Моделирование динамики процессов на основе анализа последовательности текстовых выборок

А.А. Харламов, Т.В. Ермоленко, А.А. Жонин

Введение

В сложных условиях российской действительности, характеризующейся кризисными явлениями во многих сферах жизни общества, общество проявляет повышенный интерес к стрессовым состояниям людей. В наши дни проблемы управления стрессами на работе становятся наиболее актуальными, потому что быстро изменяются социально-экономические и политические ситуации, увеличиваются нервно-психические и информационные нагрузки, происходит диверсификация производства, постоянно растет конкуренция и обостряется борьба за рынки сбыта [1, 2]. Профессии, связанные с высокой психической напряженностью, получают все более широкое распространение. Все это приводит к тому, что в современных условиях человек все чаще оказывается под влиянием биологических и психологических стрессогенных факторов, вызывающих в организме специфические и неспецифические логические, психологические и поведенческие реакции. Как и любой другой, социальный стресс представляет собой явление, имеющее множество биохимических, физиологических, психологических и социально-психологических проявлений. Последнее выражается в изменении общения при стрессе и прежде всего – его эмоциональной окраски. Очевидно, что основным источником информации для решения задач психологического анализа с целью выявления стресса и определения его динамики, являются социальные сети.

Социальные сети уже не только средство для ведения личных или корпоративных блогов, не просто платформа для общения с друзьями или виртуальными "френдами", а способ быстро обмениваться информацией, связанной с событиями общественной значимости. Из социальной сети

можно узнать о человеке довольно много – семейное и социальное положение, интересы, профессии родственников и друзей.

Все более универсальным источником информации становится Twitter, куда агентства и сетевые издания выставляют ссылки на самые последние новости и статьи. Twitter является платформой, откуда мы узнаем, например, о массовых задержаниях в ходе уличных акций, о ходе судебных прений или о выходе в свет новой книги. Иметь свои странички на Twitter считается сейчас необходимым для крупнейших политиков, кинозвезд, знаменитостей в самых разных областях. Даже у Папы римского есть несколько страничек – на разных языках. Преимуществом Twitter является его быстрота и оперативность: сообщение объемом не более 140 знаков пишутся быстро и загружаются в интернет при помощи мобильных телефонов. Не менее важны и интересны и другие социальные сети. Это, в первую очередь, Facebook, и Google+, а также популярные среди русскоязычных пользователей ВКонтакте и Живой журнал, он же Livejournal.

Огромный объем личной информации, которым делятся пользователи, отслеживается не только спецслужбами, друзьями или случайными знакомыми, за потенциальными или действующими клиентами давно уже следят банки и микрофинансовые организации, которые стремятся оценить кредитоспособность заемщика. Целью такого отслеживания является поиск и анализ новых сообщений по заданной тематике (например, касающейся указанного бренда или события) [3].

Другая цель веб-мониторинга – выявление быстро распространяющейся информации, которая копируется из одного сообщения в другое. Как правило, от программного обеспечения требуется определять сами факты быстрого распространения информации, находить основные пути ее распространения и первоначальные источники. Наконец, к задачам веб-мониторинга относится определение источников информации, имеющих существенное влияние на пользователей сети. Примерами таких источников являются блогеры, количество читателей постов которых сравнимо с

аудиторией СМИ. Каждый такой источник информации характеризуется двумя параметрами: охватом (как много пользователей читают сообщения данного источника) и степенью отклика (как много пользователей реагируют на сообщения данного источника). Цели определения влиятельных источников могут быть самыми разными, начиная от проведения маркетинговых акций и заканчивая мероприятиями, связанными с обеспечением безопасности.

Тексты блогосферы – посты и комментарии – представляют собой сферу, где ожидается выражение субъективной оценки автором того или другого явления, события, определенной группы лиц, или конкретной личности, выражение эмоций. Располагая инструментарием для автоматического определения эмоциональной и оценочной окраски текста, можно обследовать выборки текстов блогосферы значительного объема. Зная тематическую принадлежность или другие характеристики исследуемых текстов, можно определять, какие сегменты блогосферы связаны с выражением положительных или отрицательных оценок и эмоций. Таким образом, анализируя эмоциональную окраску последовательности (вчера, сегодня, завтра) текстовых выборок из социальных сетей, в результате можно получить динамику развития социально-психологического субсиндрома (улучшение, ухудшение, стабильно), формируемого при стрессе.

Выявление в документе эмоционально окрашенной лексики и эмоциональной оценки объектов автором является основной задачей анализа тональности текста или Sentiment analysis (SA) – развивающегося направления компьютерной лингвистики. Эмоциональная оценка, выраженная в тексте, также называется тональностью, или сентиментом текста 4[2]. Лексическая тональность (или лексический сентимент) – эмоциональная составляющая, выраженная на уровне лексемы, эмоциональная окраска текста определяется тональностью его составляющих, а также их взаимосвязями [4].

В анализе тональности текста считается, что текстовая информация в сети Интернет делится на два класса: факты и мнения [5]. Ключевым понятием является определение мнения. Согласно определению, данному в [5]: анализ эмоциональной окраски текста – задача автоматического анализа мнений и эмоционально окрашенной лексики, выраженных в тексте.

Эмоциональное высказывание, именуемое в дальнейшем «мнение», представляет собой кортеж из четырех элементов (entity, sentiment value, holder, time). Где entity – объект, о котором автор (holder) высказал мнение в момент времени (time). Выделяются 3 класса эмоциональной окраски (sentiment_value): позитивная, негативная и нейтральная. Под нейтральной подразумевается, что текст не содержит эмоциональной составляющей.

В общем случае классификация эмоциональной окраски может быть не только тернарной, но и бинарной (положительный / отрицательный) или ранжированной [3] и относится к задаче определения полярности текста.

