clarendon Press, 1983.
13. Wilhelm Adalbert. Handbook of Computational Statistics: Concepts and Methods / под ред. J. E. Gentle, W. Härdle, Y. Mori. Springer, 2004. С. 789–803.
14. Han Jiawei, Kamber Micheline. Data Mining: Concepts and Techniques / под ред. Jim Gray. 2 изд. Elsevier, 2006.
15. Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод [и др.]. 3 изд. БХВ-Петербург, 2009.

16. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications / G. Miner, D. Delen, J. Elder [и др.]. Elsevier, 2012.
17. C. Tetlock Paul. Giving content to investor sentiment: The role of media in the stock market // The Journal of Finance. 2007. June. T. 62, № 3. C. 1139–1168. URL: http://www0.gsb.columbia.edu/faculty/ptetlock/papers/Tetlock_JF_07_Giving_Content_to_Investor_Sentiment.pdf.
18. Archak Nikolay, Ghose Anindya, Ipeirotis Panagiotis. Deriving the pricing power of product features by mining consumer reviews // Management Science. 2011. August. T. 57, № 8. C. 1485–1509. URL: http://pages.stern.nyu.edu/~aghoste/pricingpower_print.pdf.
19. Askitas Nikolaos, Zimmermann Klaus F. Google econometrics and unemployment forecasting // Applied Economics Quarterly. 2009. April. T. 55, № 2. C. 107–120. URL: <http://ftp.iza.org/dp4201.pdf>.
20. Tausczik Yla R., Pennebaker James W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods // Journal of Language and Social Psychology. 2010. T. 29, № 1. C. 24–54. URL: <http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/Tausczik&Pennebaker2010.pdf>.
21. Golder Scott A., Macy Michael W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures // Science. 2011. September. T. 333. C. 1878–1881. URL: <http://www3.ntu.edu.sg/home/linqiu/teaching/psychoinformatics/DiurnalandSeasonalMoodVaryAcrossDiverseCultures.pdf>.
22. Mosteller F., Wallace D.L., Nerbonne J. Inference and Disputed Authorship: The Federalist. The David Hume Series. Center for the Study of Language and Information, 2008. URL: <http://books.google.ru/books?id=g7wbAQAAMAAJ>.
23. Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopedie / Russell Horton, Robert Morrissey, Mark Olsen [и др.] // Digital Humanities Quarterly. 2009. T. 3, № 2. URL: <http://www.digitalhumanities.org/dhq/vol/3/2/000044/000044.html>.
24. A latent variable model for geographic lexical variation / Jacob Eisenstein, Brendan O'Connor, Noah A. Smith [и др.] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. C. 1277–1287. URL: <http://www.cs.cmu.edu/~nasmith/papers/eisenstein+oconnor+smith+xing.emnlp10.pdf>.
25. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series / Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge [и др.] // International AAAI Conference on Weblogs and Social Media. Washington: 2010. URL: <http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf>.
26. Media coverage in times of political crisis: a text mining approach / E. Junqué de Fortuny, T. De Smedt, D. Martens [и др.] // Expert Systems with Applications. 2012. Октябрь. T. 39.
27. Causey Charles. The Battle for Bystanders: Information, Meaning Contests, and Collective Action in the Egyptian Uprising of 2011. 2012. URL: http://www.soc.washington.edu/sites/default/files/SCOPES%20Causey_Oct_12pdf.pdf.

28. Кольцова О. Ю., Павлова Ю. К методологии сбора Интернет-данных для социологического анализа. СПб, 2011. URL: <http://www.hse.ru/data/2013/06/10/1283698963/JulijaPavlova,OlesjaKolcova,Kmetodol..dljasociologicheskogoanaliza.pdf>.
29. Иудин А.А., Рюмин А.М. Контент-анализ текстов: компьютерные технологии: Учебное пособие. Н.Новгород: Нижегородский государственный университет им. Н.И.Лобачевского, 2010. URL: http://window.edu.ru/resource/004/74004/files/Ctt_3.pdf.
30. Ландэ Д.В. Основы интеграции информационных потоков. Киев: Инжиниринг, 2006.
31. Smith C. P. Content analysis and narrative analysis // Handbook of research methods in social and personality psychology / под ред. Н. Т. Reis, С. М. Judd. Cambridge, UK: Cambridge University Press, 2000. С. 313–335.
32. Почепцов Г.Г. Теория и практика коммуникации (от речей президентов до переговоров с террористами). Москва: Центр, 1998.
33. Ho Yu Chong, Jannasch-Pennell Angel, DiGangi Samuel. Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability // The Qualitative Report. Т. 16, № 3. С. 730–744. URL: <http://www.nova.edu/ssss/QR/QR16-3/yu.pdf>.
34. Аверьянов Л.Я. Контент-анализ. Москва: РГИУ, 2007.
35. Морозова В. Н. Методы политического анализа: Учебно-методическое пособие. Воронеж, 2007.
36. Рейтинг АРИ омских интернет-СМИ. Сводка. Доступ: 2014-08-25. URL: <http://omsk-journal.ru/publ/9-1-0-116>.
37. Webster Jonathan J., Kit Chunyu. Tokenization As the Initial Phase in NLP // Proceedings of the 14th Conference on Computational Linguistics - Volume 4. COLING '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. С. 1106–1110. URL: <http://dx.doi.org/10.3115/992424.992434>.