# ДЕКОМПОЗИЦИЯ ИНФОРМАЦИОННОГО СОБЫТИЯ ЧЕРЕЗ ОПРЕДЕЛЕНИЕ ИНФОРМАЦИОННЫХ ВОЛН

**Градосельская Г.В., Волгин А.Д.**

*Национальный исследовательский университет «Высшая школа экономики»,*
*Россия, г. Москва, ул. Мясницкая, д. 20*
mss981009@mail.ru, art.volgin@gmail.com

*Аннотация: Неожиданно возникающее событие влияет на информационное пространство социальных сетей и через них влияет на социальные процессы в обществе. Эти явления необходимо фиксировать, исследовать и интерпретировать. Однако существующие подходы к сбору и анализу текстовых данных выносят фокусировку на социальных результатах за исследовательские рамки. В данной статье предполагается комплексный подход к рассмотрению информационного пространства как набора информационных событий. Чтобы понять природу информационного события, необходима его декомпозиция до уровня информационных волн. Эти концепты (информационное событие и информационные волны) позволяют выстроить содержательную методологию и получить выводы о природе потоков публикаций, следующих за событием. В качестве примера информационного события анализируется взрыв в Магнитогорске, произошедший 31 декабря 2018 года и имевший широкий общественный резонанс. В статье показывается возможность комбинирования социологической методологии в качестве содержательной аналитической рамки и математико-лингвистических процедур по выявлению текстовых кластеров.*

Ключевые слова: сетевой анализ, кластеризация текстов, информационное событие, информационные волны.

## 1 Проблема декомпозиции информационного события

Рассматривая массивы текстовой информации из социальных сетей, мы исходим из того, что большая часть публикаций является следствием или реакцией на некоторые события, происходящие в реальности или искусственно создаваемые. В данной статье сконцентрируемся на той части публикаций в социальных сетях, которые являются следствием реально произошедших событий. Произошедшее в реальности событие порождает массу публикаций в СМИ и социальных сетях. Роль СМИ не следует приуменьшать, поскольку значительная часть публикаций в социальных сетях является репостами публикаций СМИ, или откликом на них. Опыт наших прикладных исследований больших текстовых массивов по социально-политической тематике показывает, что доля индивидуальной реакции невелика и составляет от 20% до 10%.

Под информационным событием в контексте данного исследования, мы будем понимать опубликованные в СМИ и социальных сетях материалы, имеющие содержательное отношение к реальному событию.

Реальное событие и его информационный след не являются чем-то однородным, они распадаются на составные части: подтемы, аспекты рассмотрения события, его участников, динамики его развития, его социально-политических последствий. Каждая тема, аспект рассмотрения реального события, порождает корпус текстов, распределенных во времени.

Под информационными волнами в контексте данного исследования, мы понимаем ряд публикаций, распределенных во времени, затрагивающих одну подтему в рамках информационного события, имеющих высокую степень сходства по речевым маркерам.

На какие информационные волны распадается информационное событие, по каким законам они распространяются и какое социальное значение имеют – это интересные вопросы, на которые мы попробуем ответить в нашем исследовании.

Для примера рассмотрим взрыв в жилом доме, который случился в Магнитогорске 31 декабря 2018 года. Весь следующий информационный месяц был посвящен рассмотрению последствий этого события, рассмотрению других событий через его призму и т.п. Мы проведем последовательную декомпозицию этого события на составляющие информационные волны и рассмотрим социальные

аспекты распространения информации. Методологической основой рассмотрения информационного события будут являться информационные волны и разработанные в рамках текстового анализа алгоритмы их выявления.

## 2 Подходы к анализу текстов при исследовании информационных событий

### 2.1 Исследование событий в информационном пространстве

Вне зависимости от канала-коммуникации, любое СМИ в процессе деятельности всегда конструирует медийные факты, которые являются лишь субъективным отражением реальности [1]. В результате этого, в процессе прохождения информации через журналиста, редактора, политику новостного агентства реальное событие превращается в информационное событие, которое существует только в медиа-пространстве. Новостной текст, рассказывающий о случившимся, нельзя воспринимать как документальное описание реального событии: это всего лишь одна из моделей действительности, созданная в процессе работы журналиста. [2].

