УДК 004.62

О ПРИМЕНИМОСТИ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДЛЯ БОРЬБЫ СО СПАМОМ В СОЦИАЛЬНЫХ СЕТЯХ

Ананченко И.В., Зудилова Т.В., Полин Я.А., Осетрова И.С.

Университет ИТМО, Санкт-Петербург, e-mail: anantchenko@yandex.ru, zudilova@ifmo.spb.ru, polin.ya@mail.ru, irina@ifmo.spb.ru

Статья посвящена рассмотрению возможности обработки спама и нежелательного контента в социальных сетях и других сервисах, где пользователи создают собственный контент (UGC – User-generated content), путем использования подхода кластеризации данных. Пользователи создают все больше контента, который нужно обрабатывать сервисам, в том числе и путем ручного модерирования. Одним из вариантов оптимизации модерирования, в том числе и ручного, является кластеризация. Задаче кластеризации данных посвящено множество научных работ. В работе осуществлена математическая постановка задачи кластеризации данных. Описан общий подход применения кластеризации данных. Произведен обзор существующих алгоритмов кластеризации: k-средних, k-медиан, expectation maximization, FOREL, c-средних, иеархической, минимально покрывающего дерева, послойный. Описаны преимущества и недостатки, а также входные данные и результаты данных алгоритмов. Рассмотрены возможные типы объектов в социальных сетях, которые можно кластеризовать для успешной борьбы с нежелательным контентом, а именно текстовая информация, изображения и учетные записи в сервисе. Приведены известные случаи применения подхода кластеризации социальных сетей и других UGC-сервисов для борьбы со спамом и нежелательным контентом.

Ключевые слова: социальные сети, спам, кластеризация, оптимизация, анти-спам системы

ON THE APPLICABILITY OF CLUSTERING ALGORITHMS FOR COMBATING SPAM IN SOCIAL NETWORKS

Ananchenko I.V., Zudilova T.V., Polin Ya.A, Osetrova I.S.

ITMO University, Saint-Petersburg, e-mail: anantchenko@yandex.ru, zudilova@ifmo.spb.ru, polin.ya@mail.ru, irina@ifmo.spb.ru

The article is devoted to the possibility of processing spam and unwanted content in social networks and other services where users create their own content (UGC – User-generated content) by using the data clustering approach. Users are creating more and more content that needs to be processed by services, including through manual moderation. One of the options for optimizing moderation, including manual moderation, is clustering. Many scientific papers are devoted to the problem of data clustering. The paper presents a mathematical formulation of the data clustering problem. A General approach to applying data clustering is described. The existing clustering algorithms are reviewed: k-means, k-medians, expectation maximization, FOREL, c-means, iearic, minimal covering tree, and layered. Advantages and disadvantages are described, as well as input data and results of these algorithms. Possible types of objects in social networks that can be clustered to successfully combat unwanted content, such as text information, images, and accounts in the service, are considered. There are well – known cases of using the clustering approach of social networks and other UGC services to combat spam and unwanted content.

Keywords: social networks, spam, clustering, optimization, anti-spam systems

Социальные сети стали неотъемлемой частью жизни многих людей. Аудитория многих социальных сетей превышает население целых стран (рисунок). Согласно исследованию, которое провело аналитическое агентство Statista, в апреле 2019 г. самой популярной социальной сетью стал Facebook с 2,32 миллиарда активных пользователей в месяц [1].

Каждая социальная сеть сталкивается со спамом и вынуждена всячески бороться с подобными злоупотреблениями. Объем контента, который генерируется пользователями, действительно очень большой. Например, количество сообщений, которыми каждый день обмениваются пользователи социальной сети «ВКонтакте», достигает 10 миллиардов [2]. Вместе с ростом количества активных пользователей, как правило, происходит и рост количества создаваемого контента.

Развитие таких технологий, как классификация, нейронные сети, во многих случаях позволяют достоверно определить спам, но остается множество подозрительного контента, решение по которому должен принять человек – модератор.