Для определения полярности текста прежде всего в нем выделяются эмотивные (субъективные) мнения. В анализе тональности текста часто встречается термин, связанный с понятием мнения - субъективность. В [5] дается определение объективного и субъективного предложений:

Объективное предложение выражает фактическую информацию о чем-либо, тогда как субъективное предложение выражает чьи-то личные чувства и предположения. Предложения, содержащие мнение, обычно являются субъективными, поэтому анализ текста на наличие субъективной информации часто является подзадачей определения полярности текста.

Итак, анализ тональности текста включает в себя следующие основные задачи: 1. Определение полярности текста. 2. Извлечение объектов из эмоционально окрашенного текста.

Методы определения полярности текста. Определение полярности предложения обычно происходит в два этапа. Сначала проводится анализ предложения на субъективность. Если предложение содержит информацию

субъективного характера, то, скорее всего, в нем выражено мнение. Далее у субъективного предложения определяется полярность. В противном случае предложение содержит фактическую информацию и далее не рассматривается.

Для отделения текстов, содержащих суждения, от документов, преимущественно описывающих факты, а также для определения полярности текста используются стандартные методы классификации. Традиционно анализ эмоциональной окраски текста осуществляется при помощи методов машинного обучения с учителем: наивный байесовский классификатор и машина опорных векторов [6].

Классами являются оценки эмоциональной окраски, а признаки извлекаются из текста. Очевидно, что основным источником информации для автоматического определения эмоциональной окраски в тексте является, прежде всего, лексика (слова и сочетания, выражающие эмоцию); также может учитываться пунктуация (например, восклицательные знаки, особенно несколько подряд) и специальные конвенции, свойственные данному типу текстов (например, эмотиконы для интернет-коммуникации). Поэтому в качестве признаков могут использоваться грамматические классы, например, части речи, структурные особенности, а также знаки препинания.

Например, составляется словарь прилагательных и наречий, как терминов, выражающих эмоцию. Кроме непосредственно слов, выражающих эмоции, существуют словосочетания, которые также содержат эмоциональную оценку. Поэтому часто в качестве признаков выбирают n-граммы слов и их грамматические характеристики. Для выявления таких случаев последовательно анализируются отдельные слова, биграммы, триграммы и т.д., оценивается их "точность". Точность n-словной цепочки – это число субъективных выражений этой цепочки, поделенное на общее число употребления этой цепочки. Употребление n-словной (n – количество слов) цепочки передает субъективность, если каждое слово этой фразы попадает в субъективный элемент. После того, как была получена оценка

точности для отдельных слов и сочетаний, используется выделение эмоционально окрашенных сочетаний: отсекаются все словосочетания, точность которых ниже установленного порога 0,1. Затем отсеиваются сочетания с точностью ниже максимальной точности слов, входящих в эти сочетания.

Существуют методы, основанные на использовании словаря эмоционально окрашенных слов [7, 8] и словаря символов, обозначающих эмоции [8]. В словарных методах каждое слово обладает весом, характеризующим его эмоциональную окрашенность. Часто словари составляются с помощью сторонних инструментов, таких как WordNet, затем термины словаря взвешиваются, например, вручную [9].

В исследованиях, основанных на словарных методах, полярность текста определяется различными способами. В [7] текст обладает тремя независимыми оценками: позитивной, негативной и субъективной. Полярность документа складывается из полярностей его предложений. Определение полярности предложения заключается в вычислении нормализованных сумм весов терминов в тексте: каждый термин имеет положительный, отрицательный и субъективный веса по шкале целых чисел от 0 до 1.

В [8] короткие текстовые сообщения независимо оцениваются по негативной и позитивной шкале: от -5 до -1 и от 1 до 5 соответственно. Каждый термин словаря имеет два веса, определенные этим способом. Положительный вес текстового сообщения равен весу термина с максимальной положительной оценкой. Если текст не содержит терминов входящих в словарь, ему присваивается минимальный положительный вес. Аналогичным образом определяется отрицательный вес сообщения.

На основе словарных методов был разработан программный продукт SentiStrength для определения тональности текста [10 - 12]. ПО SentiStrength работает практически исключительно с отдельными словами. Словарь – это набор слов-маркеров, на присутствие которых в тексте реагирует

SentiStrength. Если у слова из словаря значится положительная оценка, то и текст получит положительную оценку и наоборот. Контекст практически не влияет на оценку. Надежность применения способа зависит от надежности используемого словаря: поскобку по отдельным словам осуществляется оценка эмоциональной окраски текста.

Таким образом, сильными сторонами подобного автоматического метода исследования текстов является его способность работать с большими массивом данных и выдавать быстрый результат, по которому можно судить о наличии или отсутствии эмоций в тексте. Слабой же стороной является сложность учета всех нюансов при составлении словаря, с помощью которого работает программа и необходимость привлечения человеческих ресурсов для постоянного совершенствования словаря и проверки полученных данных.

Методы выявления значимых объектов в эмоционально окрашенном тексте. Вышеописанные задачи затрагивают проблему определения полярности мнения, но не проблему идентификации объектов (иначе говоря, аспектов, согласно приведенному определению мнения), о которых оно высказано.

На сегодняшний день данной задаче извлечения объектов из эмоционально окрашенного текста посвящено меньшее число работ, чем другим задачам, входящим в анализ тональности текста.

Задачу извлечения объектов можно рассматривать, как задачу извлечения терминов, часто употребляемых авторами мнений [13]. При ее решении популярны методы обучения без учителя и статистические методы. Для этих методов не требуется размеченных обучающих выборок, которые на сегодняшний день отсутствуют в свободном доступе.

Обучение без учителя – раздел машинного обучения, в котором изучаются задачи выявления скрытых закономерностей и взаимосвязей между объектами из неразмеченной выборки данных. Обучению без учителя

можно противопоставить методы обучения с учителем, когда для каждого объекта из обучающей выборки задан правильный ответ, и требуется найти зависимость между ответами и объектами.

