УДК 004.421

А. Г. Додонов, Д. В. Ландэ, Б. А. Березин
Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Метод построения и использования семантических моделей для мониторинга общественного мнения

Предложен метод построения и использования семантических моделей с целью непрерывного во времени мониторинга общественного мнения. Семантические модели позволяют использовать при мониторинге общественного мнения результаты лингвостатистического анализа текстов, применения методов извлечения информации, содержащейся в текстах из сети Интернет. Предложена процедура мониторинга общественного мнения, которая включает три этапа: построение и кластеризацию; отбор документов и определение тональности тематик; визуализацию результатов. Показано построение семантических моделей, применение методов кластерного анализа для определения актуальных тематик, оценивание доли и тональности отдельных подтем в составе общего тематического потока информации. Полученные результаты подтверждают возможность использования предложенного метода мониторинга общественного мнения в различных предметных областях.

Ключевые слова: *модель предметной области, семантическая модель, кластерный анализ, анализ тональности, контент-мониторинг, мониторинг общественного мнения.*

Постановка проблемы

Под семантической моделью (СМ) в рамках данной работы будем понимать модель предметной области, имеющую вид ориентированного графа, вершины которого соответствуют концептам предметной области, а дуги задают отношения между ними. Концептами могут быть понятия, события, процессы, т.е. такая семантическая модель может трактоваться как семантическая карта предметной области.

Информация, которая создается пользователями сети Интернет, отражает общественное мнение по разным вопросам и может собираться, анализироваться сис-

темами контент-мониторинга и учитываться при планировании деятельности компаний, организаций и т.п. Семантические модели позволяют использовать при мониторинге общественного мнения результаты лингвостатистического анализа текстов (Text Mining), применения методов извлечения информации (Information Extraction), содержащейся в текстах из сети Интернет. В то время как существующие проекты анализа общественного мнения больше ориентированы на разовые (статичные) исследования общественного мнения относительно объектов и явлений, в данной работе предлагается метод автоматизированного построения и использования СМ на основе непрерывного во времени мониторинга общественного мнения в сети Интернет.

Анализ публикаций

В рамках данной работы авторами предлагается проведение анализа общественного мнения на основе методов обработки естественного языка (Natural Language Processing). Такой анализ направлен на определение отношения субъекта мониторинга общественного мнения к выбранной теме. Одной из основных задач анализа общественного мнения является классификация эмоциональной окраски текста (положительной, отрицательной или нейтральной).

В обзорных работах, посвященных анализу, извлечению мнений, настроений (Sentiment Analysis — SA, Opinion Mining — OM) отмечается, что это компьютерное изучение мнений, отношения людей к объекту, концепту, который может представлять личности, события или темы [1, 2]. В этих работах выделяются уровни анализа мнений: уровень документа, уровень предложения и аспектный уровень, когда рассматривается мнение в отношении некоторого концепта. Например, в обзорах продуктов сам продукт обычно является сущностью, а все, что связано с этим продуктом (цена, качество и т.д.) являются аспектами этого продукта. Анализ бывает часто связан с поиском не только общих мнений относительно концепта, но и с обнаружением мнений об аспектах. Некоторые подходы используют фиксированный, предопределенный список аспектов, в то время как другие извлекают аспекты из анализируемого текста.

В работе [3] проанализировано отличие между общественным мнением о генетически-модифицированных организмах, представленном в ресурсах Интернет, и мнением экспертов в научных публикациях. Для этого рассматривался контент веб-сайтов из результатов поиска в Google, заголовки статей из Google News, найденные по тематическому запросу, и т.д. Для этих ресурсов были построены три семантические сети на основе анализа смежности слов, в качестве концептов использовались слова с частотой повторения выше среднего значения. В результате были определены центральные слова в каждой сети, находились общие слова в разных сетях, оценивалась тональность отдельных фрагментов сетей.

В [4] анализ общественного мнения состоял в выявлении тематической структуры в массиве комментариев при обсуждении фильма на канале Youtube. При этом сравнивались результаты семантического анализа и тематического моделирования. Для анализа на сервере службы Youtube было собрано около трех тысяч комментариев. При построении СМ в качестве вершин использовались биграммы. Тематическая структура обсуждения была выявлена с помощью кластеризации основной компоненты построенной семантической сети. Сделан вывод,

что семантический анализ может дополнять тематическое моделирование либо служить альтернативой.

