Math-Net.Ru

Общероссийский математический портал

В. Ф. Столярова, А. В. Торопова, А. Л. Тулупьев, Модель для оценки частоты публикации постов в онлайн социальной сети по неполным данным с учетом объективных детерминант поведения, *Нечеткие системы и мягкие вычисления*, 2021, том 16, выпуск 2, 77–95

DOI: <https://doi.org/10.26456/fssc81>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 52.205.19.152

26 апреля 2023 г., 11:52:27

МОДЕЛЬ ДЛЯ ОЦЕНКИ ЧАСТОТЫ ПУБЛИКАЦИИ ПОСТОВ В ОНЛАЙН СОЦИАЛЬНОЙ СЕТИ ПО НЕПОЛНЫМ ДАННЫМ С УЧЕТОМ ОБЪЕКТИВНЫХ ДЕТЕРМИНАНТ ПОВЕДЕНИЯ¹

Столярова В.Ф.* , Торопова А.В.** , Тулупьев А.Л.**,*

*Санкт-Петербургский федеральный исследовательский центр Российской Академии наук, г. Санкт-Петербург

**Санкт-Петербургский государственный университет, г. Санкт-Петербург

Поступила в редакцию 25.08.2021, после переработки 06.12.2021.

Профилирование пользователя онлайн социальной сети включает задачу оценки частоты (интенсивности) различных действий, в частности, публикации постов. Однако в силу ресурсных ограничений, может быть доступна только неполная информация о времени публикации нескольких последних постов, полученная, например, в рамках интервью. Оценка интенсивности постинга на основании таких данных востребована при анализе индивидуального риска, связанного с использованием онлайн социальных сетей. В статье предложена расширенная байесовская сеть доверия, которая использует не только информацию о времени публикации последних постов, но и объективные данные из профиля пользователя: пол, возраст, число друзей. Для обучения и демонстрации работы модели были собраны данные о публикации постов случайных пользователей в онлайн социальной сети ВКонтакте. Расширенная структура имеет более высокое значение информационного критерия Акаике по сравнению с упрощенной.

Ключевые слова: онлайн социальные сети, частота постинга, байесовские сети доверия, детерминанты поведения, профилирование пользователей.

Нечеткие системы и мягкие вычисления. 2021. Т. 16, № 2. С. 77–95.
<https://doi.org/10.26456/fssc81>

Введение

Онлайн социальные сети (ОСС) тесно связаны с повседневной деятельностью человека. Аудитория онлайн социальных сетей огромна, согласно релизу пресс-центра ВКонтакте [30], ежедневная аудитория социальной сети ВКонтакте составляет 48 миллионов пользователей; и каждый день ВКонтакте посещают 47% российских интернет-пользователей. При этом информация, которую пользователь публикует о себе в социальной сети, часто является отражением его

¹Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № 0073-2019-0003; при финансовой поддержке РФФИ, проект № 19-37-90120; проект № 20-07-00839.

© Столярова В.Ф., Торопова А.В., Тулупьев А.Л., 2021

индивидуальных личностных характеристик [25]. Таким образом, онлайн социальные сети становятся уникальной площадкой для получения данных и знаний о человеке, его поведении и распространении информации в обществе [2]. Развитие методов и технологий сбора и анализа данных из онлайн социальных сетей позволяет использовать информацию о пользователе в социологических, психологических и маркетинговых [1] исследованиях, а также при моделировании социокрибертехнических систем, в том числе при оценке защищенности информационных систем от социотехнических атак [27,31]. Поведение сотрудника может нести угрозу безопасности критичной информации организации, так как сотрудник является целью социоинженерной атаки злоумышленника [27–29]. В этом случае ряд поведенческих факторов, составляющих основу для построения профиля уязвимостей пользователя, может быть связан с риском успешной реализации социоинженерного атакующего воздействия. Оценка параметров такого поведения важна для обеспечения кибербезопасности компании. Кроме того, подобная информация может использоваться работодателями для оценки профессиональных качеств сотрудника [10].

Оценка параметров использования онлайн социальных сетей важна и с точки зрения области охраны общественного здоровья. С одной стороны использование онлайн социальных сетей предоставляет новые способы взаимодействия между пациентами и/или сотрудниками сферы здравоохранения [6], новые возможности для разработки превентивных вмешательств и анализа рискованного поведения [8]. С другой стороны использование ОСС может являться разрушающим для психического здоровья человека [15], и, соответственно, несет нагрузку на систему здравоохранения. Существует понятие *проблемного, близкого к зависимому, использования онлайн социальных сетей* (problematic Facebook use) [15], которое связано с непосредственным риском здоровью человека. С таким видом поведения связаны ряд психологических характеристик [3,5,32]. Отмечается также, что проблемные пользователи ОСС имеют статистически более высокие значения объективных численных показателей: большее число друзей, посещенных мероприятий, отправленных сообщений и прочее [16].

