

Методика мониторинга и моделирования структуры политически активного сегмента социальных сетей

С.Д. Конторович, С.В. Литвинов, В.И. Носко

Социальные сети являются удобной и популярной площадкой проведения агитационных мероприятий, на которой активно разворачивается обсуждение всех основных политических событий страны и региона, происходит информационный обмен, значимо влияющий на общественное мнение.

В последние годы наблюдается стремительное увеличение влияния событий блогосферы на политические и социальные процессы в мире, в том числе, и на политическую жизнь нашей страны. Социальные сети являются площадкой активного обсуждения всех политических событий страны, формирующей общественное мнение, и, прежде всего, молодежи – тех, в чьих руках окажется судьба страны спустя 10-15 лет. Таким образом, все более очевидной становится необходимость разработки методик и алгоритмов исследования социальной коммуникации социальных медиа и особенностей их влияния на текущие политические события [1].

В СКНЦ ВШ были разработаны методики мониторинга социальных медиа и моделирования структуры тематического сегмента блогосферы. С их помощью было проведено исследование коммуникаций в социальных медиа в контексте предстоящих выборов в Государственную Думу 4 декабря 2011 года.

Мониторинг тематического сегмента социальной сети

В соответствии с разработанной методикой мониторинг предлагается проводить в 5 этапов (см. рис. 1):

1. **Анализ предметной области.** Входящими данными для первого этапа является *Техническое задание на проведения мониторинга*. В соответствии с ним проводится анализ тематики мониторинга и поиск активных площадок блогосферы. Анализируются посты данных площадок и вычленяется массив тематического текста. Затем проводится количественный анализ данного текста.
2. **Формирование запроса.** На основании количественного представления текста формируется поисковый запрос с использованием тематических ключевых слов. Он может быть скорректирован стоп-словами для отсеивания нерелевантных результатов поиска. Ключевые слова и стоп-слова могут изменяться и / или дополняться на следующих итерациях.
3. **Поиск по блогосфере.** Исходя из рассматриваемой ситуации для проведения поиска выбирается конкретная поисковая машина. С помощью сформированного поискового запроса производится поиск с помощью выбранной поисковой машины на основании специфической семантики языка данной поисковой машины. Также возможно задание представляющего интерес для исследователя временного интервала.
4. **Анализ результатов поиска.** Осуществляется проверка шумности (количества нерелевантных результатов) выдачи. В случае высокого уровня шума необходимо на основании его анализа скорректировать ключевые слова и / или стоп-слова и вернуться к этапу 2 (формирование запроса). В случае низкого уровня шума составляется сводная таблица мнений блогосферы, включающая ссылку на публикацию, на профиль автора, указание на количество комментариев, общую тональность высказывания (позитивное, негативное, нейтральное), при необходимости – указание на тематику и т.д.
5. **Составление отчета.** Сводная таблица мнений блогосферы анализируется на ее основании составляется отчет о мониторинге.