В [5] анализировалось общественное мнение о выборах президента США в 2012 г. на основе новостных статей, опубликованных в Интернет. С помощью системы мониторинга были собрано более 81 тыс. англоязычных статей из 400 информационных агентств. На основе этих ресурсов были выделены триплеты «субъект-глагол-объект» и с их помощью построены два семантических графа, отражающих основные действующие лица, их избирательные лагеря и т.д. Результаты анализа избирательной кампании были получены путем исследования характеристик семантических графов.

Метод построения и использования семантических моделей

В данной работе предложен метод построения и использования СМ для задач мониторинга общественного мнения (МОМ) в сети Интернет, предусматривающий три этапа [6]:

— построение и кластеризацию СМ;

— отбор документов и определение тональности тематик;

— визуализацию результатов.

На первом этапе производится:

— выборка массива документов для построения СМ;

— нахождение концептов; определение связей СМ путем построения компактифицированного графа горизонтальной видимости [7];

— кластеризация графа;

— формирование запросов, соответствующих кластерам (на основе найденных кластеров экспертами выделяются тематики и формулируются запросы для отбора соответствующих документов).

На втором этапе:

— производится отбор документов, соответствующих тематикам (подтемы), из общего информационного потока с помощью запросов;

— определяется их доля в общем потоке документов;

— определяется тональность документов соответствующих тематик.

На третьем этапе тематики с тональностями:

— визуализируются на карте;

— записываются состояния в базу данных (БД) системы мониторинга для последующего получения динамики изменения результатов во времени.

Далее рассмотрены основные операции, выполняемые в составе этих трех этапов.

Этап построения и кластеризации СМ. Выборка массива документов для построения семантической модели. На основе заданного объекта мониторинга формулируется запрос на выборку массива документов.

Нахождение концептов. Входящие в массив документы проходят предварительную обработку, удаление служебной информации, а также стоп-слов, не несущих смысловой нагрузки. Может проводиться стемминг (приведение слов к основе). Затем, на основе учета частоты слов в массиве документов либо с помощью других известных метрик, например, *TFIDF* , из слов массива документов отбираются наиболее важные, имеющие наибольший вес понятия [8].

Определение связей СМ путем построения графа горизонтальной видимости. Для определения связей между концептами и построения семантической модели используется алгоритм компактифицированного графа горизонтальной видимости (КГГВ) [7], который предусматривает три шага.

1. На горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (визуально — набор вертикальных линий).

2. Строится традиционный граф горизонтальной видимости. Для этого между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию.

3. Полученная на предыдущем шаге сеть компактифицируется. Все узлы с данным словом, объединяются в один узел. Все связи таких узлов также объединяются.

Особенность использования алгоритма КГГВ в данной работе состоит в том, что его первые два шага выполняются отдельно для каждого предложения анализируемого текста. После этого полученная сеть компактифицируется. В процессе разработки предложенного метода, проводилось исследование построения СМ для документов, собираемым по темам Шелкового пути, Норд Стрим-2, ГМО и некоторым другим. Фрагмент графа семантической модели, построенной для 28 концептов темы OBOR с помощью описанного алгоритма, приведен на рис. 1.