Таким образом, в ряде исследований актуальной является задача моделирования поведения пользователя онлайн социальной сети в контексте оценки того или иного типа риска (например, риска здоровью индивида или же риска распространения социоинженерного атакующего воздействия). Большая часть доступной информации о пользователе онлайн социальной сети сосредоточена в его профиле [20], и представляет собой сложную систему как объективных данных: технических (время последнего посещения, даты публикации постов, число друзей), графических (фотографии и видео) и неточных (самоотчеты пользователя: имя, дата рождения, образование и прочее), так и данных о предпочтениях пользователя при использовании ОСС.

В центре внимания данной работы находится частота (или интенсивность) такого типа поведения в ОСС, как публикация постов. Этот показатель является составной частью различных шкал измерения проблемного, близкого к зависимому, использования онлайн социальной сети [5,7,18,32] и сам по себе является объективной косвенной характеристикой проблемного использования ОСС [16]. Отметим, что прямое измерение частоты публикации постов не всегда может быть доступно в силу дизайна проводимого исследования или ресурсных (бюджетных, временных, кадровых и т.д.) ограничений. Ресурсные ограничения возникают в ситуациях, когда требуется быстрая оценка параметров поведения индивида, на-

пример, на приеме у врача для оценки возможных рисков здоровью пациента или же при беседе с потенциальным сотрудником для оценки потенциальных уязвимостей. Кроме того, возникает проблема получения достоверной оценки интенсивности использования ОСС. В рамках психологических и социологических исследований, используются интервью и опросы [5,32] с вопросами типа: как часто Вы совершаете то или иное действие в онлайн социальной сети? В пилотных исследованиях [38], было показано, что в силу когнитивных искажений (искажение припоминания, искажение социальной ожидаемости) человек дает неточную информацию на подобные запросы. С достаточной степенью точности припоминается лишь несколько последних эпизодов и рекордные интервалы между эпизодами (минимальный и максимальный) за некоторый промежуток времени. Отметим здесь, что опросы, проводимые при помощи онлайн социальных сетей, не демонстрируют дополнительных искажений по сравнению с традиционными опросами [12] и могут использоваться при оценке различных характеристик поведения в социологических исследованиях.

В подобных ситуациях, в которых требуется наименее искаженная и наиболее быстрая оценка интенсивности поведения, обращение к самоотчетам пользователя ОСС о последних эпизодах является обоснованным.

Таким образом, возникает задача оценки частоты различных типов поведения в ОСС по ограниченным (и, возможно, неточным — в случае обращения к самоотчетам) данным о датах выхода постов пользователя и другой доступной информации из его профиля. Такая задача относится к задачам моделирования профиля пользователя онлайн социальной сети, профилирования пользователя (user profiling). Целью статьи является расширение и уточнение существующей модели оценки интенсивности поведения по данным о нескольких последних эпизодах поведения за счет объективной информации об индивиде на примере данных о датах нескольких последних постов в ОСС и информации о поле, возрасте и числе друзей пользователя ОСС, а также сравнение расширенной модели с существующей. Теоретическая значимость работы состоит в формировании расширенной (включающей дополнительную информацию о пользователе) байесовской сети доверия для решения поставленной задачи и численном сравнении качества построенной модели с уже существующей моделью и моделью, обученной на тестовых данных. Практическая значимость заключается в применении предложенной модели для оценки интенсивности постинга в онлайн социальной сети. Этот параметр является центральным при моделировании риска в ряде прикладных задач охраны общественного здоровья и социо-кибербезопасности.

1. Анализ результатов предшествующих работ

Информация об интенсивности совершения действий в онлайн социальной сети для профилирования пользователя используется для принятия решений в ряде задач человеко-машинного взаимодействия. К примеру, интенсивность постинга наравне с интенсивностью других типов поведения в ОСС является важной характеристикой при идентификации аккаунтов, распространяющих спам [9,19,26]. Кроме того, частотные характеристики профиля пользователя используются для выделения классов пользователей в маркетинговых и социологических исследованиях [1, 4, 11, 13].

Оценка частотных характеристик профиля пользователя важна в исследованиях психологической и социологической направленности. Как уже отмечалось,

интенсивность (в контексте частоты различных действий пользователя ОСС) является составной частью различных шкал измерения проблемного, близкого к зависимому, использования онлайн социальной сети [5, 7, 18, 32] и сама по себе является объективной косвенной характеристикой проблемного использования ОСС [16]. Моделирование профиля пользователя в контексте социологического исследования обладает рядом особенностей, в частности, интенсивно используется неточная информация из самоотчетов пользователя (опросы и интервью) и рассматриваются разнообразные индивидуальные характеристики наряду с экспертной информацией. Для учета этих особенностей оправданным является использование байесовских сетей доверия для моделирования многогранной и неточной информации о пользователе.