Для описания механизма распространения информационного события в социальных сетях мы задействуем понятие информационной волны. Информационная волна – это способ распространения манипулятивных моделей к определенным целевым аудиториям в информационном пространстве [3]. Информационная волна также может трактоваться как поток текстов-реакций на один информационный повод, в котором можно выделять пики и спады общественного интереса [4].

Информационные волны разделяются в зависимости от тональности сообщений (нейтральные, негативные и позитивные), по ключевому типу актора (генерируемые с помощью технических ботов, генерируемые с помощью информационных хабов и генерируемые силами реальных пользователей социальных сетей), по длине (короткие и длинные), по масштабу в зависимости от количества сообщений (малые и большие), по использованию информационных платформ (моноплатформенные и межплатформенные) и по динамике (нисходящие и флуктуационные) [3].

### 2.2 Методы кластеризации текстовых данных

Для проведения более формального, статического анализа задачу автоматического определения информационных волн можно представить, как задачу нахождения оптимального количества кластеров в массиве сообщений.

Проблема кластеризации, состоящая в обнаружении групп схожих объектов, широко исследовалась для решения широкого спектра задач в статическом анализе текста. Для текстовых данных характерен ряд отличительных признаков, которые важно учитывать при использовании методов кластеризации: (1) большая размерность в сочетании с разреженностью данных: в корпусе текста может содержаться порядка $10^5$ уникальных слов, однако отдельный документ может состоять всего лишь из нескольких слов или предложений (2) наличие корреляции слов друг с другом из-за большого корпуса документов, что требует тщательной настройки алгоритмов; (3) необходимость удаление стоп-слов и нормализации токенов [5].

Достаточно часто в задачах кластеризации текста исследователи используют классические статистические методы вроде агломеративной иерархической кластеризации или метода k-средних.

Применение иерархической кластеризации включает последовательное объединение документов на основе наилучшей попарной меры сходства между сгруппированными документами. Преимущество иерархической кластеризации заключается в возможности визуализации полученных результатов в виде дендограммы, которая может быть использована для поиска оптимального решения [6].

Метод k-средних – это итерационный метод разбиения на кластеры. Изначально каждому наблюдению присваивается наиболее близкий к ней случайно расположенный центроид кластера. После каждой итерации выполняется расчет среднего для наблюдений в каждом кластере, который заменяет центроид, определенный на предыдущей итерации, пока не будут выполнены некоторые критерии сходимости [6]. В большинстве случаев обычный метод k-средних дает лучший результат кластеризации, чем иерархический метод для текстовых данных.

Однако ни метод k-средних, ни агломеративная иерархическая кластеризация не подходят нам для решения поставленной задачи, поскольку необходимый алгоритм кластеризации, помимо всего прочего, должен соответствовать следующим критериям:

- Отсутствует необходимость в предварительном указании количества итоговых кластеров
- Может рассчитывать решение с большим количеством кластеров
- Эффективно работает с большим количеством данных

В настоящий момент существуют более продвинутые алгоритмы, которые подходят под все вышеуказанные требования. В нашей работе мы будем использовать три наиболее известных из них: DBSCAN, OPTICS и BIRCH.

DBSCAN (Density-based spatial clustering of applications with noise) – это метод кластеризации на основе плотности, который решает проблему кластеризации путем обнаружения областей высокой плотности в данных [7]. DBSCAN позволяет определять кластеры произвольной формы и повышает однородность внутри кластера путем фильтрации выбросов и шумов в данных.

Кроме того, существует популярное расширение DBSCAN – алгоритм OPTICS (Ordering points to identify the clustering structure), который позволяет рассчитывать несколько метрик расстояния одновременно: группы наблюдений, относящиеся к различным плотностям, выделяются параллельно. Этот метод не дает явной классификации наблюдений, а скорее генерирует упорядочение данных, отражающее кластерную структуру на основе плотности [8].

Также рассмотрим и алгоритм кластеризации под названием BIRCH (Balanced Iterative Reduction and Clustering using Hierarchies). В данном алгоритме используется метод сжатия, который рассчитывает ряд статистик на данных, и в дальнейшем использует его для кластеризации вместо исходных наблюдений. По сравнению с более простыми методами кластеризации BIRCH позволяет сократить время вычисления более чем на два порядка, при этом сохраняя точность на том же уровне [9]. Данный алгоритм, как и остальные методы, представленные выше, реализован в библиотеке scikit-learn на языке Python 3.7.4.

