УДК 004.67
DOI: 10.46548/21vek-2020-0951-0005

ИССЛЕДОВАНИЕ ГРУПП ПОЛЬЗОВАТЕЛЕЙ В СОЦИАЛЬНЫХ СЕТЯХ ПО ИХ ИНТЕРЕСАМ И ПОВЕДЕНИЮ НА ОСНОВЕ МНОЖЕСТВА ИСТОЧНИКОВ ДАННЫХ

Мартышкин Алексей Иванович, кандидат технических наук, доцент,
доцент кафедры «Вычислительные машины и системы»
Пензенский государственный технологический университет
(440039, Россия, Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11, e-mail: alexey314@yandex.ru)
Перекусихина Альбина Николаевна, кандидат технических наук,
доцент кафедры «Математика и математическое моделирование»
Пензенский государственный университет архитектуры и строительства
(440028, Россия, Пенза, ул. Германа Титова, д. 28)
Зоткина Алена Александровна, аспирант кафедры
«Вычислительные машины и системы»
Пензенский государственный технологический университет
(440039, Россия, Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11, e-mail: alena.zotkina.97@mail.ru)

Аннотация. В статье предложен подход к автоматизированному анализу данных социальных сетей, позволяющий определять, принадлежат ли публичные сообщества и пользователи к подгруппе групп пользователей интернет-ресурсов, имеющих социальную составляющую. Рассмотрена задача восстановления данных пользователей и подзадача выделения групп пользователей, а также описаны методы решений, основанные на различных видах данных и различных моделях. Описаны некоторые алгоритмы и методики, используемые при анализе текста, а также при решении задач структурного машинного обучения. Приведенные методики и алгоритмы использовались при решении поставленной в статье задачи. После описания процесса решения задачи показан проведенный эксперимент для апробации предлагаемого подхода. На примере задачи определения подгрупп радикальных футбольных фанатов из группы футбольных болельщиков показана состоятельность предлагаемого подхода. Результаты исследования показали, что тривиальная оценка схожести может работать на определенных данных, однако имеет большую сложность в связи с необходимостью составления правильного словаря для подгруппы. При неточном его составлении, качестве результатов значительно падает. Исследуемая модель показывает рост точности с использованием большего числа шагов, а значит большего числа узлов графа. Использование латенто-семантического анализа дало возможность улучшить результаты. Предложенный подход с использованием оценки схожести, основанной на операторе «ИЛИ» дал много ложноположительных определений членов подгруппы, однако показал абсолютную полноту. При использовании оценки схожести, основанной на линейной комбинации, получается максимальная точность. Предложенный подход позволяет при наличии малой обучающей выборки получать приемлемый результат. В заключение приведены основные выводы по проделанной работе.

Ключевые слова: социальная сеть, сообщество, социальный профиль пользователя, метод решения, обуча-ющая выборка, граф, источник данных, марковский процесс.

RESEARCH GROUPS OF USERS IN SOCIAL NETWORKS BASED ON THEIR INTERESTS AND BEHAVIOR BASED ON A VARIETY OF DATA SOURCES

Martyshkin Alexey Ivanovich, candidate of technical sciences, docent,
associate Professor of sub-department «Computers and systems»
Penza state technological University
(440039, Russia, Penza, Baydukov Proyezd / Gagarin Street, 1a/11, e-mail: alexey314@yandex.ru)
Perekushikhina Albina Nikolaevna, candidate of technical sciences,
associate Professor of sub-department «Mathematics and Mathematical Modeling»
Penza State University of Architecture and Construction
(440028, Russia, Penza, German Titov Street, 28)
Zotkina Alena Aleksandrovna, postgraduate of sub-department «Computers and systems»
Penza state technological University
(440039, Russia, Penza, Baydukov Proyezd / Gagarin Street, 1a/11, e-mail: alena.zotkina.97@mail.ru)