Задача фильтрации спама остается важной и актуальной на текущий момент, так как доля спама в почтовом трафике остается на высоком уровне, а количество пользователей социальных сетей, мессенджеров и других сервисов, где присутствует спам, продолжает расти. Пользователи в свою очередь создают все больше контента, который нужно обрабатывать сервисам, в том числе и путем ручного модерирования. В Facebook для этой цели в 2017 г. работало 4500 модераторов, а в 2018 г. их насчитывалось уже 7500 [3]. Важность отсутствия ошибок первого рода, то есть классификация хорошего объекта как плохого, не позволяет исполь-

зовать полностью автоматические системы, и часть объектов, оцененных как подозрительные, необходимо отправлять на прохождение ручного модерирования. Использование кластеризации позволит обрабатывать не каждый объект отдельно, а формировать кластеры, которые позволили бы принимать решение не по отдельно взятому документу или объекту, а по целому кластеру схожих по тем или иным параметрам документов или объектов. Такой подход позволит существенно сократить среднее время модерирования, а значит и повысить оперативность фильтрации спама.

Группировать можно не только такие объекты, как текст сообщения, изображения, но и пользователей, например на основе их действий и времени задержки между ними. Кластеризация позволяет оценивать группу объектов вместе, а не каждый объект в отдельности. В случае с ручным модерированием это позволяет существенно оптимизировать и повысить эффективность данного процесса. Группировать можно не только одинаковые объекты, но и похожие, что является как преимуществом, так и недостатком. С одной стороны, это позволяет объединить схожий спам в один кластер, но, с другой стороны, добавление к хорошему объекту элементов спама может привести к тому, что кластер будет состоять как из хороших объектов, так и из плохих.

Цель исследования: изучение возможности применения алгоритмов кластеризации для оптимизации противодействия спама в социальных сетях.

Материалы и методы исследования

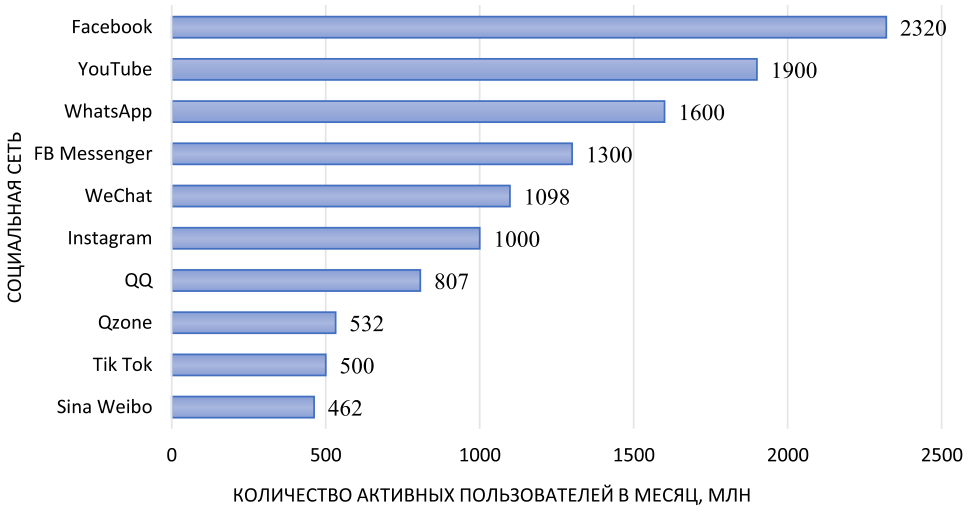
Для проведения исследования необходимо: осуществить математическую постановку задачи кластеризации, провести обзор существующих алгоритмов кластеризации, рассмотреть опыт применения кластеризации для борьбы со спамом и нежелательным контентом.

Математическая постановка задачи. Если необходимо разделить некоторую выборку на непересекающиеся друг с другом подмножества, именуемые «кластерами», то мы сталкиваемся с таким понятием, как формальная постановка задачи кластеризации. Математически ее можно описать следующим образом.