Рассмотрим один алгоритм из класса методов обучения без учителя в применении к задаче извлечения объектов, о которых высказывалось мнение. В основе методов распространения (bootstrapping) лежит идея итеративного автоматического извлечения похожие текстовых единиц (терминов) с помощью небольшого множества вручную определенных примеров (seed examples) некоторого класса, с постепенным наращиванием этого множества.

В исследовании [13] кандидатами в термины могут быть только n-граммы размером от 1 до 3 слов, содержащие только существительные, прилагательные, глаголы и наречия.

Иногда предполагают, что терминами, описывающими аспекты, могут быть одиночные существительные и словосочетания содержащие существительное, часто встречающиеся во мнениях об объектах одного и того же типа. Из всех n-грамм слов, удовлетворяющих этому требованию, выделяются те, чья частота в корпусе более одного процента.

Выделенные n-граммы, состоящие из двух и более слов, проходят проверку на компактность. Если n-грамма компактна как минимум в двух предложениях, то она попадает в список терминов.

Компактность определяется следующим образом. Пусть f – n-грамма из n слов, s – предложение, содержащее все слова из f (возможно расположенные не подряд). Если расстояние между любыми двумя словами, смежными в f, в предложении s составляет не более чем три слова, то f компактна в данном конкретном предложении.

Термины, состоящие из одного слова, также проходят статистический тест на чистоту. Отыскиваются все предложения, содержащие термин. Среди найденных предложений подсчитываются предложения, не содержащих прошедшие тест на компактность n-граммы, в которые входит этот термин.

Если число таких предложений выше некоторого экспериментально определенного порога, то термин попадает в список аспектов.

Для всех n-грамм, содержащих в себе только определенные части речи, входящих в некоторое множество документов, вычисляется их значимость, которая напрямую зависит от частоты встречаемости терминов.

К статистическим методам выделения терминов относятся способы анализа текста, реализованные на основе нейросетевых алгоритмов. Одним из хорошо зарекомендовавших себя методов статистического анализа, реализующего глобальный анализ текстов является метод на основе формализма искусственных нейронных сетей из нейроподобных элементов с временной суммацией сигналов, используемого для формирования статистического портрета текста; и формализма искусственных нейронных сетей Хопфилда, используемого для смысловой перенормировки весов ключевых понятий в тексте [14 - 16]. Этот подход реализован в системе TextAnalyst, обладает достаточным быстродействием и не зависит от языка и предметной области.

В результате анализа текста из него автоматически извлекается индекс в виде сети основных понятий и их связей с весовыми характеристиками. В качестве смыслового портрета текста рассматривается не просто список ключевых слов, а сеть понятий – множество ключевых слов или устойчивых словосочетаний связанных между собой. Каждое понятие получает некоторый вес, отражающий значимость этого понятия в тексте. Связь между понятиями тоже имеет вес. Использование связей позволяет более точно взвешивать понятия текста.

Для решения задачи определения полярности предложений и коротких сообщений эффективны как алгоритмы обучения с учителем, так и методы, основанные на словарях. Проблемой обучения с учителем является составление тренировочного корпуса с примерами из предметной области, в которой используется классификатор. Однако схожей проблемой обладают и

словарные методы: веса терминов словаря, составленного для одной предметной области, могут оказаться неадекватными для другой.

Задача извлечения объектов высказывания часто решается с помощью методов обучения без учителя и статистическими методами. Для увеличения эффективности этих методов используются лингвистические и частотные фильтры, позволяющие отсеивать слова, не имеющие отношения к аспектам.

Структурный портрет текста в виде семантической сети может быть получен на основе однородной нейросетевой обработки информации [17]. Формализм искусственных нейронных сетей на основе нейроподобных элементов с временной суммацией сигналов позволяет описать процесс структурного анализа информации и дает подход к формированию глобального семантического представления текста. Формализм сетей Хопфилда, в свою очередь, позволяет ранжировать, и, потому, выявлять ключевые фрагменты информации (ключевые понятия – в текстах).

В результате выполнения аналитического обзора по данной теме для моделирования социального процесса наиболее перспективным видится анализ последовательности (вчера, сегодня, завтра) текстовых выборок из социальных сетей с формированием их семантических портретов, в результате которого формируется последовательность семантических сетей, описывающих последовательные во времени выборки текстов. Из каждой сети фильтруется только та ее часть, которая содержит лексические и психолингвистические метки, характеризующие моделируемый процесс. Эти лексические и психолингвистические метки имеют свои смысловые веса в сети. А в последовательности сетей возникает динамика смысловых значимостей соответствующих понятий, что и моделирует социальный процесс (улучшение, ухудшение, стабильно).

1. Теоретические основы моделирования динамики процессов на основе анализа последовательности текстовых выборок

«Социальный стресс - социальное напряжение, требующее многообразных приспособительных реакций, сложного уравновешивания в системах социального поведения, взаимодействия и т.д.» [18].

«Социальные процессы играют большую роль в жизни общества, привнося в него как позитивные, так и негативные результаты» [19]. В основе их возникновения лежат противоречия, возникающие между различными социальными группами, имеющими особые корпоративные интересы, входящими в несоответствие с интересами других групп. Такое положение является естественным и позволяет обществу находить наиболее эффективный путь развития, способный консолидировать интересы большинства своих членов. Вследствие этого возникающие в обществе проблемы вызывают изменения, от которых одни категории людей получают пользу, а другие – терпят ущерб. Сами люди, являясь участниками социальных процессов, не всегда в состоянии оказать на них влияние. Причина этого в том, что вызывая изменения в обществе, люди утрачивают контроль над ними в силу неготовности или неспособности понять внутренне механизмы этих изменений. Наблюдая за происходящими в обществе изменениями, давая им оценку, не всегда можно точно предсказать последствия, к которым могут привести эти изменения. Увеличение способности общества оценивать и контролировать ход своих изменений становится составным элементом социальной культуры и является важным условием его устойчивости.