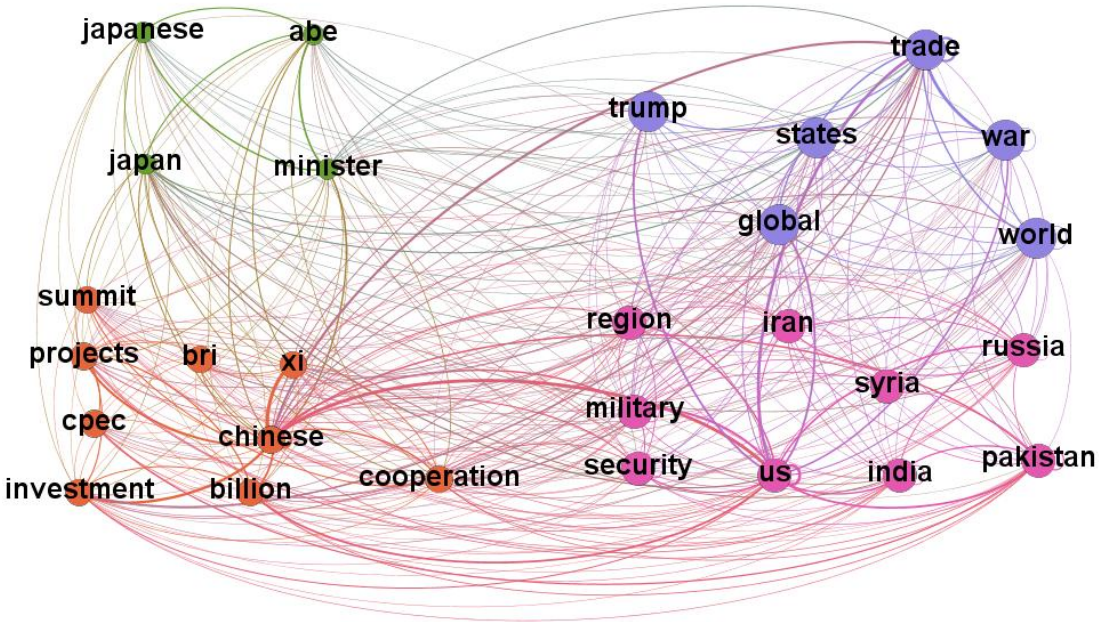

Рис. 1. Фрагмент графа семантической модели для 28 концептов темы OBOR

Кластеризация графа СМ. Учитывая актуальность аспектного уровня анализа мнений, после построения семантической модели анализируется ее сетевая структура с помощью алгоритмов кластеризации графов — выявления сообществ (clastering graph, community detection). В работе [9] сообщество определяется как плотно связанная группа узлов, которая слабо связана с остальной частью сети.

Идентификация сообществ в сети является сложной проблемой из-за существования многочисленных определений сообществ и трудоемкости алгоритмов обнаружения сообществ. В [9] рассмотрено более десятка алгоритмов кластеризации для выявления сообществ как непересекающихся, так и перекрывающихся (и сообществ обоих типов). Для кластеризации графов СМ в данной работе рассматривалось применение различных известных алгоритмов. Лучшие, наиболее осмысленные результаты, были получены при использовании алгоритмов Louvain, Leading Eigenvector, а также Walktrap.

Среди известных алгоритмов поиска сообществ в графе можно выделить алгоритм Louvain [10], в соответствии с которым в начале алгоритма каждая вершина образует отдельное сообщество. Шаг алгоритма состоит из двух фаз. На первой фазе для каждой вершины происходит попытка найти сообщество, перемещение в которое даст максимальное общее положительное изменение модулярности. Переместить вершину можно только по смежным ребрам, то есть только в те сообщества, которым принадлежат вершины, смежные с данной. Просмотр всех вершин продолжается до тех пор, пока происходит хотя бы одно перемещение вершины. На второй фазе происходит сжатие графа: вершины, входящие в одно сообщество образуют новую супервершину с соответствующим преобразованием ребер. Алгоритм останавливается, когда граф перестает изменяться.

Алгоритм Leading Eigenvector [11] использует положения спектральной теории графов. Он основан на максимизации модулярности путем разделения графа на две группы вершин, с использованием спектра графа, предложенный в работе. Алгоритм состоит в том, что находится собственный вектор, соответствующий первой компоненте спектра матрицы модулярности. Разбиение определяется оценкой ведущего собственного вектора матрицы модульности.

Алгоритм кластеризации Walktrap [9], позволяет находить плотно связанные подграфы (сообщества в графе) на основе случайных блужданий. Принцип алгоритма состоит в том, что короткие случайные блуждания, как правило, остаются в одном и том же сообществе. Утверждается, что переходы из одного кластера в другой должны происходить достаточно редко. Исходя из этого свойства, вводится метрика для схожести вершин.