Допустим, что существует множества X и Y, где X состоит из объектов, а Y содержит номера, имена или метки кластеров. Функцию расстояния между объектами внутри X зададим как $p(x, x')$. Также существует выборка объектов, которая конечна и служит для обучения – $X^m = \{x_1,..., x_m\}$. Провести кластеризацию – значит разбить начальную выборку на непересекающиеся подмножества. Они будут называться кластерами и будут разделены таким образом, чтобы каждый из них содержал объекты, близкие по метрике p. При этом объекты различных кластеров должны значительно отличаться. Во время кластеризации каждому объекту x_i из обучающей выборки X^m присваивается определенный номер y_1, являющийся номером кластера.

Общий подход применения кластеризации включает этапы:

1. Выборка объектов для последующей кластеризации.
2. Формирование переменных, по которым будет производиться оценка объектов в отобранной выборке. При необходимости значения переменных нормализуют.
3. Производится вычисление мер сходства объектов.



Рейтинг социальных сетей по количеству активных пользователей в месяц

4. Применяется непосредственно кластерный анализ – формируются группы (кластеры) близких объектов.

5. Результаты анализа представляются в том или ином виде.

6. Корректируются отобранные метрики или метод кластерного анализа, чтобы получить оптимальный результат. Этап производится по необходимости.

Для оценки мер схожести объектов используют расстояние между ними. Расстояние между парами объектов можно измерять множеством метрик. Самые распространенные: евклидово расстояние, квадрат евклидова расстояния, манхэттенское расстояние, иначе называемое расстоянием городских кварталов, расстояние Чебышева и степенное расстояние.

Рассмотрим алгоритмы и методы формирования кластеров.

Метод k-средних. Алгоритм разбивает элементы множества на известное число кластеров, при этом минимизирует суммарное квадратичное отклонение точек кластеров от их центров. Каждое новое разделение на кластеры происходит с вычислением центра масс кластеров из предыдущей итерации. После чего происходит следующее разбиение на кластеры и так до тех пор, пока центр масс не изменяется.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2,$$

где k – число кластеров, S_i – полученные кластеры, $i = 1, 2, \ldots, k$ и μ_i – центры масс векторов $x_j \in S_i$.

Недостатками такого способа можно считать тот факт, что при таком разбиении на кластеры V – глобальный минимум суммарного квадратичного отклонения – не достигается. Достижимы только локальные минимумы. Кроме того, начальный выбор центров каждого из кластеров оказывает решающее влияние на конечный результат. Оптимального же метода выбора центров не существует. Еще одним недостатком такого подхода можно назвать необходимость знать заранее число кластеров для разбиения.

Метод k-медиан. В методе k-медиан для определения центра кластера вместо среднего значения вычисляется медиана множества. Задача разбиения кластеров по методу k-медиан заключается в нахождении центров кластеров так, чтобы полученные кластеры были компактны. Центры кластеров выбираются так, чтобы сумма расстояний между точками данных до центров кластеров была минимальна.

EM (expectation maximization). Существует алгоритм, перешедший в кластерный анализ из вероятностных моделей. Так называемый EM-алгоритм. Применяется тогда, когда модель зависит от скрытых переменных и находит оценки максимального правдоподобия параметров модели. Он состоит из итераций, каждый из которых производится в два шага – E-шаг и M-шаг. На первом шаге рассчитывается значение функции правдоподобия. Скрытые переменные при этом рассматриваются как наблюдаемые. На втором шаге находится оценка максимального правдоподобия. После чего это значение используется для первого шага на второй итерации.