В роли эталона для сравнения качества работы предложенных алгоритмов будет использоваться экспертная кодировка, осуществляемая по принципу “от информационного повода к информационному событию”. Схожесть полученных кластерных решений с экспертной кодировка будет проверена с помощью ARI (Adjusted Rand Index) и F-меры.

## 3 Исследование информационного события на примере взрыва в Магнитогорске в 2018 г.

В качестве примера декомпозиции информационного события был выбран резонансный инцидент взрыва дома в Магнитогорске, случившийся 31 декабря 2018 года. Для получения доступа к медиа-сообщениям об этом событии мы воспользовались услугами компании по мониторингу и анализу СМИ “Медиалогия”.

Всего был скачано 50,547 статей и 397,107 сообщений о взрыве в Магнитогорске, которые были опубликованы на новостных сайтах и в различных социальных сетях (включая “ВКонтакте”, “Одноклассники” и “Facebook”) с 31 декабря 2018 года по 29 января 2019 года.

Релевантные сообщения об инциденте были отделены от общего информационного потока с помощью следующего поискового запроса: “Магнитогорск & (взрыв | катастрофа | обрушение | теракт | трагедия | спасатели | Путин | ИГИЛ | газ | Шахты | маршрутка | траур | переселение | расселение | завал | террорист | Газель | Фокин | эвакуация | Следственный комитет | ФСБ | кошмар | ужас | пранкер | соболезнования | пожертвования)”.

Для выявления информационных волн был использован метод определения сообществ FastGreedy [10] на матрице попарной схожести между сообщениями. В качестве меры схожести использовался коэффициент Жаккара.

Как уже опоминалось выше, возможны оба направления консолидации элементарных подволн: и от событий к информационным волнам, и от волн – к информационным событиям.

Всего после взрыва в Магнитогорске за следующий месяц было зафиксировано более 200 волн – по привязке к разным подтемам-событиям. Число репостов в каждой волне составляло от 20,000 до 300. Каждая из этих волн по смысловому сходству разбивается еще на несколько подволн (в среднем около 5 в каждой волне).

Можно привести и другой пример, когда определенные технически подволны объединялись по отнесению к одному событию. Например, в СМИ и социальных сетях широко освещалась история спасения мальчика Вани Фокина из-под завалов.

Условно назовем все это событие «Спасение 10-месячного ребенка из-под завалов», оно распадается на 68 подсобытий-подволн. Основной шлейф события длился 30 дней, которые освещались 23 подволнами.

Распределение информационных волн, описывающих одно событие («Спасение 10-месячного ребенка из-под завалов») по времени показаны ниже на рисунке 1.

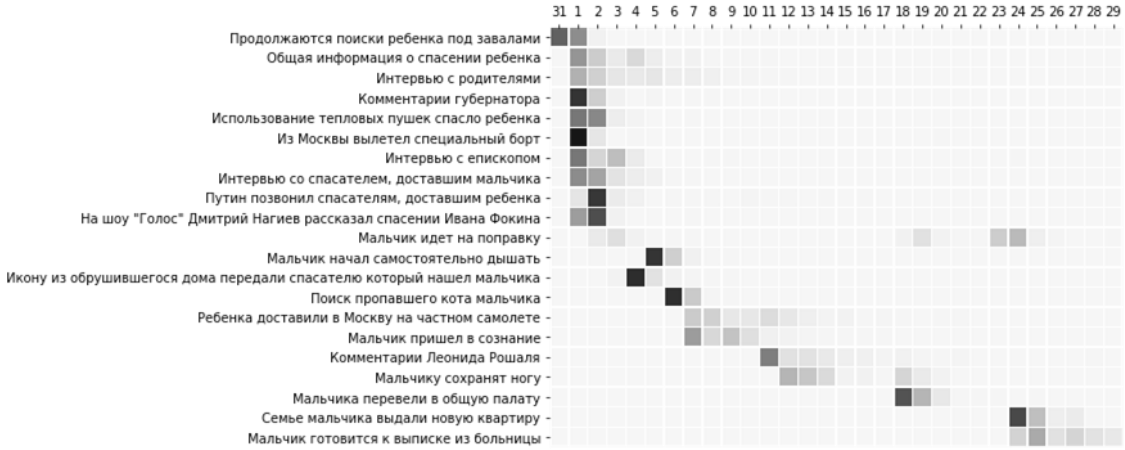


*Рис 1. Распределение информационного события о спасении ребенка по социальным сетям*