Процесс – любой вид движения, модификации, трансформации, чередования или «эволюции», то есть любое изменение изучаемого объекта в течение определенного времени, будь то изменение его места в пространстве либо модификация его количественных характеристик.

Социально-экономические и политические процессы – это изменения в обществе, отражающиеся на его благосостоянии, политической и экономической стабильности, условиях безопасности. Это социально значимые изменения в обществе, вызванные стремлением различных групп

влиять на сложившиеся в социуме условия с целью удовлетворения определенного интереса.

Процесс имеет выраженную временную составляющую, позволяющую рассматривать все свойства процесса в зависимости от времени.

Процесс характеризуется масштабом, **направленностью, интенсивностью**, составом и характером стимуляции.

Масштаб процесса – это степень вовлеченности в него субъектов. Охват вовлеченных в процесс индивидов или отдельных социальных групп означает микроуровень, а если в качестве субъекта процесса выступает государства, народы, этносы, культуры – это макроуровень .

Направленность процесса характеризуется его вектором, выражающим ориентацию процесса на определенный исход.

Интенсивность процесса задается осознанным значением его результата для вовлеченных в него участников. Фактически это значение может быть задано через освещение этого процесса в СМИ, осознанием глобальности его последствий для социального субъекта» [19].

Предлагаемый в работе количественный анализ направленности процесса социального стресса основывается на двух основных подходах.

1. Для анализа процесса используется автоматическое формирование семантической сети текста (корпуса текстов), содержащего информацию, касающуюся процесса. При этом автоматически извлекаемые из текста, среди прочих ключевых понятий, лексические и психолингвистические метки ранжируются в зависимости от их значимости в тексте. Суммарный ранг этих меток определяет состояние процесса (хорошо-нейтрально-плохо).
2. Рассматриваются последовательные срезы текстовых корпусов, с построением их семантических сетей (и ранжированием меток), что позволяет выявлять динамику исследуемого процесса как изменение

суммарного ранга меток, характеризующих процесс, от среза к срезу.

Рассмотрим эти два подхода более подробно.

1.1 Формирование семантической сети текста

Научно-производственным инновационным центром «Микросистемы», г. Москва была разработана технология для автоматического смыслового анализа текстовой информации TextAnalyst [17]. Разработанная технология обработки текстовой информации основана на использовании структурных свойств языка и текста, которые могут быть выявлены с помощью статистического анализа, включающего в свой состав некоторые элементы лингвистических представлений. Эта технология позволяет на основе анализа статистики слов и их связей в тексте, реконструировать внутреннюю структуру текста, и таким образом, реализовать автоматическое формирование описания семантики предметной области текста.

Статистический анализ выявляет ключевые понятия текста - слова или устойчивые словосочетания с их частотой встречаемости в тексте. Важной особенностью используемого подхода, является возможность автоматически устанавливать взаимосвязи между выявленными ключевыми понятиями текста. При выявлении связей учитывается статистика попарного появления слов в смысловых фрагментах исследуемого материала. Далее статистические показатели ключевых слов пересчитываются в семантические веса, при этом учитываются подобные характеристики ключевых понятий с ними связанных, а также учитываются численные показатели связей.

После пересчета статистических характеристик текста в семантические, ключевые понятия, которые не релевантны структуре текста, получают малый вес, а наиболее представительные наделяются высоким рангом. Полученная семантическая сеть позволяет производить различные виды анализа текстовой информации. Сеть отражает внутреннюю структуру текста, значимость выделенных ключевых понятий, а также, показывает

степень связанности понятий в тексте. Такое представление текста получается полностью автоматически.

Технология реализует следующую обработку информации.

1. Сегментацию слов и предложений текста на основе графематических правил.
2. Нормализацию грамматических форм слов и вариаций словосочетаний. Выявление корневых основ ключевых понятий.
3. Фильтрацию в тексте семантически несущественных, вспомогательных слов: удаляются предлоги, числительные и самые общеупотребимые слова с широким значением.
4. Выделение ключевых понятий текста (слов и словосочетаний), и их взаимосвязей в тексте.
5. Вычисление их относительной (в тексте) значимости – рангов ключевых понятий.
6. Формирование представления семантики текста в форме семантической сети.

До собственно статистического (с элементами лингвистики) анализа текста осуществляется его первичная обработка. Задачей первичной обработки текста является подготовка его к статистическому анализу. Подготовка текста заключается в очистке его от нетекстовых символов, а также в корректной обработке таких единиц текста как аббревиатуры, инициалы, заголовки, адреса, номера, даты, указатели времени.

Сегментация предложений позволяет разбить текст на участки, которые могут содержать терминологические словосочетания предметной области и избежать выделения неадекватных словосочетаний на стыках таких участков. В результате предобработки (с использованием морфологического анализа) близкие по форме слова и словосочетания приводятся к одинаковой форме (нормализуются).

Ключевые понятия предметной области (слова и словосочетания) выделяются с использованием частотного анализа текста. В процессе

формирования частотного портрета текста подсчитывается частота встречаемости слов в тексте.

Сформированное таким образом представление лексики текста подвергается затем пороговому преобразованию по частоте встречаемости. Порог отражает степень детальности описания текста. В процессе статистического анализа выделяются устойчивые термины и терминологические словосочетания, которые служат далее в качестве элементов для построения семантической сети. При этом в составе частотного портрета текста общеупотребительные слова, а также словосочетания, содержащие только общеупотребительные слова, опускаются.