Abstract. The article offers an approach to automated analysis of social network data that allows determining whether public communities and users belong to a subgroup of groups of users of Internet resources that have a social component. The problem of user data recovery and the subproblem of user group allocation are considered, and the methods of solutions based on different types of data and different models are described. Describes some of the algorithms and techniques used in the analysis of the text, and also when solving problems of structural machine learning. These methods and algorithms were used to solve the problem set in the article. After describing the process of solving the problem, an

experiment is shown to test the proposed approach. The consistency of the proposed approach is shown by the example of the problem of determining subgroups of radical football fans from a group of football fans. The results of the study showed that a trivial similarity assessment can work on certain data, but it is more difficult due to the need to compile the correct dictionary for a subgroup. If it is not accurately compiled, the quality of the results significantly decreases. The model under study shows an increase in accuracy using a larger number of steps, which means a larger number of graph nodes. The use of latent semantic analysis made it possible to improve the results. The proposed approach using similarity estimation based on the "OR" operator gave many false-positive definitions of subgroup members, but showed absolute completeness. When using a similarity score based on a linear combination, maximum accuracy is obtained. The proposed approach allows you to get an acceptable result if you have a small training sample. In conclusion, the main conclusions on the work done are presented.

Keywords: social network, community, user's social profile, solution methods, training sample, graph, data sources, Markov process.

Введение. В последнее время социальные медиа набрали огромную популярность. Такие социальные сети, как *Facebook.com, Vk.com, Twitter.com* обладают огромной аудиторией. В совокупности размер аудитории существующих социальных сетей составляет более двух миллиардов пользователей, и постоянно растет. Данные, порождаемые пользователями, могут быть использованы для определения интересов, предпочтений и иных личных свойств пользователя. Часть подобной информации, пол, возраст, местоположение, увлечения, может быть указана в профиле пользователя. Получение данных о пользователе может быть полезно как бизнесу, так и государству [1]. Имея дополнительные данные об увлечениях людей, можно определять потенциальных преступников, что позволит предотвращать нарушения или прогнозировать конфликты [2]. Существуют исследования о восстановлении данных, явно неуказанных в профилях пользователей [3 – 6]. Для решения этой задачи используют данные о социальных связях пользователей. Показано, что они влияют на поведение человека, на его взгляды [7, 8]. В статье предложен подход для выделения подгруппы пользователей из определенной группы.

Материалы исследования. Зачастую из-за необязательности заполнения некоторых данных возникает проблема восстановления характеристик пользователей. Бывает, что необязательно заполнять пол, возраст, физические данные, тогда как они могут быть очень важны для выделения и оценки определенного рода ресурсов [4 – 6]. Существуют исследования определяющие психотип пользователя, используя лишь данные о них из их же аккаунтов с социальных медиа [4]. Определение психотипа, хронотипа – задача, сводящаяся к выделению группы пользователей, ярким примером которой может служить проблема определения принадлежности пользователя к политическому движению [9 – 13]. Одним из популярнейших решений является подход, использующий известные признаки пользователей, взятые из профилей. Так, например, опишем подход из [14]: для каждого пользователя собирались все публичные характеристики его профиля, такие как, пол, возраст и т.д. Наглядным примером является задача определению пола по имени [15]. Другим методом приведения текста к численными данными является латентный семантический анализ [16]. Важной группой данных при восстановлении характеристик пользователя являются медиа данные, такие, как фотографии, видеозаписи, музыка. В качестве примера использования фотографий рассмотрим исследование [17], определяющее гендерную принадлежность пользователя используя яркость фотографий. Минус такого подхода заключается в возможности изменения метаинформации на не соответствующую действительности. Существует множество исследований, использующих в качестве основы своей модели информацию о музыке пользователя [18, 19], отрицательным качеством которых является слабая точность результатов без дополнительных параметров. Иногда пользователи оставляют информацию о своих аккаунтах на других сайтах. Использование подобной информации требует решения дополнительной проблемы, определения правдивости принадлежности аккаунтов одному человеку [20].

Результаты исследования. Дадим формальную постановку задачи. Пусть G, $G_{sub}:G_{sub} \subset G$, $f(g_i, g_j) = [0, 1]$, где 1 обозначает связь между элементами $g_i, g_j \in G$. Тогда, зная, что $g_1, g_2 g_3 .. g_k \in G_{sub}$, нужно уметь определять входит ли $g \in G$ в G_{sub}. Имеющиеся данные можно представить в виде смешанного графа. Узел – либо пользователь, либо публичная страница, с сопутствующей информацией, а ребра между узлами – отношение подписка-подписчик. Сопутствующая информация группы есть ее публичные сообщения и список пользователей, одобривших сообщения.В этой задаче имеем три типа данных: информацию о социальных связях, тестовую информацию, а также отношение пользователя к тестовым сообщениям.