Существующая модель байесовской сети доверия для оценки интенсивности поведения по данным о трех последних его эпизодах была опробована на данных о постинге в онлайн социальных сетях [22, 23]. Однако расширение такой модели при помощи объективных данных из профиля пользователя является новым.

2. Байесовские сети доверия

«Байесовская сеть доверия — это пара $\langle G, P \rangle$, где G — ациклический направленный граф (то есть граф, в котором нет направленных циклов, ненаправленные циклы допускаются), P — совместное распределение вероятностей всех случайных элементов, приписанных узлам графа (множество тензоров условных вероятностей), причем для любого узла соответствующий ему случайный элемент условно независим от всех своих непотомков, которые также не являются его родителями, при заданном означивании своих родителей». [37]

Байесовская сеть доверия позволяет моделировать причинно-следственные взаимосвязи предметной области наглядным образом (в виде графа), допускает естественное сочетание экспертной и статистической информации при проведении вывода [14]. Кроме того, что байесовские сети позволяют представлять данные с неполнотой вычислительно приемлемым образом, они позволяют проводить апостериорный вероятностный вывод, который заключается в пропагации (распространении по сети) поступившего свидетельства — некоторых новых обуславливающих данных, на основе которых строятся апостериорные оценки означиваний случайных элементов. Алгоритмы вывода в байесовских сетях доверия основаны на теореме Байеса и правиле декомпозиции, примеры можно найти в [17].

На Рис. 1 приведен пример структуры байесовской сети доверия, таблицы представляют собой таблицы условной вероятности для этой структуры, то есть над всеми случайными элементами задано совместное распределение вероятностей $P = p(a), p(b|a), p(c|a, b)$ (для простоты будем считать случайные элементы бинарными).

Таблица 1: Таблицы условной вероятности для P

a	0.6
$\neg a$	0.4

	a	$\neg a$
b	0.7	0.4
$\neg b$	0.3	0.6

	a		$\neg a$	
	b	$\neg b$	b	$\neg b$
c	0.5	0.3	0.9	0.6
$\neg c$	0.5	0.7	0.1	0.4

Рис. 1: Пример структуры байесовской сети доверия.

3. Метод последних эпизодов и рекордных интервалов в оценке интенсивности поведения

В ряде случаев необходимая информация об эпизодах публикации постов может оказаться недоступной в силу дизайна исследования (когда не предполагается обращение к технической информации из профиля пользователя). В этом случае исследователю доступна информация лишь о нескольких эпизодах такого поведения. При обращении к самоотчетам человека о его поведении, наименее подвержена когнитивным искажениям информация о нескольких последних эпизодах поведения [38]. Чтобы снизить влияние неопределенности на численные значения, был предложен метод последних эпизодов для оценки частоты поведения по таким неполным данным [24, 33, 38].

Метод последних эпизодов поведения [33] опирается на опросный инструмент, который содержит вопросы особого типа: *когда в последний (предпоследний) раз Вы совершали то или иное действие?*, и на математическую модель поведения индивида, в основе которой лежит точечный процесс Пуассона [24, 34, 38]. Формализация задачи восстановления интенсивности поведения по данным о последних эпизодах состоит в следующем: пусть в момент интервью I исследователь собирает информацию о времени реализации последнего, предпоследнего и пред-предпоследнего за определенный промежуток времени T эпизодов некоторого типа поведения: t_1, t_2, t_3 . Отметим, что в рамках интервью возможно получение различного числа данных о времени реализации эпизодов поведения. К примеру, часть респондентов может припоминать только последний эпизод поведения, а часть вспомнить более трех эпизодов. В данной работе используется модель с тремя возможными эпизодами поведения, однако она может быть расширена до большего числа эпизодов.

В пилотных исследованиях показано, что охотнее всего респонденты припоминают два-три последних эпизода обыденного поведения, (к примеру, участия в эпизодах рискованного поведения для групп лиц с девиантным поведением) [HIVmono]. В качестве ключевого параметра выступает частота (или интенсивность) поведения, представляющая число эпизодов за определенный промежуток времени наблюдения $\lambda = N/T$. Для решения задачи оценки λ по неполным данным о времени реализации эпизодов была построена математическая модель поведения, предложены оценки максимального правдоподобия различных систем ответов [34]. В рамках этой модели предполагается, что эпизоды процесса реализуются согласно вехам точечного процесса Пуассона.