Первичная (частотная) сеть формируется из выявленных на предыдущем этапе ключевых понятий за счет использования ассоциативных связей (попарной встречаемости) этих слов в смысловых фрагментах текста. В качестве критерия для определения наличия ассоциативной связи между парой понятий используется частота их совместной встречаемости в одном смысловом фрагменте текста (например, в предложении). Превышение частотой попарной встречаемости ключевых понятий некоторого порога позволяет говорить о наличии между понятиями ассоциативной (семантической) связи, а совместные вхождения понятий в предложения с частотой меньше порога считаются просто случайными.

Элементы полученного таким образом частотного портрета текста (однородной семантической – ассоциативной – сети) и их связи имеют числовые характеристики, отражающие их относительный вес в данном тексте, соответствующий частоте их встречаемости в тексте.

При достаточно большом объеме текста значения частот встречаемости ключевых понятий отражают соответствующие семантические (субъективно оцениваемые) веса этих понятий в тексте. Однако, для небольших корпусов текстов, в частности, при анализе отдельного текста, не все частотные характеристики соответствуют действительным семантическим весам -

важности понятий в тексте. Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого “семантического сгущения”. В результате итеративной процедуры перенормировки наибольшие веса получают ключевые понятия, связанные с наибольшим числом других понятий с большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста. Полученные таким образом смысловые веса ключевых понятий показывают значимость этих понятий в тексте.

1.2 Выявление динамики процесса

Полученная сеть представляет собой семантический (структурный) портрет текста (корпуса текстов). Если текст, или корпус текстов описывает некоторую структуру (научную разработку, предметную область, социологическую ситуацию), то сформированная таким образом семантическая сеть представляет собой семантический срез этой структуры в момент написания текста.

Семантическая сеть, построенная на тексте, написанном позже, и описывающем ту же структуру, может отличаться от первой, поскольку представляет текст, релевантный состоянию описываемого процесса на момент времени позжий, чем предыдущее. Сеть может содержать те же ключевые понятия, но может не содержать некоторых из них, которые выбыли из описываемой структуры, а может включать в себя другие понятия, которые появились в описываемой текстом структуре за это время. И, главное, весовые характеристики содержащихся в сети понятий могут отличаться от их весовых характеристик, какие были в первой сети.

Соединим одинаковые ключевые понятия обеих сетей связями, толщина которых будет пропорциональна весу ключевого понятия. Если понятия в обеих сетях имеют одинаковый вес, связь имеет одинаковую толщину от сети к сети. Если понятия имеют разные веса, связь, их соединяющая, либо

утолщается, либо утоньшается, демонстрируя динамику состояний ключевых понятий, и, таким образом, динамику состояний сети в целом.

Если мы возьмем тексты следующего временного среза, и построим еще одну сеть, и присоединим ее к двум предыдущим, то будем иметь картину разворачивания структуры (научной разработки, предметной области, социологической ситуации) во времени. И так сколько угодно шагов. Такая модель динамики процесса наглядна, удобна для исследования (сеть как статический смысловой срез исследуемой структуры представляет собой удобный для навигации по нему объект в силу ассоциативности связей между ключевыми понятиями), и обладает числовыми характеристиками, что делает ее удобной для аналитического исследования процессов, и, как следствие, удобной для автоматического анализа.

Наконец, для того, чтобы исследовать конкретный процесс в его динамике, например процесс социального стресса, выберем ключевые понятия семантической сети, которые являются лексическими и психолингвистическими метками этого процесса, и будем исследовать динамику развития количественных характеристик этих понятий. Как и другие ключевые понятия, эти могут появляться вновь, появляться последовательно на разных временных срезах, и наконец, исчезать. Могут также меняться от временного среза к срезу их численные характеристики. То есть они ведут себя как обычные ключевые понятия в динамике.

Удалим все ключевые понятия всех сетей, кроме упомянутых меток. В этом случае оставшаяся часть модели динамики текстов становится моделью динамики исследуемого процесса. Причем, суммарная числовая характеристика оставшихся ключевых понятий сети характеризует состояние процесса в текущий момент времени, а их изменение, от временного среза к срезу, характеризует динамику процесса во времени.

1.3 Формализм подхода

Для формирования однородной семантической (ассоциатиативной) сети формируется частотный портрет текста, содержащий информацию о частоте встречаемости ключевых понятий текста, представленных как корневые основы соответствующих слов, или их устойчивых сочетаний, встречающихся в тексте, а также об их совместной (попарной) встречаемости в смысловых фрагментах текста (например, в предложениях). Частотный портрет, таким образом, содержит информацию о частоте встречаемости ключевых понятий и их попарной (в терминах их ассоциативной связи) встречаемости в тексте. Использование хопфилдоподобного алгоритма позволяет перейти от частоты встречаемости к смысловому весу (вес связей при этом остается неизменным).

Эта обработка включает несколько этапов. На этапе первичной обработки из текста удаляется нетекстовая информация, текст сегментируется на слова и предложения, из текста удаляются стоп-слова, рабочие и общеупотребимые слова, а оставшиеся слова подвергаются морфологической обработке. Морфологическая обработка производится с использованием заранее подготовленного морфологического словаря (словаря флективных морфем) – словаря первого уровня - $\{B_i\}_1$. И формируется словарь второго уровня – $\{B_i\}_2$ – словарь корневых основ (и устойчивых словосочетаний).

На следующем этапе строится частотный портрет текста, то есть выявляются частоты p_i встречаемости корневых основ B_{i2} ключевых понятий (полученных в результате морфологического анализа) и их устойчивых сочетаний, и частоты p_{ij} их попарной встречаемости в предложениях текста. Одновременно формируется словарь третьего уровня $\{B_i\}_3$ - словарь пар слов.

На третьем этапе, частоты встречаемости перенормировываются в смысловые веса с использованием итеративной процедуры. В результате итеративной процедуры перенормировки наибольшие веса получают ключевые понятия, связанные с наибольшим числом других понятий с

большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста.