Алгоритмы FOREL. Идея алгоритма заключается в объединении объектов в один кластер в зонах их сгущения. Разбиваются объекты таким образом, чтобы сумма расстояний от объектов кластеров до центров была минимальной по всем кластерам. На каждом шаге выбирается объект, раздувается вокруг него сфера, внутри которой выбирается центр тяжести, становящийся центром новой сферы. Так при каждой итерации центр смещается в сторону сгущения объектов. В тот момент, когда центр станет стабильным, объекты, находящиеся внутри полученной сферы, считаются кластеризованными и убираются из рассмотрения. Процесс повторяется, пока все объекты множества не будут кластеризованы. Минимизируемый алгоритмом функционал качества:

$$F = \sum_{j=1}^{k} \sum_{x \in K_j} p(x, W_j).$$

Нечеткая кластеризация C-средних. Подход кластеризации по c-средним позволяет разбивать множество на заданное количество нечетких множеств. В методе используется нечеткая матрица принадлежности с некоторыми элементами, которые определяют принадлежность определенного элемента начального множества конкретному кластеру.

Иерархическая кластеризация. Кластеризация с помощью иерархических алгоритмов позволяет использовать теорию графов для разделения множества объектов. В результате такого разбиения получается иерархия, иначе называемая дендрограммой.

Существует несколько подходов к иерархическому разбиению. В основном они характеризуются направленностью роста иерархий. Различают одиночную и полную связь. Одиночная связь (single-link), или метод ближайшего соседа: каждый шаг объединяет два кластера с самым маленьким расстоянием между двумя объектами. Полная связь (complete-link) иначе называется

методом дальнего соседа. Характеризуется объединением двух расположенных максимально далеко друг от друга представителей. Различают еще метод средней связи (middle-link), являющийся комбинацией двух предыдущих подходов.

Алгоритм минимального покрывающего дерева. При реализации алгоритма минимального покрывающего дерева все экземпляры сначала помещаются в общий кластер, после чего происходит несколько итераций кластеризации. На каждой из них кластеры разделяются на два таким образом, чтобы расстояние между ними было наибольшим. Алгоритм наиболее подходит для выделения кластеров типа сгущений или лент. Другим недостатком данного алгоритма является высокая трудоемкость.

Послойная кластеризация. Еще одним примером графовой кластеризации является послойная кластеризация. В этом методе выделяются связанные компоненты. Из начальных данных формируется поочередно несколько подграфов, отражающих структурную иерархию связей. Связанные компоненты графа определяются порогом расстояния, которое как раз формирует новые кластеры. Изменяя уровень расстояний, можно варьировать степень глубины иерархии новых кластеров. Итоговые результаты сравнения алгоритмов представлены в таблице.

Рассмотрим, какие объекты в социальных сетях можно кластеризовать для успешной борьбы со спамом.

– Текстовые сообщения. Это наиболее очевидный объект, который можно кластеризовать, так как в социальных сетях текстовые сообщения используются в личных сообщениях, статусах, групповых публикациях, информации об аккаунте и т.д. Для кластеризации текстовых документов часто используют алгоритм k-means [4].

– Изображения. Со временем спамеры стали использовать «графический» спам, что значительно затрудняло анализ, но достаточно быстро появились ответные меры для борьбы с этим, в том числе распознавание текста с изображений [5]. Кластеризовать можно как сами изображения, группируя как полностью одинаковые изображения, так и схожие. Кроме того, при комбинировании кластеризации текста с распознаванием текста на изображениях становится возможным кластеризовать разные изображения на основе одинакового или схожего текста на них.

– Пользователи. Это ключевой элемент в любой спам-рассылке, так как без аккаунтов рассылка становится невозможной. Выделяют различные группы аккаунтов, которые могут быть использованы для рассылок спама. Сервисы обладают очень обширной информацией о существующих аккаунтах, что позволяет выделить множество атрибутов для последующей кластеризации пользователей.

Для кластеризации спама в социальных сетях подходит множество алгоритмов. Выбор конкретного алгоритма может зависеть от используемого стека технологий в конкретной социальной сети. Кроме того, наиболее важным для качественной кластеризации является подбор входных данных, то есть атрибутов, которые и будет необходимо кластеризовать, особенно это касается задачи кластеризации пользователей.