Важной частью исследования является анализ сообщений, оставленных в публичных сообществах. Приведем понятия, использующиеся в дальнейшем описании [21]. Термин (слово) – атомарная лингвистическая единица. Документ – конечный набор терминов. В контексте настоящей статьи, документ – публичное текстовое сообщение. Коллекция – набор документов $D = \{D_1, D_2, ..., D_n\}$, где D – коллекция, а D_i – документ. Словарь – набор всех терминов, встречающихся во всех коллекциях.

$$T = \left\{ t : t \in \bigcup_{j=1}^{n} D_j \right\} = \{t_i\}_{i=1}^{m} . \quad (1)$$

Для описания подхода удобно использовать дву-

мерную матрицу, в которой каждый столбец будет представлять вектор, соответствующий документу. Для вычисления значений элементов матрицы как правило используют формулу *TF-IDF*, она имеет следующий вид

$$d_{ij} = tf_{ij} \cdot \log \frac{n}{df_i}, \quad (2)$$

где tf_{ij} – число встреч термина i в документе j,

df_i – число документов, в которых встречается термин i, а gf_i – число встреч термина i во всей коллекции.

Выражение (2) может меняться в зависимости от исследования. Кроме того, могут применяться и иные подходы, нов рамках настоящего исследования они не применяются.

Очевидным решением задачи определения принадлежности документа к определенной тематике является подсчет вхождения терминов из предварительно составленного словаря, вмещающего в себя термины искомой тематики. Таким образом, при использовании данного метода коллекция хранится как таблица, состоящая из строк – документов, столбцов – терминах составленного словаря и ячейки содержащей, информацию, обозначающую принадлежность соответствующего термина в соответствующий документ.

Латентно-семантический анализ – метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами в них встречающимися, сопоставляющий тематики всем документам и терминам. Таким образом автоматически решают задачу определения тематики терминов.

Пусть $D \in R^{m \times n}$ – матрица «термин-документ», вычисленная каким-либо образом. Требуется выполнить разложение данной матрицы. $D=U \cdot V^T$, $U \in R^{m \times k}, V \in R^{n \times k}$, где U – матрица «термин-тема», V – матрица «документ-тема», а k – число тем. Строка матрицы U под номером i характеризует «степень принадлежности» термина i каждой из тем. Строка матрицы V под номером j обозначает «степень принадлежности» документа j каждой из тем. Фактически данный метод можно рассматривать как нечеткую кластеризацию. Латентно-семантический анализ позволяет уменьшить набор терминов, что существенно облегчает задачу. Существуют два подвида данной задачи, один использует вероятностную модель данных, в ячейках матрицы хранятся вероятности, другие использую особые метрики. Формально вероятностную модель данных можно описать так.

$$p(d, w) = \sum_{t \in T} p(t)\, p(w \mid t)\, p(d \mid t), \quad (3)$$

где T – множество тем, $p(d, w)$ – вероятность возникновения термина w в документе d, $p(t)$ – вероятность выбрать тему t, $p(w|t)$ – вероятность выбора термина w из темы t, а $p(d|t)$ – вероятность выбора документа d, при условии, что выбрана тема t.

К вероятностным алгоритмам латентно-семантического анализа относят *PLSA* [22] и *LDA* [23]. Из не вероятностных алгоритмов отметим *LSI* [24]. К минусам модели можно отнести сложность при интерпретации данных.

Случайные марковские поля [25] – метод, широко применяемый в различных областях искусственного интеллекта. Его успешно используют при распознавании речи и образов, а также в обработке текста [26]. Марковским случайным полем (марковской сетью) называют графовую модель, которая используется для представления совместных распределений набора нескольких случайных переменных. Считаем, что если вершины не смежны, то они являются условно независимыми случайными величинами. Совместное распределение набора случайных величин в марковском случайном поле вычисляется по выражению

$$P(x) = 1 / z \prod_{t \in T} p(X_D), \quad (4)$$

где $p(X_D)$ – потенциальная функция, описывающая состояние случайных величин в k клике; z – коэффициент нормализации, $z = \sum_{x \in X} \prod_k p_k(X_k)$.