Результаты кластеризации графов СМ, построенных для документов, собранных по теме Нового шелкового пути («One Belt, One Road» — OBOR) с помощью рассмотренных алгоритмов, отражены в табл. 1.

Формирование запросов, соответствующих кластерам. В результате кластеризации графа семантической модели были найдены множества наиболее связанных вершин графа, соответствующих выявленным кластерам, т.е. множества близких понятий. На основе этих понятий в общем потоке документов, характеризующих анализируемую предметную область, выделяются тематики, аспекты. Эксперты в данной предметной области дают названия этим тематикам и формулируют запросы для последующего отбора с помощью информационно-поисковой системы документов, соответствующих тематикам, из общего потока документов, характеризующих общую предметную область.

Для рассматриваемой темы OBOR, на основе найденных кластеров, экспертами были сформулированы четыре тематики и соответствующие им запросы, приведенные в табл. 2. На этом, периодически повторяемый этап обучения систе-

мы мониторинга (период повторения от нескольких часов до суток), реализуемый при помощи алгоритма построения и кластеризации СМ, заканчивается.

Этап отбора документов и определения тональности тематик. Отбор документов тематик из общего информационного потока с помощью запросов. Из общего потока документов, формируемого по поисковому запросу, характеризующему предметную область, отбираются документы тематик с помощью поисковых запросов, сформулированных экспертами на основе выявленных в семантической модели кластеров. Для отобранных документов каждой из тематик определяется их доля в общем потоке документов.

Для потока документов, формируемого по поисковому запросу к системе контент-мониторинга InfoStream ***(one-road)&(one-belt)&china***, характеризующего тему OBOR**,** названия тематик и соответствующие запросы приведены в табл. 2. Доли документов тематик, отобранных с помощью сформулированных запросов из общего потока, приведены в табл. 3.

Определение тональности документов тематик. Для документов каждой из выявленных тематик определяется тональность — позитивная, негативная, нейтральная на основе анализа слов, входящих в состав документов, относящихся к темам. Для определения тональности могут быть использованы алгоритмы, предложенные в [5]. Под тональностью текста в данном случае понимаются позитивная, негативная или нейтральная эмоциональные окраски как всего текстового документа, так и отдельных его частей, имеющих отношения к определенным понятиям, таким как персоны, организации, бренды и т.п. В задаче определения тональности проверяется как минимум три показателя эмоциональной окраски: позитивная, негативная, нейтральная и, зачастую, существует потребность также в проверке комбинации этих гипотез (например, для выявления уровня «экспрессивности» текста). Тональности документов тематик, отобранных из общего потока документов темы OBOR, приведены в табл. 3.

Этап визуализации результатов. На данном этапе выполняется визуализация найденных тематик с тональностями на карте. Результаты мониторинга визуализируются в режиме реального времени на географической карте с привязкой к конкретным объектам. По каждой выявленной в общем потоке документов тематике отображается диаграмма с указанием названия тематики и доли документов общего потока, приходящейся на эту тематику, а также с отображением доли документов положительной, отрицательной и нейтральной тональности внутри тематик (табл. 3, рис. 2). На географической карте показываются: тематики, выявленные в потоке входных документов; доля документов по каждой тематике; тональность документов по тематикам, а также динамика изменения результатов во времени. Найденные в процессе мониторинга состояния записываются в БД мониторинга для последующего получения динамики изменения результатов во времени. Изменение доли документов и тональности сформулированных тематик (подтем) главной темы OBOR по неделям приведено на рис. 3–5.

Операции, выполняемые в составе трех этапов рассмотренного метода реализованы с помощью средств программного пакета Gephi (http://gephi.org), а также с помощью программных средств, разработанных на языке программирования для статистических расчетов *R*. Результаты использования предложенного метода приведены ниже.