4. Байесовская сеть доверия для оценки интенсивности поведения по данным о последних эпизодах и дополнительной информации о пользователе ОСС

Однако для работы с данными о поведении человека необходимо учитывать различную вариабельность входящих в модель переменных. В частности, в ходе исследования могут быть доступны различные наборы данных об эпизодах поведения (всего один эпизод или несколько). Кроме того, поведение человека само по себе обусловлено индивидуальными характеристиками, параметр λ варьируется между индивидами (является случайной величиной). Поэтому для практических приложений, связанных с принятием решений в условиях неопределенности (оценкой риска), была предложена структура байесовской сети доверия [35, 36], которая учитывает как вариабельность модели, так и сопутствующую самоотчетам неопределенность ответов [35].

Множество вершин для такой сети (Рис. 2) есть $V = \{\lambda, \tau_1, \tau_2, \tau_3\}$, где λ представляет собой искомый параметр интенсивности (частоты) поведения, τ_1 есть интервалы между моментом интервью и последним эпизодом поведения, τ_2 — между последним и предпоследним эпизодами, τ_3 — между предпоследним и предпредпоследним эпизодами.

Если же для оценки интенсивности поведения важно учитывать дополнительную объективную информацию о пользователе ОСС, то такие характеристики будут оказывать влияние на интервалы между эпизодами поведения посредством интенсивности λ (Рис. 2).

Рис. 2: Экспертно заданная структура байесовской сети доверия для задачи оценки интенсивности поведения λ по данным о трех последних эпизодах поведения τ_1, τ_2, τ_3 ; **слева:** изначальная, упрощенная структура без учета объективных данных о пользователе ОСС; **справа:** расширенная структура, включающая объективные данные о поле *sex*, возрасте *age* и числе друзей пользователя *friendscount*.

5. Данные

Для обучения и проверки моделей были собраны данные о профилях пользователей онлайн социальной сети ВКонтакте: 9140 пользователей в тестовой

выборке и 9004 в тренировочной. Для всех открытых профилей были загружены значения следующих атрибутов: дата рождения, пол, время последнего посещения социальной сети и количество друзей, а также даты публикаций 100 последних записей пользователя на стене (больше 100 записей собрать невозможно в связи с ограничениями API ВКонтакте). Так как цель работы заключается в сравнении предлагаемых расширенных моделей с упрощенной, то для тестовой и тренировочной выборок были взяты только полные наблюдения без пропущенных значений. После удаления из тестовой выборки пользователей, у которых скрыта информация в аккаунте или не указана дата рождения, а также отсутствовали хотя бы 3 поста в течение года, ее объем составил 1472 пользователя, а тренировочной: 1490. Отметим, что пользователи младше 18 лет также исключались из исследования по этическим соображениям.

По данным о времени публикации постов на стене пользователя была вычислена частота публикации постов за год (как отношение количества постов к 365) и значения временных интервалов между последним появлением пользователя в сети и временем последнего поста, между последним и предпоследним постами и между предпоследним и пред-предпоследним постами в днях. Для использования в байесовской сети доверия для непрерывных переменных были выбраны следующие интервалы дискретизации: для интенсивности постинга — $[0, 0.015)$, $[0.015, 0.03)$, $[0.03, 0.06)$, $[0.06, 0.14)$, $[0.14, \infty)$, для временных интервалов — $[0, 1)$, $[1, 7)$, $[7, 14)$, $[14, 30)$, $[30, 90)$, $[90, \infty)$ (единица измерения дни). Возраст был разбит на два интервала: младше 35 лет и старше 35 лет, такая дискретизация обусловлена установленным в Российской Федерации возрастом молодежи.

Статистически объективные показатели действительно связаны с частотой публикации постов (Рис. 3). Интенсивность постинга статистически значимо различается между мужчинами и женщинами (с поправкой на возраст): в рассматриваемой выборке женщины чаще публикуют посты в социальных сетях, чем мужчины. Для каждого диапазона интенсивности постинга, кроме самого высокого, число друзей у мужчин и женщин статистически значимо отличается.

6. Обучение байесовских сетей доверия

Тензоры условных вероятностей в упрощенной и экспертной расширенной структурах БСД (Рис. 2) были оценены при помощи метода максимального правдоподобия [17, 37] на данных тренировочной выборки. Чтобы проверить истинность предположений об условной независимости в экспертной структуре (Рис. 2), при помощи алгоритма оптимизации Hill-Climbing (HC) [21] была выбрана оптимальная структура сети для тренировочной выборки. В качестве основной меры качества модели использовался информационный критерий Акаике (AIC). Полученная структура представлена на Рис. 4.

Для оценки силы взаимосвязи переменных в полученной структуре использовалась разница между правдоподобием моделей с дугой и без неё [17, 21]. В таблице 2 приведены значения силы для каждой дуги, чем меньше значение — тем сильнее связь между соответствующими вершинами.