Одной из разновидностей метода случайных марковских полей является метод скрытых марковских полей (*CRF*) [27, 28]. У метода есть недостатки, такие как вычислительная сложность анализа обучающей выборки, это затрудняет обновление модели с обновлением обучающих данных. Поэтому было решено опробовать собственный упрощенный метод. Вернемся к представлению данных в виде графа. Введем для каждого узла характеристику p – вероятность отнесения его к определенной подгруппе. Пусть множество узлов M – множество узлов с приписанной вручную p. Далее для каждого смежного узла x вычисляется значение его p.

$$p = F_{h \in H} k \cdot h_x, \quad (5)$$

где H – множество признаков, таких как, например, текстовое сходство, поведенческое сходство и так далее, F – функция, считающая суммарный вклад признаков, а k – коэффициент, определяющий важность параметра.

В результате получаем множество размеченных узлов M_1. Можно обучиться на выборке данных, чтобы понять какой из параметров наиболее влиятельный и представлять F в виде линейной комбинации. Этот процесс повторяется в рамках множества M_1. Пересчет предлагается остановить, когда норма Фробениуса станет меньше либо равна E [29]. Важной оценкой качества такого подхода является проверка изменений p размеченных узлов. Далее алгоритм повторяет последовательность действий, до состояния полного покрытия сети. Этот процесс крайне ресурсоемкий, поэтому в рамках исследования использованы меньшие объемы данных. Рассматривался рост в три шага.

На рисунке 1 показана общая схема предлагаемого подхода.

Опишем последовательно схему представляемого подхода. Из поставленной задачи имеем проблему

определения группы пользователей. На первом этапе, исходя из тематики группы, необходимо собрать набор публичных страниц, придерживающихся данной тематики. Для этого предлагается сделать аналог лингвистической экспертизы. Имея набор групп определенной тематики, мы собираем всю информацию, связанную с этим группами: тексты подписчики и т.д. Собрав все необходимые данные, строим по ним граф социальных связей. Преобразуем его к модификации случайного марковского поля, используя дополнительную информацию. Узлом графа может являться публичная страница или же сам пользователь. Ребро обозначает подписку на публичную страницу или двусторонние отношения, определяемые как взаимная подписка или отношение – дружба.

Рисунок 1 – Схема решения задачи выделения подгруппы пользователей

На рисунке 2 показана схема графа социальных связей.

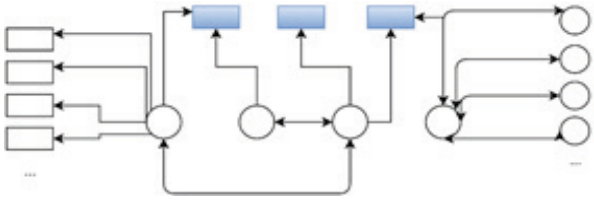

Рисунок 2 – Граф социальных связей

На рисунке 2 приняты следующие обозначения: – публичная страница; – пользователь; – публичная страница, отнесенная к тематике.

Сбор данных. В эксперименте решается задача для группы футбольных болельщиков и подгруппы радикальных фанатов. В рамках исследования используются данные из социальной сети *Вконтакте*. Пользователи имеют связи одновременно с группами и другими пользователями, группы же связаны только с пользователями. Вначале необходимо создать список публичных страниц тематики искомой группы. Для группы футбольных болельщиков: группа должна быть посвящена определенному футбольному клубу, это определялась по текстовым сообщениям, если в них присутствовали новости о футбольной команде, страница входит в группу. Для исследования взята группа страниц, посвященных футбольному клубу "*Зенит*". Для определения публичных страниц, входящих в подгруппу радикальных фанатов, из собранных групп выбирались те, которые содержали негативные отзывы о командах соперников, ненормативные высказывания в адрес болельщиков других команд. Всего собрано 211 групп посвященных этой тематике, 10 из которых были о футбольных хулиганах. Выбирались примерно одинаковые по числу подписчиков сообщества, размер которых не превышал 5000 пользователей и был выше 500. Всего собрано 225230 пользователей, 74% из них оказались мужчинами. Максимальное число подписок из перечня сообществ составляло 7. Сообщества сильно разнились по числу публичных сообщений. Суммарно собрано 178022 публичных сообщений. Число лайков и репостов – 7199. В ходе сбора подписчиков и подписок с только что добавленных узлов, граф очень быстро растет. В рамках исследования не проводились эксперименты с большим числом обновлений выборки. Пользователи добавлялись не более 2 раз, группы – не более 3.