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\overline{E}), \qquad (1)$$

здесь $w_i(0) = p_i$; $w_{ij} = p_{ij} / p_j$ и $\sigma(\overline{E}) = 1/(1+e^{-k\overline{E}})$ - функция, нормирующая на среднее значение энергии всех вершин сети $\overline{E}$, где p_i – частота встречаемости i-го слова в тексте, p_{ij} – частота совместной встречаемости i-го и j-го слов в фрагментах текста. В дальнейшем эта информация используется для выявления предложений текста, содержащих наиболее важную информацию в тексте.

В результате получается так называемая ассоциативная (однородная) семантическая сеть N как совокупность несимметричных пар понятий $< c_i c_j >$, где c_i и c_j – ключевые понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста). Иначе семантическую сеть можно представить в виде множества звездочек $< c_i < c_j >>$, где $< c_j >$ - множество ближайших ассоциантов ключевого понятия c_i.

Под семантической сетью N. понимается множество несимметричных $< c_i c_j > \neq < c_j c_i >$ пар ключевых понятий $<< c_i c_j >>$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некоторой ситуации):

$$N \cong \{< c_i c_j >\}. \qquad (2)$$

Семантическая сеть, описанная таким образом, может быть переописана как множество так называемых звездочек $< c_i < c_j >>$:

$$N \cong \{z_i\} = \{< c_i < c_j >>\} \quad (3)$$

Под звездочкой $< c_i < c_j >>$ понимается конструкция, включающая главное событие c_i , связанное с множеством событий-ассоциантов c_j , которые являются семантическими признаками главного события, отстоящими от главного события на одну связь. Связи направлены от главного события к событиям-ассоциантам.

Последовательность одноименных звездочек, принадлежащих разным временным смысловым срезам – семантическим сетям, называется элементарным процессом π:

$$\pi = z_i(t_1) \Rightarrow z_i(t_2) \Rightarrow z_i(t_3) \Rightarrow \ldots, \quad (4)$$

где $z_i(t_k)$ – конкретная звездочка в момент времени t_k. Вес ключевого понятия в текущий момент времени, определяющий его ранг в семантической сети - $w_i(t_k)$.

События-ассоцианты c_j главного понятия звездочки c_i являются его семантическими признаками, и позволяют интерпретировать его содержательно на каждом шаге процесса.

2. Описание информационной модели оценки направленности процесса социального стресса на основе однородной семантической (ассоциативной) сети

Для реализации интеллектуального анализа полуструктурированной информации и информационного моделирования направленности процессов социального стресса на основе данных из открытых источников была реализована информационная модель, включающая в себя все этапы

обработки информации в процессе решения поставленной задачи. Модель включает в свой состав:

1) этап поиска релевантной исходной информации (текстов из открытых источников);
2) этап извлечения из подготовленных текстов семантической сети;
3) этап оценки направленности социального стресса.

Этап поиска релевантных для обработки текстов начинается с поиска по заданным источникам текстов, удовлетворяющих условиям поставленной задачи. В том числе, задается регион, для которого предполагается проведение исследований, временной промежуток $\Delta T_l = \left(t_{lend} - t_{lbeg}\right), l = 1..L$, который принимается за l-й временной срез (один из L), а также лексические и психолингвистические маркеры M, характеризующие предметную область исследуемого процесса. Последние задаются экспертом, характерны строго для своего социального процесса (или нескольких процессов, если они исследуются совместно), и, в конченом итоге, определяют качество анализа. Зато информационная модель абсолютно не зависит от предметной области. Ей все равно, что исследовать.

Процесс поиска текстов для каждого маркера $M_k, k = 1..K$ осуществляется отдельно, при заданных остальных (место и время) параметрах поиска. Полученные на этапе поиска тексты обрабатываются (по отдельности) с формированием семантической сети N (см. следующий этап), исключительно для ранжирования их в корпусе текстов – результатов поиска относительно релевантности именно этому маркеру. Для этого в каждом тексте вычисляется смысловой вес $r_j = w_i$ заданного маркера. Он определяет релевантность текста к этому маркеру.

Далее, для каждого маркера отбираются тексты, ранг которых по этому маркеру оказывается выше заданного порога $r_j \geq h_{отбора}$. Эти тексты, в совокупности по всем маркерам, составляют корпус текстов, подлежащих обработке на следующем этапе.

Этап формирования семантической сети N предусматривает несколько процедур в своем составе. Исходные тексты претерпевают предобработку, в процессе которой из них удаляется нетекстовая информация, удаляются также стоп-слова, рабочие и общеупотребимые слова, то есть слова, которые не несут смысла в этом корпусе текстов. Помимо этого, в процессе морфологического анализа, все словоформы приводятся к своим корневым основам, чтобы увеличить достоверность результатов последующей статистической обработки

На этапе частотного анализа формируется частотный портрет текста в виде первичной сети, в которой участвуют в качестве вершин оставшиеся после предобработки леммы слов, связанные между собой связями, полученными из анализа попарной встречаемости слов в смысловых фрагментах текста (например, предложениях). Как вершины сети, так и их связи имеют числовые характеристики – частоты их (вершин - p_i, и связей - p_{ij}) встречаемости в анализируемом тексте. Эти числовые характеристики нам понадобятся после перенормировки для оценки рангов наших маркеров описываемого текстом процесса.

На этапе перенормировки, в процессе итеративной процедуры, частотные характеристики вершин сети (ключевых понятий текста) пересчитываются в смысловые веса таким образом, что понятия (вершины сети), связанные с большим числом других понятий, увеличивают свой вес, в ущерб весам других понятий. Понятия, несущие в тексте максимальную смысловую нагрузку, становятся максимально весомыми. Они как бы стягивают на себя структуру текста. Становятся главными темами текста.

Далее, реализуется последний этап обработки информации, связанный с выявлением численных характеристик маркеров, их суммированием для конкретного временного среза текстов, и выявлением динамики изменений полученных таким образом обобщенных характеристик процесса от среза к срезу.