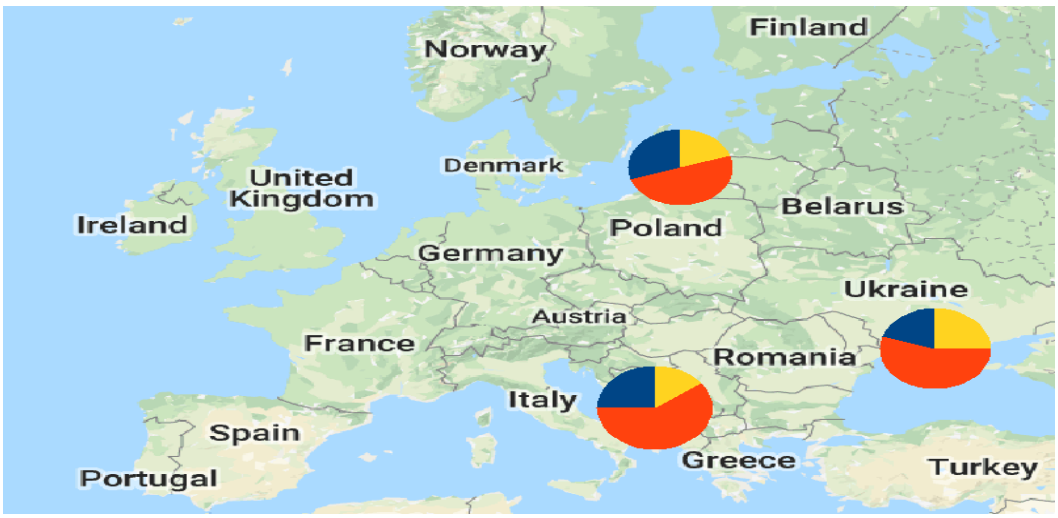

Рис. 2. Визуализация результатов мониторинга на географической карте.
По каждой из найденных тематик на карте отображаются доля ее документов
в общем входном потоке и тональность документов

Полученные результаты

Возможности применения предложенного метода построения и использования СМ для мониторинга общественного мнения анализировались на основе результатов мониторинга интернет-ресурсов по нескольким темам:

One Belt, One Road (OBOR) — инициатива Китайской Народной Республики по объединенным проектам «Экономического пояса Шелкового пути» и «Морского Шелкового пути XXI века»;

Nord Stream — проект газопровода из России в Германию через Балтийское море;

GMO — генетически модифицированные организмы и некоторые другие темы. Для мониторинга общественного мнения по теме OBOR анализировался массив из 1000 англоязычных документов (период с 30.11.2018 по 25.07.2018), собранных при помощи запроса ***(one-road)&(one-belt)&china*** с использованием системы InfoStream.

На первом этапе, после выборки массива документов и его предварительной обработки, нахождения концептов (на основе частоты использования терминов, а также на основе показателя *TFIDF*) были построены соответствующие СМ (рис. 1). Кластеры понятий, получаемые с помощью алгоритмов Louvain, Leading Eigenvector и Walktrap на основе построенных СМ, приведены в табл. 1.

В первой строке табл. 1 приведены названия трех алгоритмов, которые применялись для кластеризации СМ. В столбцах 1, 2 — результаты применения алгоритмов кластеризации к СМ, построенной на концептах, отобранных по частоте использования в документах, а в столбцах 3–5 — результаты кластеризации СМ, построенной на концептах, отобранных при помощи показателя TFIDF. В столбцах фрагмента таблицы кластеров приведены взаимосвязанные совокупности понятий анализируемой темы, найденные с помощью трех алгоритмов кластеризации. На основе сопоставления полученных кластеров, экспертами в данной пред-

метной области могут быть сформулированы названия соответствующих им тематик, а также запросы для выборки документов, относящихся к тематикам, определения их доли и тональности (табл. 2).

Таблица 1. Фрагмент таблицы кластеров, полученных при использовании разных алгоритмов кластеризации графа СМ для документов темы OBOR