7. Сравнение моделей

Для трех заданных моделей (упрощенной, экспертной расширенной и оптимальной расширенной) были вычислены значения информационных критериев,

Рис. 3: Статистические характеристики тестовой выборки. А: зависимость дискретизированной частоты постинга от числа друзей с группировкой по полу. В: зависимость рассчитанной частоты постинга от пола с группировкой по возрасту.

Рис. 4: Обученная структура модели оценки интенсивности с факторами поведения.

Таблица 2: Сила связей дуг между узлами оптимальной для тренировочных данных структуры

Начало дуги	Конец дуги	Сила
λ	τ_1	-299,54
λ	τ_2	-205,381
λ	τ_3	-190,153
<i>friends count</i>	λ	-24,478
<i>age</i>	<i>friends count</i>	-15,039
<i>age</i>	λ	-5,121
<i>age</i>	<i>sex</i>	-2,549

Таблица 3: Значения информационных критериев

Модель	тренировочная выборка			тестовая выборка		
	AIC	BIC	loglik	AIC	BIC	loglik
упрощенная	-9505	-9715	-9426	-9298	-9507	-9219
расширенная	-13542	-13924	-13398	-13223	-13604	-13398
с обученной структурой	-13510	-13818	-13394	-13196	-13503	-13079

а также показатели качества классификации. В таблице 3 представлены значения информационных критериев для трех моделей на двух выборках: тестовой и тренировочной. Были использованы следующие критерии: информационный критерий Акаике (AIC), Байесовский информационный критерий (BIC) и критерий отношения правдоподобия (loglik). Все эти критерии можно посчитать с помощью функции правдоподобия: $\loglik = \log L$, где L — это функция правдоподобия, $AIC = -2\loglik + 2k$, k — количество использованных параметров, $BIC = -2\loglik + 2\ln(n)$, где n — размер выборки. По сравнению с BIC, AIC может выбирать более сложные модели, что означает, что она может уделять больше внимания производительности модели на обучающем наборе данных, однако, если выбор моделей-кандидатов включает истинную модель для набора данных, то вероятность того, что BIC выберет истинную модель, увеличивается с ростом размера обучающего набора данных, что нельзя сказать о показателе AIC. Расширенные структуры показали лучшие результаты по сравнению с упрощенной моделью. Обе расширенные модели показали похожие результаты.

На следующем шаге были рассчитаны показатели качества классификации трех моделей на тестовой выборке. Для этого были получены предсказания частоты постинга по данным в моделях. Точность (ассигасу) представляет собой отношение правильно сделанных предсказаний к общему числу предсказаний. Мера каппа (карра) используется для измерения эффективности классификаторов, ее значения расположены в интервале от -1 (полное несогласие) до 1 (полное согласие) и проходят через 0 (случайная классификация). Средняя точность (average assigasy) — это доля правильно классифицированных экземпляров в матрицах смежности для каждого класса. Все три модели оказались сопоставимы, это можно объяснить тем, что связи с индивидуальными характеристиками пользователя являются довольно слабыми (см. таблицу 4).

Таблица 4: Критерии качества предсказаний моделей

Модель	Точность	95% Дове- рительный интервал	Каппа	Средняя точность
упрощенная	0.475	(0.4491, 0.5007)	0.334	0.788
расширенная	0.481	(0.4552, 0.5069)	0.342	0.791
с обученной структурой	0.477	(0.4511, 0.5028)	0.336	0.789

Заключение

В ряде задач оценки риска, которые связаны с моделированием поведения человека, важно учитывать индивидуальные характеристики, такие как пол и возраст. В статье предложена расширенная модель интенсивности постинга в онлайн социальной сети индивида по данным о последних постах. Эта характеристика важна, например, при оценке индивидуального риска здоровью человека (в контексте проблемного использования ОСС). Для учета влияния на интенсивность постинга пола, возраста и числа друзей пользователя были предложены экспертная структура байесовской сети доверия, а также обученная по данным из ОСС ВКонтакте структура. Согласно информационным критериям, обе расширенные структуры похожи, и представляют лучшее отражение информации о пользователе ОСС. Однако по качеству классификации все три рассматриваемые модели (исходная без дополнительных характеристик пользователя, экспертная расширенная и обученная расширенная) сопоставимы. Это связано с тем, что взаимосвязь интенсивности поведения с объективными характеристиками профиля пользователя имеется (Рис. 4), но является структурно слабой (Таблица 2).

Таким образом, использование дополнительной информации важно с точки зрения полноты представления информации, однако требуется дополнительные исследования для идентификации наиболее релевантных для поставленной задачи характеристик пользователя.