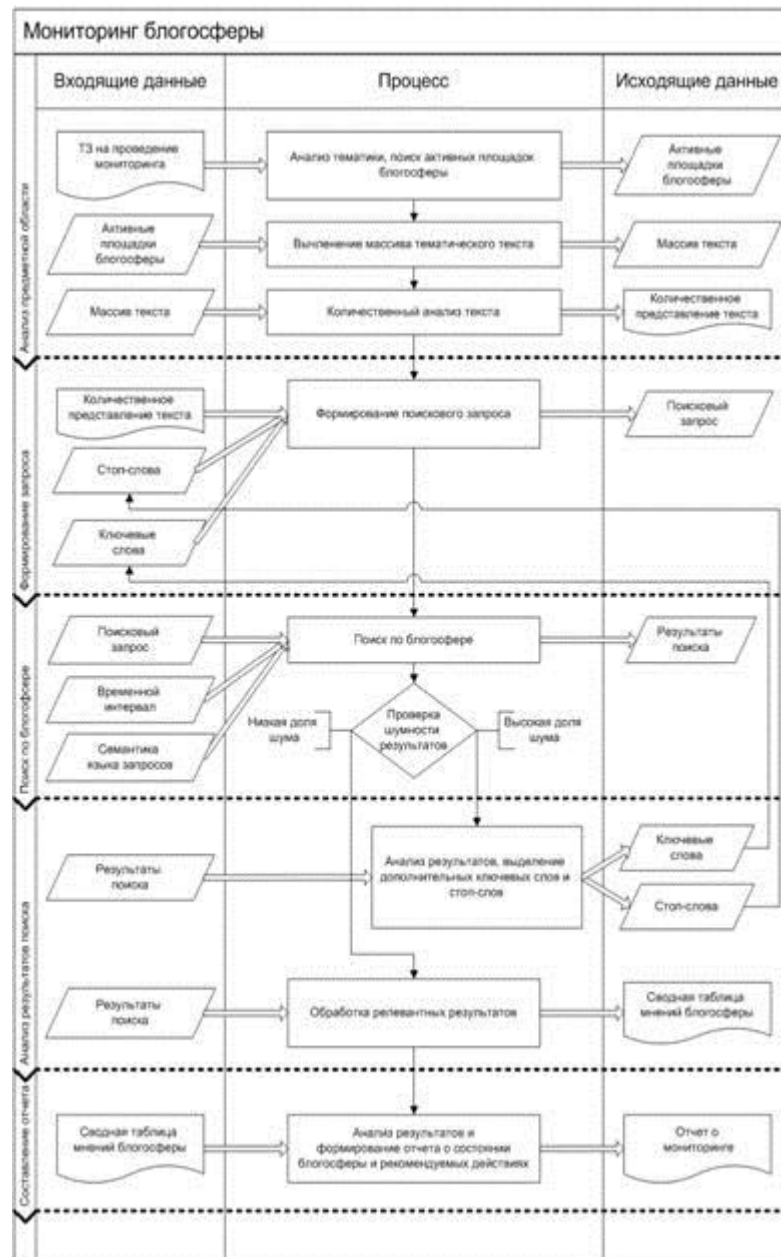


Рисунок 1 - Алгоритм мониторинга тематического сегмента блогосферы

При исследовании результатов мониторинга используется семантический анализ текста, обращаясь к содержательной стороне упоминаний: анализу подвергается смысл текстов сообщений, виды, сложность и набор характеристик объектов анализа, типы связей объектов между собой и т.д. для получения полноты представлений о спектре мнений пользователей блогосферы по конкретному поводу. В целом, анализ текстов блогосферы позволяет ответить в рамках конкретной темы на следующие вопросы:

- какая информация о событии курсирует в блогосфере прежде всего, какие события оказывают влияние на объем дискуссий;
- какова динамика объемов обсуждений темы и подтем, существующих в ее рамках;
- какие вопросы занимают аудиторию, в связи с темой, прежде всего. Как распределены темы и подтемы в рамках целевой аудитории;
- какое количество представителей аудитории оказывает влияние на обсуждение темы и почему;
- какова доля позитивных/нейтральных/негативных комментариев в общем объеме обсуждения. Что именно вызывает положительную и отрицательную реакцию аудитории;
- Особенности дискурса в рамках обсуждаемой темы (мемы, жаргонизмы и т.д.);

При проведении конкретного исследования интерес могут представлять не все перечисленные цели.

Поскольку прежде всего мониторинг ориентирован на оперативное отслеживание изменений во мнениях, суждениях относительно совершающихся событий и т.д., то основной объект интереса – оценочные высказывания по интересующему информационному поводу. Простейшая и наиболее удобная для дальнейшего практического применения операционализация отслеживания этих оценочных суждений связана с простым разделением их на позитивные, негативные и нейтральные. Для этого необходимо присутствие в текстовом массиве эмоционально окрашенной лексики в отношении объектов, персон и событий, представляющих интерес для исследователя.