Louvain	Walktrap	Louvain (TFIDF)	Walktrap (TFIDF)	Leading eigenvector (TFIDF)
economy military world countries country state global war trade projects economic cooperation security political infrastructure investment development international support national		**india** military power **pakistan** country **us** strategic might russia american political iran russian defense syria turkey		india world relations war united states russia japan think iran defense syria
chinese president people government market **xi** business million debt including beijing billion summit leaders		sri chinese president people government **projects xi** debt trump project **investment** cpec billion port bri		sri chinese president obor people government **projects xi** cooperation debt project **investment** beijing summit port bri
india china **south** asia power **region pakistan** relations influence **us** strategic east europe russia american trump japan africa part sea		**south region** relations indian xinjiang economic cooperation africa **development** beijing summit pacific taiwan		china's military **south** power **region** state global strategic might economic american security political africa research uighur pacific taiwan
prime **minister** foreign policy	prime **minister** foreign	prime **minister** foreign **japanese abe** mahathir malaysia	prime **minister japanese abe** mahathir malaysia	prime **minister** foreign **japanese abe** mahathir russian
china's road belt initiative project silk	road belt initiative silk	china's global obor market research		

Таблица 2. Тематики и запросы, сформулированные на основе найденных кластеров

Номер	Ключевые слова тематики, запрос	Название тематики
1	xi projects investment	Председатель КНР (президент) Си Цзиньпин об инвестиционных проектах в составе инициативы bri
2	india pakistan us	Отношение Индии, Пакистана, США и др стран к инициативе
3	south region development	Отношение стран к развитию южного региона в рамках инициативы
4	japan minister abe	Премьер министр Японии Абэ об инициативе

На втором этапе, доли документов тематик, отобранных с помощью сформулированных запросов из общего потока, приведены в табл. 3. Тут же приведены тональности документов сформулированных тематик, отобранных из общего потока документов темы OBOR.

Таблица 3. Доля документов сформулированных тематик в общем потоке темы «OBOR» и их тональность

Тематика	Доля	Негативная	Нейтральная	Позитивная
xi projects investment	10,1 % (101)	2 % (2)	0	98 % (99)
japan minister abe	9,2 % (92)	1 % (1)	0	99 % (91)
india pakistan us	8 % (80)	5 % (4)	0	95 % (76)
south region development	15,3 % (153)	5 % (8)	1 % (1)	94 % (144)

На третьем этапе выполняется визуализация полученных результатов. Общий вид интерфейса для визуализации найденных тематик, их доли в общем потоке и тональностей приведен на рис. 2. Динамика изменения доли документов и тональности сформулированных тематик в составе документов темы OBOR по неделям показана на рис. 3–5. На рис. 3 показано изменение доли документов сформулированных тематик (Тематика 1 – Тематика 4, четыре нижних графика) в составе документов темы OBOR (Тема, верхний график) по неделям. На рис. 4 отражены графики изменения количества документов с позитивной, негативной и нейтральной тональностью в тематике «South Region Development» по неделям. На рис. 5 показаны графики изменения количества документов с позитивной, негативной и нейтральной тональностью в тематике «India Pakistan US» по неделям.