Список литературы

- [1] Alalwan A.A., Rana N.P., Dwivedi Y.K., Algharabat R. Social media in marketing: A review and analysis of the existing literature // Telematics and Informatics. 2017. Vol. 34, № 7. Pp. 1177–1190.
- [2] Basaran S., Ejimogu O.H. A Neural Network Approach for Predicting Personality From Facebook Data // SAGE Open. 2021. Vol. 11, № 3. ID 21582440211032156.
- [3] Busalim A.H., Masrom M., Binti Wan Zakaria W.N. The impact of Facebook Addiction and self-esteem on students' academic performance: A multi-group analysis // Computers Education. 2019. Vol. 142. ID 103651.
- [4] Van Dam J.W., Van De Velden M. Online profiling and clustering of Facebook users // Decision Support Systems. 2015. Vol. 70. Pp. 60–72.
- [5] Ellison N.B., Steinfield C., Lampe C. The Benefits of Facebook "Friends": Social Capital and College Students' Use of Online Social Network Sites // Journal of Computer-Mediated Communication. 2007. Vol. 4, № 12. Pp. 1143–1168.

- [6] Giustini D., Ali S.M., Fraser M., Boulos M.N.K. Effective uses of social media in public health and medicine: a systematic review of systematic reviews // *Online journal of public health informatics*. 2018. Vol. 10, № 2. ID e215.
- [7] Gosling S.D., Augustine A.A., Vazire S., Holtzman N., Gaddis S. Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information // *Cyberpsychology, Behavior, and Social Networking*. 2011. Vol. 14, № 9. Pp. 483–488.
- [8] Ilakkuvan V., Johnson A., Villanti A.C., Evans W.D., Turner M. Patterns of social media use and their relationship to health risks among young adults // *Journal of Adolescent Health*. 2019. Vol. 64, № 2. Pp. 158–164.
- [9] Inuwa-Dutse I., Liptrott M., Korkontzelos I. Detection of spam-posting accounts on Twitter // *Neurocomputing*. 2018. Vol. 315. Pp. 496–511.
- [10] Junco R. Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance // *Computers in human behavior*. 2012. Vol. 28, № 1. Pp. 187–198.
- [11] Kachamas P., Akkaradamrongrat S., Sinthupinyo S., Chandrachai A. Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis // *International Journal of Machine Learning and Computing*. 2019. Vol. 9, № 1. Pp. 91–97.
- [12] Kalimeri K., Beiro M.G., Bonanomi A., Rosina A., Cattuto C. Traditional versus Facebook-based surveys // *Demographic research*. 2020. Vol. 42. Pp. 133–148.
- [13] Khlobystova A.O., Abramov M.V., Tulupyeva T.V. Application of the Alternatives Method Probabilities in Construction of Intensity of User Communications Estimates // 2020 XXIII IEEE International Conference on Soft Computing and Measurements, SCM (St. Petersburg, Russia. 2020). Pp. 22–24. <https://doi.org/10.1109/SCM50615.2020.9198751>
- [14] Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniques. The MIT Press, 2009. 1231 p.
- [15] Marino C., Gini G., Vieno A., Spada M.M. A comprehensive meta-analysis on problematic Facebook use // *Computers in Human Behavior*. 2018. Vol. 83. Pp. 262–277.
- [16] Marino C., Finos L., Vieno A., Lenzi M., Spada M.M. Objective Facebook behaviour: Differences between problematic and non-problematic users // *Computers in Human Behavior*. 2017. Vol. 73. Pp. 541–546.
- [17] Neapolitan R.E. Learning Bayesian Networks. Pearson Prentice Hall, 2003. 674 p.
- [18] Orosz G., Toth-Kiraly I., Bothe B. Four facets of Facebook intensity — The development of the Multidimensional Facebook Intensity Scale // *Personality and Individual Differences*. 2016. Vol. 100. Pp. 95–104. <https://doi.org/10.1016/j.paid.2015.11.038>
- [19] Rathore S., Loia V., Park J.H. SpamSpotter: an efficient spammer detection framework based on intelligent decision support system on facebook // *Applied Soft Computing*. 2018. Vol. 67. Pp. 920–932.