На сегодняшний день существует некоторое количество автоматизированных систем классификации и анализа интернет-текстов. Основаны такие системы, как правило, на соотношении имеющего текстового фрагмента с заранее составленными тональными словарями. По совокупности выявленной эмотивной лексики текст оценивается как позитивный или негативный. Однако, ввиду особенностей используемого в постах языка с его незакрепленным порядком слов в предложении, обилием разговорной и ненормативной лексики с самыми неожиданными контекстуальными значениями, с двусмысленностями и юмором, понятными исходя из анализа подтекстов и диалогов, но не из фактического словарного смысла сказанного, к которому может обратиться автоматизированная система, можно сказать, что на сегодняшний день достаточно адекватной и эффективной автоматизированной системы анализа интернет-контента не существует [2]. Безусловно, это направление будет интенсивно развиваться в ближайшие годы, но на данный момент наиболее адекватным инструментом анализа представляется ручная обработка данных мониторинга социальных медиа.

Таким образом, целесообразно в сводной таблице, куда заносятся результаты мониторинга, оставить пустую графу «тональность», которая будет заполняться вручную при повторном прохождении таблицы. Такой способ организации данных позволяет быстро и легко обратиться к тем или иным оценкам (например, рассмотреть только негативные упоминания для выявления основных конфликтных зон, способов выражения негатива, используемых словарей и степенью эмоциональной вовлеченности авторов) ознакомиться с аудиторией, являющейся поставщиком тех или иных мнений и т.д. Проводимая в рамках регулярного мониторинга, такая процедура позволяет быстро увидеть изменения в настроениях пользователей и выявить зоны возможного влияния на них.

Модуль импорта в систему комментариев пользователей разработан таким образом, что позволяет разделять собираемую информацию на элементарные части, каждая из которых представляет собой отдельное поле в базе данных. Импорт производится гранулярно, что впоследствии дает возможность гибко фильтровать результаты. В набор собираемых данных входят:

- заголовок комментария;
- автор комментария – никнейм пользователя в сети или блоге;
- автор журнала – запись, к которой относится комментарий;
- комментарий – собственно анализируемый текст;
- ссылка на комментарий;
- журнал комментатора;
- журнал автора поста;
- дата комментирования – актуальная дата размещения комментария в сети.

В дальнейшем данная информация используется для моделирования и анализа структуры политически активного сегмента блогосферы.

Моделирование структуры тематического сегмента социальной сети

С точки зрения исследователя взаимодействий в социальных сетях блогосфера представляет собой совокупность социальных акторов и набор связей между ними. Таким образом, структура выделенного сегмента социальной сети может быть промоделирована графом, вершинами которого являются участники сети, а ребрами – отношения между ними. В качестве таких отношений в модели могут выступать связи дружбы («френдирование»), связь «автор–комментатор» и т.п. Такой подход позволяет применять

стандартные методы теории графов при анализе структуры сегмента блогосферы.

При исследовании структуры рассматриваемого сегмента социальной сети основными являются следующие задачи:

1. Определение наиболее авторитетных и влиятельных участников рассматриваемого сегмента социальной сети. Наблюдения психологов показывают, что участники социальной сети зачастую не располагают достаточной для принятия решений информацией или не могут самостоятельно обработать ее, поэтому их решения в значительной степени могут быть подвержены социальному влиянию, то есть, основываться на наблюдаемых ими решениях или представлениях других агентов [3].
2. Нахождение проводников информации – пользователей, играющих наиболее важную роль в распространении информации внутри сети. Промоделировав и визуализировав определенный сегмент социальной сети, становится возможным выделить пользователей, занимающих максимально выгодную с точки зрения коммуникации позицию; такие пользователи, не являясь авторитетами, способствуют, тем не менее, распространению информации и настроений в сети.
3. Выделение в рассматриваемом сегменте социальной сети кластеров – слабосвязанных между собой групп со схожими точками зрения и сильными связями внутри себя.
4. Обнаружение информационных брокеров – пользователей, обеспечивающих связь между кластерами социальной сети. Без этих пользователей сегмент социальной сети распался бы на несвязанные друг с другом кластеры.