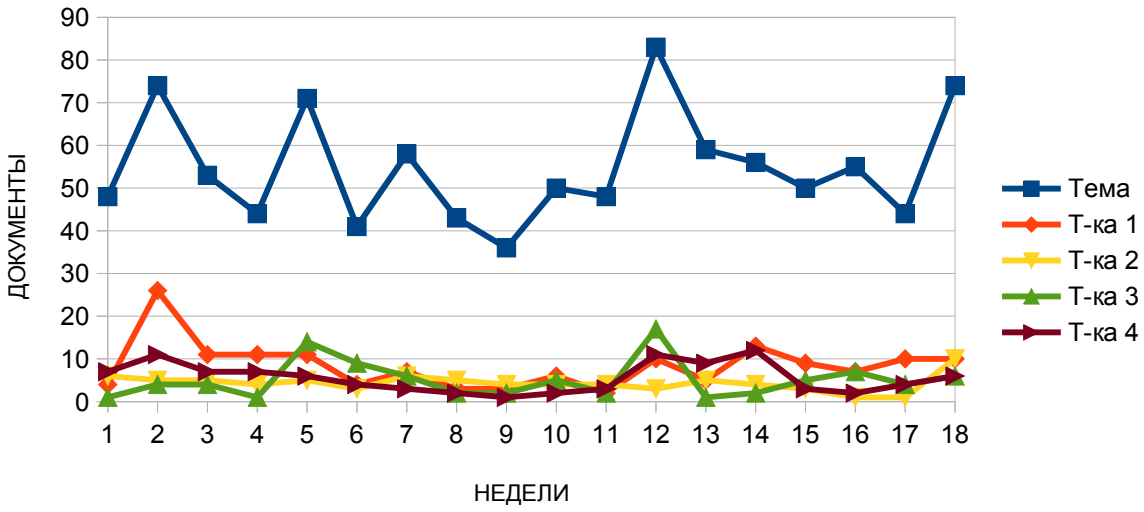

Рис. 3. Изменение доли документов сформулированных тематик (Тематика 1 – Тематика 4 — четыре нижних графика) в составе документов темы OBOR (Тема, верхний график) по неделям

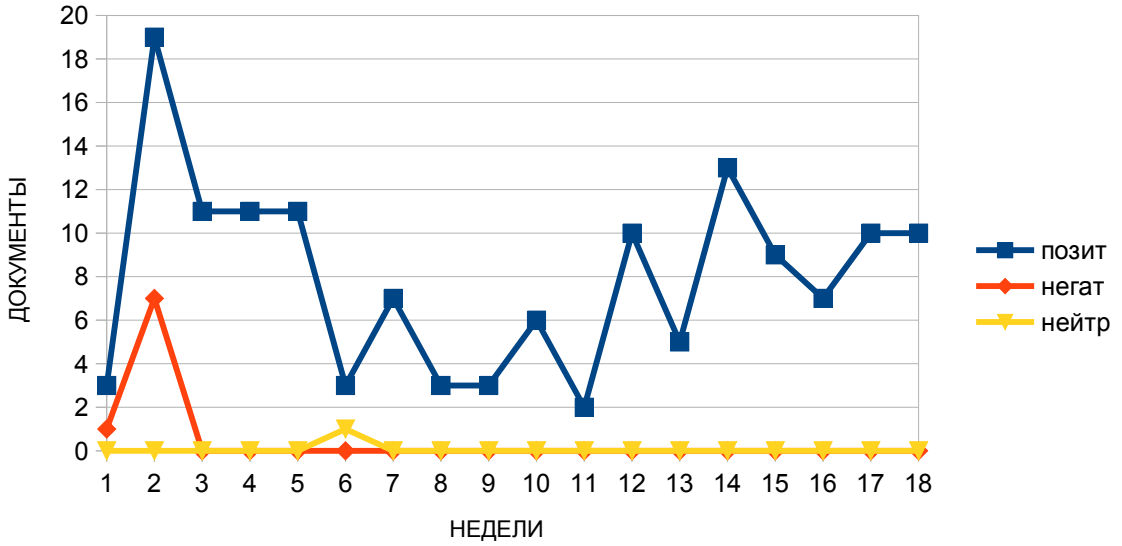

Рис. 4. Изменение количества документов с позитивной, негативной и нейтральной тональностью в тематике «South Region Development» по неделям