Использование алгоритмов. Условия задачи ставят определенные ограничения на используемые методы. Так имеется крайне ограниченная выборка, состоящая из небольшого набора публичных страниц. Поэтому предлагается воспользоваться подходом, схожим с предложенным в [25]. Так можно значительно увеличить объем данных, всегда можно оценить качество результатов. Для эксперимента было выбрано две модификации описанного подхода. Множества признаков H будет одинаковым для двух модификаций, однако, оно будет отличаться от типа узла. Для узла пользователя определим два признака: наличие одобрения содержимого из смежных публичных страниц, принадлежащих к группе, и влияние характеристик смежных узлов. Для групп же – текстовая схожесть с текстами из подгруппы и влияние характеристик смежных узлов.

Использование методов, основанных на данных смежных узлов, обусловлено предположением, что пользователь, имеющий больше социальных связей с членами подгруппы, с большей вероятностью сам будет принадлежать к этой подгруппе. Использование признака одобрения контента основывается на предположении, что пользователь, одобривший что-то, действительно склонен одобрять данного рода информацию. Модификации будут отличаться вычислением влияния смежных узлов. В первом случае *F* будет рассчитываться так:

$$F_{user}(x) = isApproved(x) or \sum_{y \in M_{adjacents}} p_y / n_{notnullable} > 0,5 \quad (6),$$

$$F_{publicpage}(x) = textSimilarity(x) or \sum_{y \in M_{adjacents}} p_y / n_{notnullable} > 0,5 \quad (7).$$

Во втором – как линейная комбинация:

$$F_{group}(x) = (k_1 \cdot isApproved(x) + k_2 \cdot \sum_{y \in M_{adjacents}} p_y / n_{notnullable}) / (k_1 + k_2) \quad (8),$$

$$F_{group}(x) = (k_3 \cdot textSimilarity(x) + k_4 \cdot \sum_{y \in M_{adjacents}} p_y / n_{notnullable}) / (k_3 + k_4) \quad (9).$$

Процесс пересчета стоит продолжать до некоего предела E. В нашем случае было принято E = 0,2%. При использовании упрощенной модели вычисляется норма Фробениуса первого рода, которая получает максимальную разность в матрицах. В результате такого подхода вероятности для групп могут измениться относительно изначальных. Путем сравнения с изначальными данными можно проверить качество моде-

ли. На каждом шагу увеличения выборки проверяем первоначальные данные. Далее следует обход графа социальных связей. На каждом этапе добавляются новые подписчики только что добавленных узлов графа и подписки, если они у узла есть.

Результаты. Важным критерием оценки является корректное определение групп из обучающей выборки после нескольких этапов добавления новых узлов в граф социальных связей. Также интересен результат для пользователей, определенных как администраторы сообществ. До начала эксперимента считается, что они будут принадлежать к той же группе, что и сообщество. Для оценки основным был выбран метод кросс-валидации. Выбранные группы делились на 10 частей, 9 групп входили в обучающую выборку. Оставшаяся часть использовалась для тестирования. Процедура повторялась 10 раз, для каждой части. Алгоритм работает в несколько шагов, так что целесообразно проверять качество работы на разных этапах.

Оценка результатов. Приведем результаты применения различных вариаций алгоритмов на группе футбольных болельщиков и радикальных футбольных фанатов. Т.к. в ходе исследования не было обнаружено аналогично поставленных задач, результаты подхода сравним с результатами тривиальных алгоритмов.

Результаты тривиальной оценки схожести текстов. Данный метод казался перспективным ввиду того, что подгруппы пользователей, включенные в эксперимент, обладали собственным сленгом. Однако оказалось, что применение сленга довольно популярно и термины из словаря сленговых выражений встречались, как и в подгруппе, так и в группе, что не позволило использовать данный метод. Средняя оценка точности не превышала 60%. Поэтому подробное рассмотрение результатов этого метода нецелесообразно. Однако, используя латенто-семантический анализ, удалось добиться хорошего разделения текста по тематикам. На рисунке 3 серыми кругами обозначены термины из группы обычных фанатов, звездочками – из радикальных, черными квадратами – фанатские группы, белыми кругами – радикальные фанатские группы. На рисунке 3 приведены данные для шести радикальных групп и шести нерадикальных.