Мониторинг с помощью программной платформы регионального сегмента блогосферы и социальных сетей производится в несколько этапов, на каждом из которых решается конкретная инженерная задача.

Моделирование предлагается проводить в 4 этапа (рис. 2):

1. **Получение технического задания.** В набор данных для технического задания входят набор ключевых слов, по которым осуществляется мониторинг а также список площадок социальных сетей и блогосферы. Указывается диапазон дат и рабочая гипотеза, которую в процессе мониторинга необходимо подтвердить или опровергнуть.
2. **Сбор данных.** На втором этапе с помощью специальных программных модулей производится парсинг страниц, содержащих данные о пользователях, в том числе их комментарии. Производится проверка набора полученных данных на адекватность и полноту.
3. **Анализ графов.** На третьем этапе производится импорт полученных данных в программную платформу Gephi, в которой выполняется расчет структурных параметров графа, в соответствии с поставленными задачами выявляются следующие метрики и свойства: входящее и выходящее распределение степеней в графе, промежуточность, коэффициент кластеризации и определение количества кластеров нужного размера.
4. **Формирование единой базы данных.** На заключительном этапе все собранные данные, таблицы, графики, сам граф импортируются в информационную систему для последующего повторного использования.

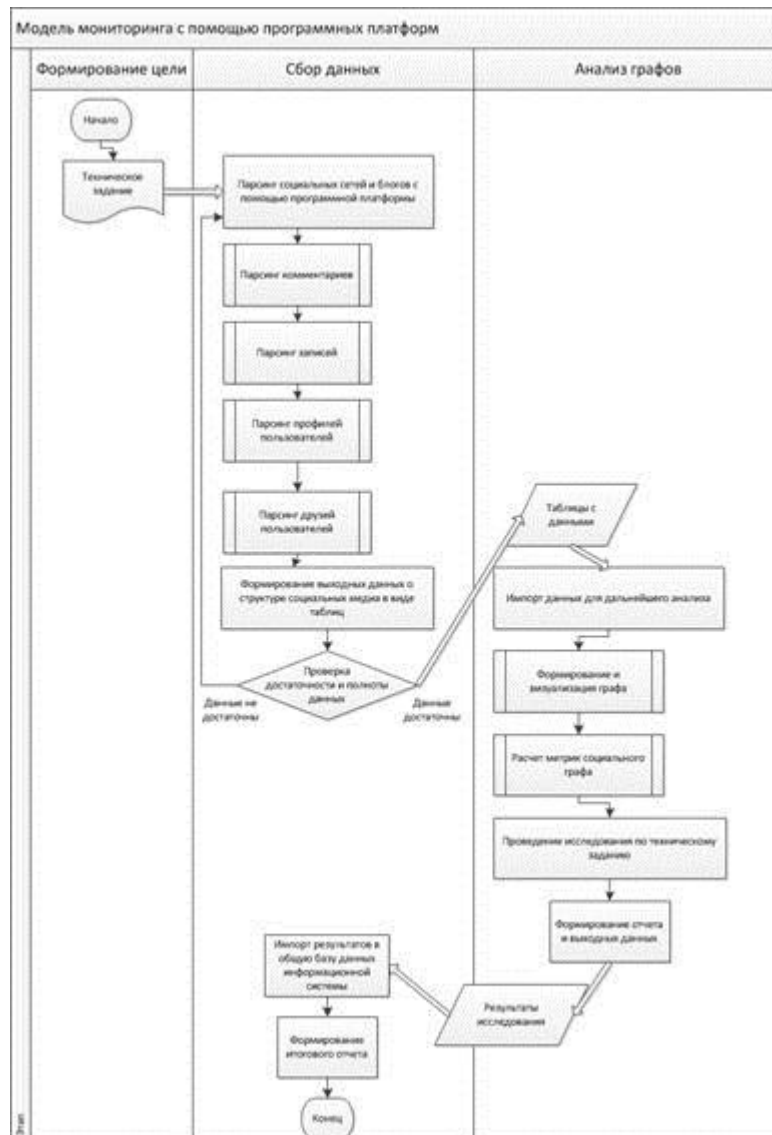


Рисунок 2 - Модель мониторинга с помощью программных платформ

Моделирование структуры сегмента социальной сети предлагается проводить в 4 этапа (см. рис. 3):