- [20] Schiaffino S., Amandi A. Intelligent user profiling // Artificial Intelligence An International Perspective. Ed. by M. Bramer. Series: Lecture Notes in Artificial Intelligence. Vol. 5640. Berlin, Heidelberg: Springer, 2009. Pp. 193–216. https://doi.org/10.1007/978-3-642-03226-4_11
- [21] Scutari M. Learning Bayesian Networks with the bnlearn R Package // Journal of Statistical Software. 2010. Vol. 35, № 3. Pp. 1–22. <https://doi.org/10.18637/jss.v035.i03>
- [22] Stoliarova V.F., Tulupyev A.L. Regression Model for the Problem of Parameter Estimation in the Gamma Poisson Model of Behavior: an Application to the Online Social Media Posting Data // Conference proceedings of XXI International conference on Soft Computing and Measurements, SCM-2021 (St. Petersburg). (to appear)
- [23] Toropova A., Tulupyeva T. Comparison of Behavior Rate Models Based on Bayesian Belief Network // Recent Research in Control Engineering and Decision Making. ICIT 2020. Eds. by Dolinina O. et al.. Series: Studies in Systems, Decision and Control. Vol. 337. Cham: Springer, 2021. Pp. 510–521. https://doi.org/10.1007/978-3-030-65283-8_42
- [24] Tulupyev A., Suvorova A., Sousa J., Zelterman D. Beta prime regression with application to risky behavior frequency screening // Statistics in medicine. 2013. Vol. 32, № 23. Pp. 4044–4056.
- [25] Wilson R.E., Gosling S.D., Graham L.T. A review of Facebook research in the social sciences // Perspectives on psychological science. 2012. Vol. 7, № 3. Pp. 203–220.
- [26] Wu T., Wen S., Xiang Y., Zhou W. Twitter spam detection: Survey of new approaches and comparative study // Computers & Security. 2018. Vol. 76. Pp. 265–284.
- [27] Абрамов М.В. Автоматизация анализа социальных сетей для оценивания защищённости от социоинженерных атак // Автоматизация процессов управления. 2018. Т. 1, № 51. С. 34–40.
- [28] Абрамов М.В., Тулупьева Т.В., Тулупьев А.Л. Социоинженерные атаки: социальные сети и оценки защищенности пользователей. СПб.: ГУАП, 2018. 266 с.
- [29] Азаров А.А., Тулупьева Т.В., Суворова А.В., Тулупьев А.Л., Абрамов М.В., Юсупов Р.М. Социоинженерные атаки: проблемы анализа. СПб.: Наука, 2016. 349 с.
- [30] ВКонтakte подвела итоги первого квартала 2021 года: выручка компании выросла на 21%, а аудитория — на 6% [Электронный ресурс] // <https://vk.com/press/q1-2021-results>. URL: .
- [31] Корепанова А.А., Абрамов М.В., Тулупьева Т.В. Идентификация аккаунтов пользователей в социальных сетях «ВКонтакте» и «Одноклассники» // Сборник научных трудов семнадцатой национальной конференции по искусственному интеллекту с международным участием, КИИ-2019 (21–25 октября 2019 г., г. Ульяновск, Россия). Ульяновск: УлГТУ, 2019. С. 153–163.

- [32] Корниенко Д.С., Горбушина Е.А. Особенности профилей пользователей, интенсивность и навязчивость в использовании социальной сети и перфекционистская самопрезентация // Вестник Пермского государственного гуманитарно-педагогического университета. Серия № 1. Психологические и педагогические науки. 2020. № 1. С. 14–25.
- [33] Пашенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В. Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здоровоохранение Российской Федерации. 2010. № 2. С. 32–35.
- [34] Степанов Д.В., Мусина В.Ф., Суворова А.В., Тулупьев А.Л., Сироткин А.В., Тулупьева Т.В. Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита // Информатика и автоматизация. 2012. № 23. С. 157–184.
- [35] Суворова А.В. Модели и алгоритмы анализа сверхкоротких гранулярных временных рядов на основе байесовских сетей доверия: Дисс. ... канд. физ.-мат. наук. 2013.
- [36] Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления. 2014. Т. 9, № 2. С. 115–129.
- [37] Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей. Учебник. СПб.: Изд-во С.Петербург. ун-та, 2019. 399 с.
- [38] Тулупьева Т.В., Пашенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С. Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. Монография. 2008. 346 с.

Образец цитирования

Столярова В.Ф., Торопова А.В., Тулупьев А.Л. Модель для оценки частоты публикации постов в онлайн социальной сети по неполным данным с учетом объективных детерминант поведения // Нечеткие системы и мягкие вычисления. 2021. Т. 16, № 2. С. 77–95. <https://doi.org/10.26456/fssc81>

Сведения об авторах

1. Столярова Валерия Фуатовна

младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Санкт-Петербургского федерального исследовательского центра Российской Академии наук.

199178, Санкт-Петербург, 14-я линия В.О., д. 39, СПб ФИЦ РАН.

2. Торопова Александра Витальевна

младший научный сотрудник кафедры информатики Санкт-Петербургского государственного университета.

199034, Санкт-Петербург, Университетская наб., д. 7-9, СПбГУ.

3. Тулупьев Александр Львович

профессор кафедры информатики Санкт-Петербургского государственного университета; главный научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Санкт-Петербургского федерального исследовательского центра Российской Академии наук.