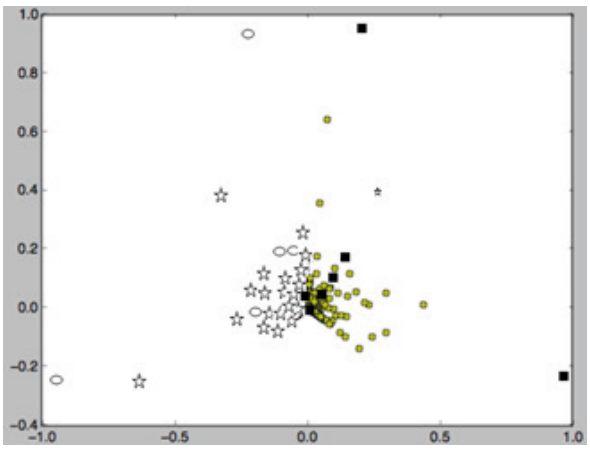

Рисунок 3 – Разделение на тематики

Тривиальные подходы. Подход всегда определяющий узел как нерадикальный. При использовании подобного подхода, точность вычислений становится 95%. Подход, определяющий узел как радикальный с вероятностью 10 к 211, обладает точностью 91%, а подход, определяющий узел как радикальный с вероятностью 50% –точностью 50%.

Метод оценки схожести, основанный на операторе «ИЛИ». Кросс-валидация показала результаты средней точности 63%. Важным критерием является возможность определять членов подгруппы. Т.к. за счет несбалансированности размеров классов, алгоритм, возвращающий наиболее часто встречающийся результат, будет давать точность лучшую с увеличением класса.

Заключение. В статье продемонстрирован подход, позволяющий определять, принадлежат ли публичные сообщества и пользователи к подгруппе групп пользователей интернет-ресурсов, имеющих социальную составляющую. В ходе эксперимента выделялась подгруппа радикальных фанатов из группы болельщиков футбольного клуба «*Зенит*». Результаты показали, что наилучшего результата удалось добиться с использованием подхода, использующего латентно-семантический анализ и линейную комбинацию признаков.

Стоит отметить, что предложенный подход выделения подгруппы пользователей может использовать не представленные в исследовании признаки схожести. Кроме текстовых данных могут использоваться данные о геолокации, данные из других социальных сетей, полученные по идентификатору, указанному в профиле, данные из медиа контента, опубликованного пользователями и сообществами. Используя метод, основанный на линейной комбинации, можно улучшить результаты, пересчитывая на обучающих данных коэффициенты признаков. Предложенные признаки дают высокую полноту, а при применении других алгоритмов семантического анализа или других более точных алгоритмов получения тематики текста, возможно, смогут дать более точное выделение подгруппы.

К недостатку предлагаемого метода можно отнести то, что для описанных в примере признаков качество результатов может сильно ухудшаться для некоторых видов подгрупп. Использование других модификаций случайных марковских полей также может улучшить результат. Результаты может также улучшить усовершенствование существующих признаков. Развитию данного подхода может поспособствовать апробирование представленного метода на других группах и подгруппах.

СПИСОК ЛИТЕРАТУРЫ:

1. Beating the news' with EMBERS: Forecasting civil unrest using open source indicators / N. Ramakrishnan [идр.] // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM. 2014. – PP. 1799–1808.

2. Grothoff C., Porup J. The NSA's SKYNET program may be killing thousands of innocent people // HAL Inria. – 2016.

3. Blachnio A., Przepiorka A., D´az-Morales J. F. Facebook use and chronotype: Results of a cross-sectional study // Chronobiology international. – 2015. – T. 32, № 9. – PP. 1315–1319.

4. Personality, gender, and age in the language of social media: The open- vocabulary approach / H. A. Schwartz [и др.] // PloS one. – 2013. – T. 8, № 9. – e73791.DOI: 10.1371/journal.pone.0073791.