Другим не менее важным фактором является производительность алгоритма. Это действительно является важным параметром, так как объем контента и данных в социальных сетях очень большой.

Сравнительная таблица алгоритмов кластеризации

Алгоритм кластеризации	Форма кластеров	Входные данные	Результаты
k-средних	Гиперсфера	Число кластеров	Центры кластеров
k-медиан	Гиперсфера	Число кластеров	Центры кластеров
EM	Гиперсфера	Число кластеров	Матрицы, содержащие обновляемые параметры смеси
FOREL	Гиперсфера	Радиус поиска локальных сгущений	Центры кластеров
C-средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
Минимально покрывающего дерева	Произвольная	Число кластеров или порог расстояния для удаления ребер	Древовидная структура кластеров
Послойный	Произвольная	Последовательность порогов расстояния	Древовидная структура кластеров с разными уровнями иерархии

Рассмотрим опыт применения кластеризации данных для борьбы со спамом и другим нежелательным контентом. Социальная сеть Badoo рассказала об использовании кластеризации текстовых сообщений с использование k-means [6]. Facebook для борьбы с дезинформацией, так называемыми фейковыми новостями, использует в том числе и кластеризацию [7]. Linkedin использует кластеризацию для задачи предотвращения злоупотреблений [8]. «Юла» проводит кластеризацию по каждой сущности объявления, чтобы найти дубликаты [9].

Заключение

В данной работе были описаны различные алгоритмы кластеризации данных. В частности, рассмотрены возможности применения алгоритмов кластеризации данных для борьбы со спамом в социальных сетях. В результате проведенного исследования удалось определить, что алгоритмы кластеризации данных являются мощными помощниками при проведении анализа полученных наблюдений и характеристик в борьбе со спамом и другим нежелательным контентом в различных сервисах.

Список литературы

1. Global social media ranking 2019/ Statistic: Statista – The Statistics Portal for Market Data, Market Research and Market Studies [Электронный ресурс]. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (дата обращения: 15.02.2020).

2. Соцсеть ВКонтакте достигла рекордных 10 миллиардов сообщений в сутки: АиФ Санкт-Петербург. [Электронный ресурс]. URL: https://spb.aif.ru/society/people/socset_vkontakte_dostigla_rekordnyh_10_milliardov_soobshcheniy_v_sutki (дата обращения: 15.02.2020).

3. Количество модераторов Facebook превышает размер всего штата Twitter: Rusbase. Медиа, которое решает задачи предпринимателей. [Электронный ресурс]. URL: https://rb.ru/story/facebook-moderation/ (дата обращения: 21.02.2020).

4. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. 2017. Т. 29. № 2. С. 161–200.

5. Эволюция спама: Энциклопедия «Касперского» [Электронный ресурс]. URL: https://encyclopedia.kaspersky.ru/knowledge/the-evolution-of-spam/ (дата обращения: 21.02.2020).

6. Вычисляем по IP: как бороться со спамом в социальной сети: Хабр. [Электронный ресурс]. URL: https://habr.com/ru/company/oleg-bunin/blog/426843/ (дата обращения: 22.02.2020).

7. FightingAbuse @Scale – MichaelMcNallyandLauren-Bose: @Scale. [Электронный ресурс]. URL: https://atscale-conference.com/videos/fighting-abuse-scale-michael-mcnally-and-lauren-bose/ (дата обращения: 23.02.2020).

8. Fighting Abuse @Scale 2019: Preventing Abuse Using Unsupervised Learning: @Scale [Электронный ресурс]. URL: https://atscaleconference.com/videos/fighting-abuse-scale-2019-preventing-abuse-using-unsupervised-learning/ (дата обращения: 23.02.2020).

9. Как мы модерируем объявления / Блог компании Юла: Хабр. [Электронный ресурс]. URL: https://habr.com/ru/company/youla/blog/455128/ (дата обращения: 24.02.2020).