1. **Визуализация графа.** Входными данными для первого этапа является *Техническое задание на проведение моделирования* и *Данные парсинга* (см. аналитический отчет «Методика мониторинга регионального сегмента социальных медиа в период избирательных кампаний»). Производится импорт ранее полученных данных в программную платформу. Затем производится настройка и визуализация полученного графа.
2. **Расчет метрик вершин.** На втором этапе проводится сопоставление и анализ параметров вершин социального графа, которые отражают активность актора в сети: количество друзей пользователя, количество тематических комментариев, количество и качество записей пользователя. Делается расчет метрик вершин и ребер в графе.
3. **Анализ модели.** На третьем этапе выполняется расчет структурных параметров графа, в соответствии с поставленными задачами выявляются следующие метрики и свойства: входящее и выходящее распределение степеней в графе, промежуточность, коэффициент кластеризации и определение количества кластеров нужного размера. В результате анализа набирается достаточное количество информации для выработки сценария дальнейших действий.
4. **Формирование отчета.** Подводится итог, анализируются результаты моделирования, составляются рекомендации.

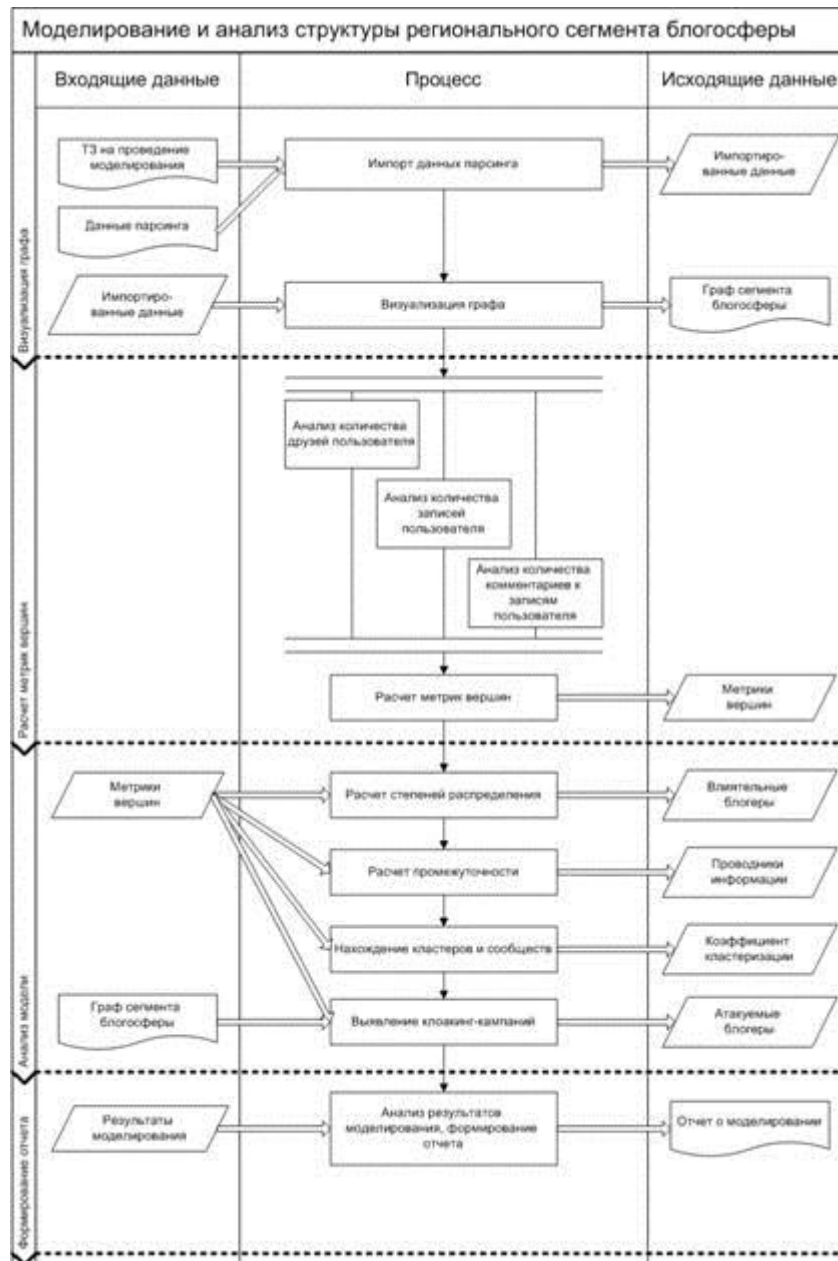


Рисунок 3 - Алгоритм моделирования регионального сегмента блогосферы

Результаты исследования

В качестве платформы исследуемого сегмента социальных медиа была выбрана блог-платформа LiveJournal («Живой журнал»). Такой выбор обусловлен, во-первых, направленностью данной сети в первую очередь именно на ведение открытых публичных дискуссий: «Живой журнал» на сегодняшний день стал основной площадкой для «гражданской журналистики» [4]. Во-вторых, «Живой журнал» характеризуется сравнительной простотой получения интересующей исследователя информации (т.н. парсинга). В настоящее время аудитория «Живого журнала» составляет 8,6 миллионов русскоязычных пользователей. При этом за последние 3 года наблюдается тенденция роста популярности LiveJournal, при этом среднегодовой рост составляет 40%. [5]. О популярности «Живого журнала» в качестве площадки политических дискуссий свидетельствует также значительное число блогов политических и общественных деятелей современной России.