Как видим около 40% всех подволн приходится на 2-4 день события (не на первый). В первый день публикуется и распространяется основное фактологическое сообщение. За этот же день формируется реакция журналистов, и потенциальные подтемы развития этого события. Во второй день публиковались в основном рестроспективные расширения исходной темы (интервью с губернатором, родителями, спасателями). Потом происходит резкий спад интереса, и последовательно освещаются сопутствующие этапы развития события (перевозка в Москву, пребывание в детской больнице, комментарии врача и т.п.).

Ниже показано распределение подволн по событию по двум социальным сетям (Вконтакте и Одноклассники).

Обратим внимание, что сети фактически специализируются на разных темах. Во Вконтакте показаны более политические аспекты видения события (реакция Путина, Губернатора, справки о здоровье ребенка). Фактически, этот конкретный случай стал основой для политического позиционирования. В Однокласниках, наоборот, более широкое распространение получили семейно-психологические подтемы (новая квартира семьи, интервью с родителями, комментарии врача и т.п.).


*Рис 2. Распределение информационного события о спасении ребенка по социальным сетям*

## 4 Выводы и рекомендации

Практические результаты исследований распространяющейся текстовой информации показывают, что традиционные математико-лингвистические подходы к исследованию больших массивов текстов не дают желаемую глубину анализа.

Современная информация в своей основе производится и распространяется по весьма определенным социальным законам. Поэтому для понимания информационных процессов невозможно ограничиваться традиционным кластерно-лингвистическим анализом текстовых массивов.

Исследования по Магнитогорску и другие наши исследования текстовых массивов по разным темам показывают, что свыше 80% распространяемой информации являются информационными

волнами. Из этого следуют, как минимум, два вывода: во-первых, информационное пространство по большей части является манипулятивным; во-вторых, манипулятивная часть развивается и распространяется по определенным социальным правилам, которые необходимо исследовать.

В качестве резюме можно предположить, что число факторов влияния намного меньше, чем объем текстового массива, и их взаимосвязь зачастую может быть вовне массива – т.е. установить прямую причинно-следственную связь может быть затруднительно.

Можно рекомендовать при исследовании больших массивов текстов по социально-политической тематике и из социальных источников интегрировать социальные и математические подходы, разрабатывать новую методологию, направленную на понимание глубинных причинно-следственных процессов, происходящих в информационном пространстве.

## Литература


*1. Ермоленкина Л.И.* Информационное событие как репрезентация картины мира радиодискурса // Вестн. Том. гос. ун-та. Филология. 2010. №3 (11).

*2. Рабкина Н.В.* Медиасобытие в интернет-хронотопе: В. В. Путин и «Кольцо рыбака» // Вестник КемГУ. 2015. №2-3 (62).

*3. Градосельская Г.В., Щеглова Т.Е., Карпов И.А*. Информационные волны в социальных сетях: проблематизация, определение, механизмы распространения – М.: Системы высокой доступности, 2018. Т. 14. № 3. С. 87-91.

*4. Gradoselskaya G., Shcheglova T., Karpov I.* Information Waves on Social Networks: Problematization, Definition, Distribution Mechanisms, Eleventh International Conference "Management of large-scale system development" (MLSD-2018). M.: IEEE. 2018. – P. 1-4.

*5. Aggarwal C.C., Zhai C.* A Survey of Text Clustering Algorithms // Mining Text Data. 2012. – P.77-128.

*6. Alnajran N.* Cluster Analysis of Twitter Data: A Review of Algorithms // 9th International Conference on Agents and Artificial Intelligence. 2017 – P.239-249.

7. *Ester M., Kriegel H., Xi X.* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // Spatial, Text & Multimedia. 1996. – P.226-231.

*8. Ankerst M., Sander J.* OPTICS: ordering points to identify the clustering structure // ACM SIGMOD. 1999.

*9. Berkhin P.* A Survey of Clustering Data Mining Techniques // Grouping Multidimensional Data. 2006. – P.25-71.

*10. Newman M.E.* Fast algorithm for detecting community structure in networks // Physical Review E. 2004.