5. Определение демографических атрибутов пользователей микроблогов / Д. Турдаков [и др.] // Труды Института системного программирования РАН. – 2013. – Т. 25. – С. 179–192.

6. Peersman C., Daelemans W., Van Vaerenbergh L. Predicting age and gender in online social networks // Proceedings of the 3rd international workshop on Search and mining user-generated contents. – ACM. 2011. – PP. 37–44.

7. Trusov M., Bodapati A. V., Bucklin R. E. Determining influential users in internet social networks // Journal of Marketing Research. – 2010. – T. 47, № 4. – PP. 643–658.

8. A 61-million-person experiment in social influence and political mobilization / R. M. Bond [и др.] // Nature. – 2012. – T. 489, № 7415. – PP. 295–298.

9. Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber? / Pablo Barberá, John T Jost, Jonathan Nagler Joshua A Tucker, Richard Bonneau // Psychological science. 2015 Oct;26(10):1531-42. doi: 10.1177/0956797615594620.

10. Yardi S., Boyd D. Dynamic debates: An analysis of group polarization over time on twitter // Bulletin of Science, Technology & Society. – 2010. – T. 30, № 5. – PP. 316–327.

11. Lo J., Proksch S.-O., Gschwend T. A common left-right scale for voters and parties in Europe // Political Analysis. – 2014. – T. 22, № 2. – PP. 205–223.

12. Bonica A. Ideology and interests in the political marketplace // American Journal of Political Science. – 2013. – T. 57, № 2. – PP. 294–311.

13. Gruzd A., Roy J. Investigating political polarization on Twitter: A Canadian perspective // Policy & Internet. – 2014. – T. 6, № 1. – PP. 28–45.

14. Golbeck J., Robles C., Turner K. Predicting personality with social media // CHI'11 Extended Abstracts on Human Factors in Computing Systems. – ACM. 2011. – PP. 253–262.

15. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter / L. Sloan [и др.] // Sociological research online. – 2013. – T. 18, № 3. – P. 7.

16. Harvesting multiple sources for user profile learning: a big data study / A. Farseev [и др.] // Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. – ACM.2015. – PP.235–242.

17. Baluja S., Rowley H. A. Boosting sex identification performance // International Journal of computer vision. – 2007. – T. 71, № 1. – PP. 111–119.

18. Wu M.-J., Jang J.-S. R., Lu C.-H. Gender Identification and Age Estimation of Users Based on Music Metadata. // ISMIR. – 2014. – PP. 555–560.

19. Liu J.-Y., Yang Y.-H. Inferring personal traits from music listening history // Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. – ACM. 2012. – PP. 31–36.

20. Matching Entities Across Online Social Networks / O. Peled [идр.] // arXiv preprint arXiv:1410.6717. – 2014.

21. Introduction to information retrieval. T. 1 / C. D. Manning, P. Raghavan, H. Schutze [и др.]. – Cambridge university press Cambridge, 2008.–P. 504.

22. Chemudugunta C., Steyvers P. S. M. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model // Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. T. 19. – MIT Press. 2007. – P. 241.

23. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // the Journal of machine Learning research. – 2003. – T. 3. – PP. 993–1022.

24. Indexing by latent semantic analysis / S. Deerwester [идр.] // Journal of the American society for information science. – 1990. – T. 41, № 6. – P. 391.

25. Kindermann R., Snell J. L. [и др.] Markov random fields and their applications. T. 1. – American Mathematical Society Providence, RI, 1980. – P. 147.

26. Li S. Z. Markov random field modeling in image analysis. – 3rd Edition. Springer Science & Business Media, 2009. – P. 371.

27. Lafferty J., McCallum A., Pereira F. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". Proceedings of the 18th International Conference on Machine Learning. Pages 282–289. 2001.

28. Антонова А.Ю., Соловьев А.Н. Метод условных случайных полей в задачах обработки русскоязычных текстов. «Информационные технологии и системы — 2013», Калининград, 2013. [Электронный ресурс]. URLhttp://itas2013.iitp.ru/pdf/1569759547.pdf(датаобращения: 04.10.2020).

29. Теория матриц/ П. Ланкастер: пер. с англ. С. П. Демушкина М.: Наука. Главная редакция физико-математической литературы, 1978. – 280 с.

Статья поступила в редакцию 11.11.2020
Статья принята к публикации 11.12.2020