Исследование проводилось в середине ноября 2011 года. В ходе исследования был проведен анализ октябрьских-ноябрьских дискуссий из Живого Журнала, касающихся предстоящих выборов в Государственную Думу 4 декабря.

В результате исследования было установлено, что подавляющее большинство упоминаний политического

процесса в России являются резко отрицательными. Причем уровень негатива и агрессии, выраженной в слове, зачастую очень высок, эмоционально ярко окрашен и представляет явную личную значимость для авторов. Личного позитивного отношения к политикам, за исключением единичных эпизодов, практически не было отмечено.

Спектр тем, обсуждавшихся в блогосфере в преддверии выборов, достаточно узок. Безусловным лидером по количеству и остроте дискуссий является тема нечестности предстоящих 4 декабря выборов. Интернет-аудитория задолго до проведения выборов сходится в том, что результаты выборов предрешены, выборы являются «фарсом», «спектаклем» и т.д. Общая доля упоминаний подобного рода в проанализированных постах превышает 90%. Лейтмотив – «Единая Россия» все равно победит, вне зависимости от реального положения дел. «Реальное» же положение дел в текстах живого Журнала – единица динамическая. В зависимости от автора поста, оно выглядит подчас диаметрально противоположно – либо «Единая Россия» предстает властителем дум львиной доли населения, полагающей наличествующими улучшения качества жизни и международного положения Российской Федерации, либо партия власти предстает агрессором, узурпировавшим власть, несмотря на то, что не имеет реальной поддержки народа.

В рассматриваемом в исследовании временном интервале (октябрь-ноябрь 2011 года) основными подтемами в рамках обозначенной доминанты стали:

- Обсуждения результатов интернет-голосования;
- Обсуждение работы сайта ktonarushil.ru;
- Обсуждение шантажа по отношению к ветеранам, учителям и другим зависимым от госбюджета гражданам членами Единой России;
- Призыв к различным формам протеста против «захвата власти»;
- Обсуждение невыполнения партией «Единая Россия» своих прошлых предвыборных обещаний;
- Обсуждение несовершенства системы власти и выборной системы России;
- Обсуждение иных нарушений предвыборного процесса, допущенных партией «Единая Россия».

На основании собранных данных была построена и проанализирована модель структуры выделенного сегмента блогосферы.