На этом этапе выбранные экспертом лексические и психолингвистические маркеры исследуемого процесса M_k, которые были заданы экспертом на первом этапе работы модели, фильтруют полученную на предыдущем этапе семантическую сеть $N = \{< c_i < c_j >>\}$. Из нее удаляются все вершины, кроме понятий-маркеров c_i, а также ближайших ассоциантов маркеров – верщин семантической сети (понятий), отстоящих от понятий-маркеров на один шаг $< c_j >$.

Каждому маркеру на каждом временном срезе ставится в соответствие его смысловой вес $w_j = 0..100$, полученный на предыдущем этапе, который становится рангом этого маркера M_j для этого временного среза l.

Для всех маркеров вычисляется произведение $П_j$ статуса маркера («хорошо-нейтрально-плохо») $S_j = (-1, 0, +1)$ на его ранг:

$$П_j = S_j * r_j.$$

И полученные для каждого маркера M_j произведения $П_j$ суммируются по всем маркерам:

$$П = \sum П_j.$$

Таким образом, получается суммарная характеристика П(l) временного среза l оцениваемого процесса.

Далее строится график суммарных характеристик П(l) срезов. Или, по-другому, визуализируется ряд семантических сетей, включающих помимо маркеров также их ближайшие ассоцианты $c_i < c_j >$. При этом центральное понятие-маркер c_i в сети имеет размер, пропорциональный его рангу, цвет – соответствующий его статусу «хорошо-нейтрально-плохо», и все центральные понятия связываются между собой временными связями своего цвета.

Таким образом, на графике видны основные тенденции каждого маркера во времени, а их ближайшие ассоцианты позволяют проинтерпретировать получившуюся картину.

3 Информационное моделирование оценки направленности процесса социального стресса

В качестве примера представлено построение модели оценки направленности процесса социального стресса на примере текстов по тематике внутренней политики РФ на основе новостных материалов портала newsru.com в свете отношений правительства и общества. Модель включает в свой состав:

1) тексты из открытых источников, относящиеся к одной теме и разным временным срезам;
2) семантические сети, построенные для корпусов текстов каждого среза;
3) оценку направленности социального стресса, основанную на характеристиках лексических и психолингвистических маркеров.

Этап поиска релевантных для обработки текстов начинается с поиска по заданным источникам текстов, удовлетворяющих условиям поставленной задачи. Процесс поиска текстов для каждого маркера $M_k, k = 1..K$ осуществляется отдельно, при заданных остальных (место и время) параметрах поиска. Полученные на этапе поиска тексты обрабатываются (по отдельности) с формированием семантической сети N (см. следующий этап), исключительно для ранжирования их в корпусе текстов – результатов поиска относительно релевантности именно этому маркеру. Для этого в каждом тексте вычисляется смысловой вес $r_j = w_i$ заданного маркера. Он определяет релевантность текста к этому маркеру.

Задачей моделирования оценки направленности процесса социального стресса будет являться построение модели оценки направленности процесса социального стресса на примере текстов по тематике внутренней политики РФ на основе новостных материалов портала newsru.com в свете отношений правительства и общества.

Перечень выбранных экспертным путём маркеров: "конфликт" (ранг - 2), "консенсус" (ранг 1), "согласие" (ранг 2), "бесконфликтность" (ранг 2). Ранг отражает, с содержательной стороны, степень выраженности отношений

согласия между правительством и обществом, чем выше ранг - тем более выражено согласие. С формальной вычислительной стороны ранг отражает вклад веса каждого маркера в итоговую интегральную оценку процесса социального стресса.

Тематические термины: "правительство", "общество".

Временной срез: сентябрь и октябрь 2013 года.

Выбранные тексты: выбрано 17 новостных текстов за октябрь 2013 и 22 новостных текста за сентябрь 2013. Не во всех текстах встречаются все выбранные маркеры. Надёжность оценки выбранного среза прямо зависит от объема корпуса текстов с выбранными маркерами в данном срезе.

Следующий этап обработки заключается в объединении всех текстов для каждого среза в единый текст для каждого среза и обработке модулем автоматического анализа полуструктурированной информации с удалением стоп-слов. В результате обработки выводится итоговая семантическая сеть, содержащая среди понятий маркеры, которым приписаны веса. Эти веса используются для вычислений на следующем этапе.

Последний этап обработки информации связан с выявлением численных характеристик маркеров, их суммированием для конкретного временного среза текстов, и выявлением динамики изменений полученных таким образом обобщенных характеристик процесса от среза к срезу.

Таблица 1. Оценка направленности процесса.

Маркер	Ранг	Вес в семантической сети среза для 09.2013	Вклад в оценку стресса 09.2013	Вес в семантической сети среза для 10.2013	Вклад в оценку стресса 10.2013
конфликт	-2	65	-130	79	-158
консенсус	1	72	72	98	98
согласие	2	54	108	93	186
бесконфликтность	2	83	166	52	104

Итоговая сумма:			216		230

Итоговые суммы являются интегральной оценкой социального стресса по данной тематике для данного временного среза. Однако, бессмысленно рассматривать интегральные оценки в отрыве от динамики их изменения, поскольку сумма рангов участвующих в рассмотрении маркеров несбалансированна и почти наверняка имеет отклонение в ту или иную сторону, что приведёт к отклонению интегральной оценки в случайном направлении. Состав маркеров одинаков для разных временных срезов и, поэтому, разница интегральных оценок между срезами не подвержена этому эффекту - случайные компоненты вклада взаимоуничтожаются.

Среднеквадратичное отклонение интегральной оценки, вычисленной методом бутстрепа (формирование случайной подвыборки) составляет 8.7. Оценка динамики, оцененное на данной выборке с данным распределением и с данным СКО является достоверной.