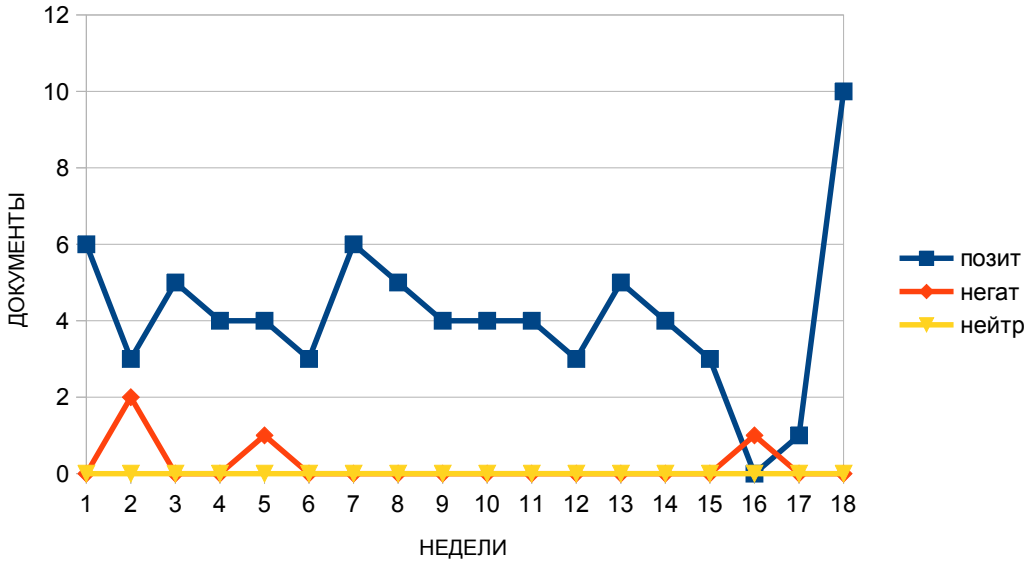

Рис. 5. Изменение количества документов с позитивной, негативной и нейтральной тональностями в тематике «India Pakistan US» по неделям

Кроме темы OBOR рассматривалось использование предложенного метода мониторинга общественного мнения для тем Nord Stream, GMO и других. Например, для темы Nord Stream, на основе массива из 1000 англоязычных документов, (собранных в период с 02.11.2018 по 18.08.2018) была построена соответствующая СМ, найдены кластеры и сформулированы тематики: merkel putin meeting (о встрече Меркель и Путина); gas transit ukraine (о транспортировке газа через Украину); european security energy market (о безопасности европейского энергетиче-

ского рынка), poland united states (об отношении Польши и США к проекту Nord Stream).

Выводы

Предложен метод построения и использования СМ для мониторинга общественного мнения, включающий три этапа: построение и кластеризацию СМ; отбор документов и определение тональности тематик; визуализацию результатов.

Показано построение СМ с помощью алгоритма компактифицированного графа горизонтальной видимости, применение методов кластерного анализа для определения актуальных тематик, оценивание доли и тональности отдельных подтем в составе общего тематического потока информации.

Полученные результаты подтверждают возможность использования предложенного метода мониторинга общественного мнения в различных предметных областях.

1. Schouten K., Frasincar F. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering. 2016. Issue 28(3). P. 813–830.

2. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey // *Ain Shams Engineering Journal*. 2014. Issue 5(4). P. 1093–1113.

3. Jiang K., Anderton B., Ronald, P., Barnett G. Semantic Network Analysis Reveals Opposing Online Representations of the Search Term «GMO». *Global Challenges*. 2018. Issue 2(1). P. 1700082. DOI: https://doi.org/10.1002/gch2.201700082.

4. Юдина Д.И., Дудина В.И. Семантическая сеть на биграммах как метод валидизации результатов тематического моделирования в социологическом исследовании. *Журнал социологии и социальной антропологии*. 2016. Вып. 19(4). С. 71–83.

5. Sudhahar S., Veltri G., Cristianini N. Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*. 2015. Issue 2(1). P. 21–49.

6. Додонов А.Г., Ландэ Д.В., Березин Б.А. Построение и использование семантических моделей для мониторинга общественного мнения. Міжнародна науково-практична конференція «Інтелектуальні технології лінгвістичного аналізу»: тези доповідей. Київ: НАУ, 2018. С. 6.

7. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text. 12th Mexican International Conference on Artificial Intelligence (MICAI), 2013. P. 209–215. DOI: 10.1109/MICAI.2013.33

8. Ланде Д.В. Підходи до автоматичного визначення термінологічних основ онтологій. Тези доповідей. Міжнародної науково-технічної конференції «Інтелектуальні технології лінгвістичного аналізу». Київ: НАУ, 2014. С. 7–9.

9. Harenberg S., Bello G., Gjeltema L., Ranshous S., Harlalka J., Seay R., Samatova N. Community detection in large-scale networks: a survey and empirical evaluation. Wiley Interdisciplinary Reviews: Computational Statistics, 2014. Issue 6(6). P. 426–439.

10. Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008. Issue 10. P. P10008

11. Newman M.E. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*. 2006. Iss. 74(3). P. 036104.

12. Lande D.V. Identification of information tonality based on Bayesian approach and neural networks. E-preprint arXiv: 0806.2738 (2008).

Поступила в редакцию 14.12.2018