Для определения авторитетных пользователей была применена метрика входящих степеней. Если вершина обладает высокой входящей степенью, значит, записи данного пользователя собирают большое число комментариев, что означает высокую степень интереса к нему со стороны сообщества. Как правило, такие пользователи являются лидерами мнений и проводниками новых идей, которые вызывают активные дискуссии в обществе. Исследование показало, что распределение входящих степеней подчиняется степенному закону и резко убывает с ростом числа комментаторов.

На рисунке 4 приводится модель политически активного сегмента блогосферы с выделенными вершинами, соответствующими лидерам мнений:

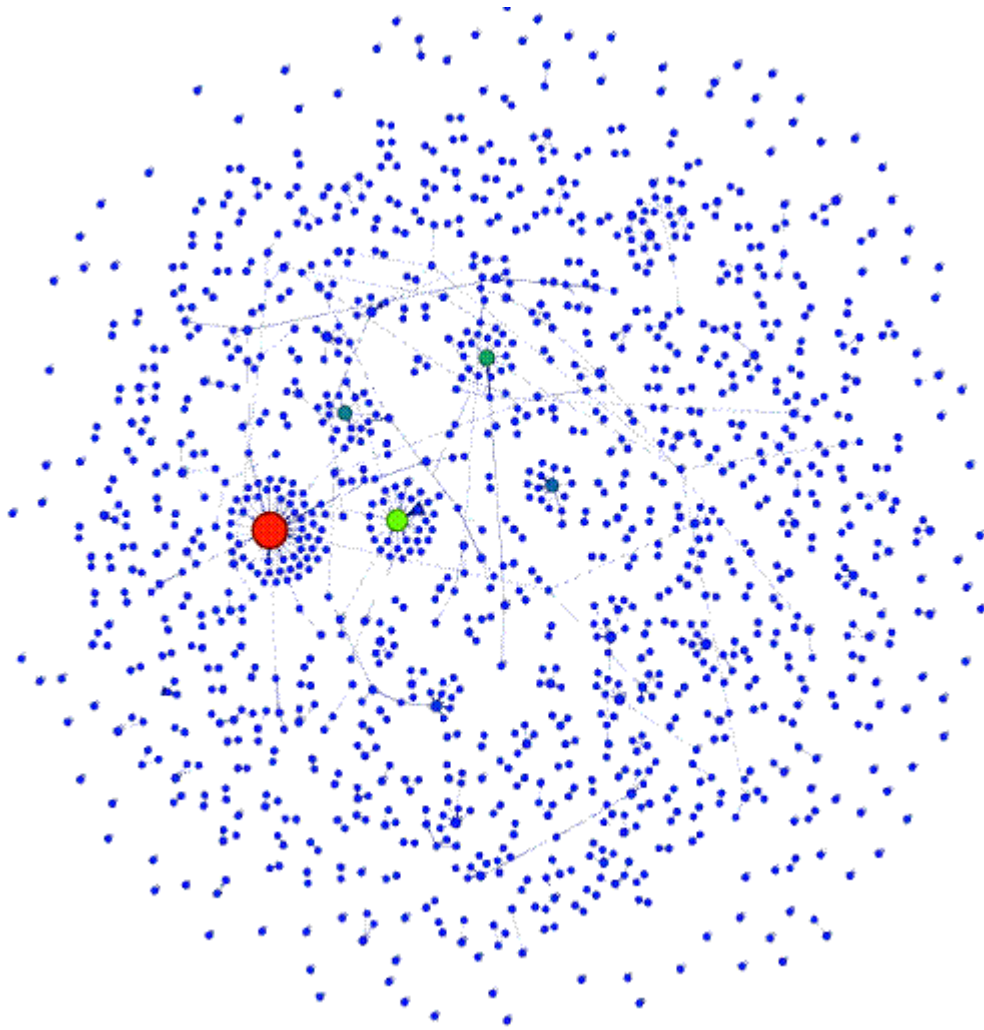


Рисунок 4 – Модель положения лидеров мнений в структуре выделенного тематического сегмента сети

Анализ выходящей степени в графе показывает, что, как правило, люди комментирующие лидеров мнений сами являются лидерами по количеству комментариев.