Разница интегральных оценок есть оценка направленности процесса социального стресса. Показателем качества модели является устойчивость оценки направленности относительно добавления или удаления маркера сходного с имеющимися типа, что означает что предпочтительны модели с большим числом оцениваемых маркеров и большее число тематических текстов должно подвергаться анализу. Возможна "калибровка" модели с целью вычисления среднего отклонения интегральной оценки при наличии устойчивости модели (признаваемой экспертно). Для этого следует принять некоторый временный срез за "0", точку отсчету и значение интегральной оценки в этом срезе вычитать в дальнейшем из интегральных оценок прочих временных срезов.

Вывод и оценка модели: данная модель имеет несбалансированный вклад множества рассматриваемых маркеров (общий вклад маркеров равен +3, средний +0.75) и на рассматриваемой выборке новостных текстов по тематике взаимоотношений общества и правительства (задаваемых

ключевыми словами "правительство" и "общество") обнаруживает положительную направленность процесса социального стресса от временного среза за сентябрь 2013 г. до временного среза за октябрь 2013 г.

ЗАКЛЮЧЕНИЕ

Описанный в статье подход к моделированию динамики процессов основан на хорошо зарекомендовавшей себя технологии автоматического смыслового анализа текстовой информации. В процессе обработки текста формируется ассоциативная сеть, ключевые понятия которой, в том числе, лексические и психолингвистические маркеры анализируемого процесса, ранжируются их смысловым весом. Умноженный на статус маркера на шкале «хорошо-плохо», этот вес дает значение вклада маркера в характеристику состояния процесса. Изменение суммарной для всех маркеров характеристики процесса от временного среза к временному срезу и характеризует направленность процесса. Приведенный пример не демонстрирует качество обработки, а лишь иллюстрирует работу механизма оценки. Предлагаемый подход является для эксперта инструментом моделирования процесса, настраивая который, подстраивая под свои представления, под реальный процесс, можно добиться адекватности моделирования.

Работа была выполнена в рамках НИР «Исследование и разработка программного обеспечения понимания неструктурированной текстовой информации на русском и английском языках на базе создания методов компьютерного полного лингвистического анализа» (При финансовой поддержке Министерства образования и науки Российской Федерации по Госконтракту от 10 октября 2013г. № 14.514.11.4114).

Литература:

1. Конторович С.Д., Литвинов С.В., Носко В.И. Методика мониторинга и моделирования структуры политически активного сегмента социальных сетей [Электронный ресурс] / С.Д. Конторович, С.В. Литвинов,

В. И. Носко // «Инженерный вестник Дона», 2011, №4. – Режим доступа: http://ivdon.ru/magazine/archive/n4y2011/642/2/1428 (доступ свободный) – Загл. с экрана. – Яз. рус.

2. Розин М.Д., Свечкарев В.П., Конторович С.Д., Литвинов С.В., Носко В.И. Исследование социальных сетей как площадки социальной коммуникации рунета, используемой в целях предвыборной агитации / М.Д. Розин, В.П. Свечкарев, С.Д. Конторович, С.В. Литвинов, В.И. Носко // «Инженерный вестник Дона», 2011, №1. – Режим доступа: http://http://ivdon.ru/magazine/archive/n1y2011/397 (доступ свободный) – Загл. с экрана. – Яз. рус.

3. Ahuja, V. and Medury, Y. (2011) Corporate blog as e-CRM tools Building Consumer engagement through content management. Journal of Database Marketing & Customer Strategy Management 17 (2): 91–105.

4. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке: труды международной конференции "Диалог, 2011". С.510–522.

5. Pang B. & Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - pp.1-135.

6. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques// Language. 2002. pp: 79-86.

7. Kerstin Denecke Using SentiWordNet for multilingual sentiment analysis// IEEE 24th International Conference on Data Engineering Workshop. 2008 pp: 507-512.

8. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. Sentiment strength detection in short informal text.// Journal of the American Society for Information Science and Technology, Vol., 2544–2558. 2010.

9. M.Taboada, J. Brooke, M.Tofiloski, K. Voll, M. Stede. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267-307 (2011).

10. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), P. 2544–2558.

11. M. Thelwall, K. Buckley, G. Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology. 63, 163–173 (2012).

12. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai. Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology. 61, 2544–2558 (2010).

13. Thelen, M., & Riloff, E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts.// Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP), 214-221. Morristown, NJ, USA,2002.

14. Харламов, А.А. Автоматический структурный анализ текстов [Электронный ресурс] / А.А. Харламов // Открытые системы. – 2002 – № 10. – Режим доступа: http://www.osp.ru/os/2002/10/ 182010/. – 20.12.2011 г.

15. Харламов А.А. Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды. / А.А. Харламов, В.В. Раевский // Речевые технологии, N 3, 2008. Стр. 27-35.

16. Харламов А. А. Семантические сети как формальная основа решения проблемы интеграции интеллектуальных систем. Формализм автоматического формирования семантической сети с помощью преобразования в многомерное пространство / А. А. Харламов, Ермоленко Т. В. //Материалы международной научно-технической конференции OSTIS-2011, 87-96 стр. Минск БГУИР.

17. Технология для автоматической смысловой обработки текстов [Электронный ресурс] / Сайт компании ООО «Научно-производственный инновационный центр Микросистемы»

http://www.analyst.ru/index.php?lang=eng&dir=content/downloads/ – Режим доступа: (доступ свободный) – Загл. с экрана. – Яз. рус.

18. Глоссарий [Электронный ресурс] / Глоссарий.ru – Режим доступа: http://www.glossary.ru/cgi-bin/gl_sch2.cgi?RRywlxx: (доступ свободный) – Загл. с экрана. – Яз. рус.

19. Социальные процессы [Электронный ресурс] / Новосибирский государственный аграрный университет – Режим доступа: http://nsau.edu.ru (доступ свободный) – Загл. с экрана. – Яз. рус.