Взвешенная степень характеризует нормализованное распределение степеней в диапазоне от 1 до 100. Безоговорочными лидерами являются А. Навальный, Г. Явлинский, сообщество ru_politics. Также в списке присутствуют экономист Хазин и движение Солидарность. Интересным результатом стало то, что в списке нет таких политиков и деятелей как Г. Зюганов, В. Жириновский, М. Прохоров, что отчасти может объясняться тем, что основные дискуссии ведутся их сторонниками на других площадках, в частности, официальных сайтах.

Также интересным результатом можно считать то, что в списке нет региональных политических сообществ. Такие региональные сообщества, хотя и имеют читателей, не привлекают активных комментаторов. Как правило, в них содержатся только тексты и перепосты, но практически отсутствует политическая активность и дискуссии.

Нахождение проводников информации осуществляется путем определения промежуточности каждой из вершин графа. Проведенное исследование показало, что высокой степенью промежуточности обладает небольшое количество узлов – всего 0.5% из анализируемой выборки. Это означает, что в политическом сегменте Рунета не наблюдается сложной разветвленной сети со множеством больших кластеров и сообществ. Как правило, пользователи-проводники информации имеют возможность передавать информацию, общаясь одновременно в 2-4 различных кругах политических мнений.

Выводы

Реализация разработанной технологии мониторинга агитационных действий будет полезна на разных

этапах избирательного процесса – как во время избирательных кампаний, так и в периоды между ними. Знание основных тенденций в предпочтениях избирателей, возможность быстрого получения информации об изменениях в настроениях различных групп регионального электората, определение уровня и особенностей конкуренции в виртуальной среде, знание истории, продолжительности и качества деятельности всех кандидатов, их сторонников и оппонентов, возможность оценки влияния конкурентов в социальных сетях – все это является мощным инструментом планирования и осуществления поддержки избирательного процесса, и может существенно повысить как уровень избирательной активности, так и качество диалога между электоратом и властью. Также, практический аспект использования результатов исследования может выражаться в повышении узнаваемости кандидатов, увеличении лояльности Интернет-аудитории к конкретным кандидатам путем конструирования близкой ей политической позиции. Таким образом, в долгосрочной перспективе, использование результатов исследования может способствовать формированию атмосферы доверия целевой аудитории и Интернет-сообщества как к отдельным кандидатам, так и выборному процессу в целом.

Литература

1. М.Д. Розин, В.П. Свечкарев, С.Д. Конторович, С.В. Литвинов, В.И. Носко. Исследование социальных сетей как площадки социальной коммуникации рунета, используемой в целях предвыборной агитации // Инженерный вестник Дона, 2011. №1.
<http://www.ivdon.ru/magazine/archive/n1y2011/397> (доступ свободный) — Загл. с экрана. — Яз. рус.
2. Jason Falls. Why You Shouldn't Trust Automated Sentiment Scoring [Электронный ресурс] / 2011. Режим доступа: <http://www.socialmediaexplorer.com/social-media-monitoring/trusting-automated-sentiment-scoring> — Загл. с экрана. — Яз.англ.
3. Губанов Д.А. Социальные сети: модели информационного влияния, управления и противоборства [Текст] / Под ред. чл.-кор. РАН Д.А. Новикова / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили. – М.: Изд-во физ.-мат. лит., 2010. – 228 с.
4. Особенности чтения Живого журнала [Электронный ресурс] / Электронное периодическое издание «РИАН.Ру», 2011. Режим доступа: <http://ria.ru/analytics/20110725/407023471.html>, свободный. — Загл. с экрана. — Яз.рус.
5. Живой журнал [Электронный ресурс] / «Ведомости» Новости, аналитика, прогнозы. 2011. Режим доступа: www.vedomosti.ru/newspaper/article/269022/poluzhivoj_zhurnal, свободный. — Загл. с экрана. — Яз.рус.