199034, Санкт-Петербург, Университетская наб., д. 7-9, СПбГУ.

A MODEL FOR ESTIMATING THE POSTING FREQUENCY IN AN ONLINE SOCIAL MEDIA WITH INCOMPLETE DATA USING OBJECTIVE DETERMINANTS OF USERS' BEHAVIOUR

Stoliarova Valeria Fuatovna

Junior Researcher at the Laboratory of Theoretical and Interdisciplinary Problems
of Computer Science, St. Petersburg Federal Research Center of the Russian
Academy of Sciences
Russia, 199178, St. Petersburg, 14th line V.O., 39, St. Petersburg FIS RAS.

Toropova Aleksandra Vitalyevna

Junior Researcher at the Department of Computer Science, St. Petersburg State
University
*Russia, 199034, Saint Petersburg, 7-9 Universitetskaya nab., St. Petersburg State
University.*

Tulupyev Aleksandr L'vovich

Professor at the Department of Computer Science, St. Petersburg State University
Chief Scientific Officer at the Laboratory of Theoretical and Interdisciplinary
Problems of Computer Science, St. Petersburg Federal Research Center of the
Russian Academy of Sciences
*Russia, 199034, Saint Petersburg, 7-9 Universitetskaya nab., St. Petersburg State
University.*

Received 25.08.2021, revised 06.12.2021.

User profiling is related to the problem of estimation of frequency of certain user's actions in an online social media, like posting. But due to limited resources the only information available may be imprecise information on several last episodes of posting, that can be gathered via an interview. The frequency of posting estimates with such limited data may be used in the individual risk assessment that is connected with the use of online social media, for example, in medicine or cybersecurity. In the paper the Bayes belief network (BBN) for this problem is constructed, that incorporates not only the limited data on times of several last posts in an online social media, but the objective data about the user's profile: age, sex, and friends count. With the training dataset gathered via API VKontakte we estimated conditional probability tables for two expert BBN structures (existing reduced structure based only on dates of several last posts and novel extended structure with objective behavior determinants incorporated) and automatically learned the optimal structure for the training data. Both extended models (expert and learned) showed lower values of the information criteria (Akaike information criteria and bayesian information criteria). Then with the test dataset the classification problem of the true frequency value was assessed. All three models showed similar results based on accuracy, kappa and average accuracy characteristics. This result is related to the weak strength of arcs between frequency variable and objective behavior determinants. But nevertheless the use of such variables is important in the application in order to construct the comprehensive structure of the knowledge in the area of

interest. The practical significance of the work lies in the possibility of applying the proposed model to assess the posting frequency in the online social network, in particular in the tasks of modeling risk in the field of public health and socio-cybersecurity.

Keywords: online social networks, posting frequency, Bayesian belief networks, behaviour determinants, user profiling.

Citation

Stoliarova V.F., Toropova A.V., Tulupyev A.L., “A Model for Estimating the Posting Frequency in an Online Social Media with Incomplete Data Using Objective Determinants of Users’ Behaviour”, *Nechetkie Sistemy i Myagkie Vychisleniya [Fuzzy Systems and Soft Computing]*, **16:2** (2021), 77–95 (in Russian). <https://doi.org/10.26456/fssc81>

References

- [1] Alalwan A.A., Rana N.P., Dwivedi Y.K., Algharabat R., “Social media in marketing: A review and analysis of the existing literature”, *Telematics and Informatics*, **34:7** (2017), 1177–1190.
- [2] Basaran S., Ejimogu O.H., “A Neural Network Approach for Predicting Personality From Facebook Data”, *SAGE Open*, **11:3** (2021), 21582440211032156.
- [3] Busalim A.H., Masrom M., Binti Wan Zakaria W.N., “The impact of Facebook Addiction and self-esteem on students’ academic performance: A multi-group analysis”, *Computers Education*, **142** (2019), 103651.
- [4] Van Dam J.W., Van De Velden M., “Online profiling and clustering of Facebook users”, *Decision Support Systems*, **70** (2015), 60–72.
- [5] Ellison N.B., Steinfield C., Lampe C., “The Benefits of Facebook ”Friends”: Social Capital and College Students’ Use of Online Social Network Sites”, *Journal of Computer-Mediated Communication*, **4:12** (2007), 1143–1168.
- [6] Giustini D., Ali S.M., Fraser M., Boulos M.N.K., “Effective uses of social media in public health and medicine: a systematic review of systematic reviews”, *Online journal of public health informatics*, **10:2** (2018), e215.
- [7] Gosling S.D., Augustine A.A., Vazire S., Holtzman N., Gaddis S., “Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information”, *Cyberpsychology, Behavior, and Social Networking*, **14:9** (2011), 483–488.
- [8] Ilakkuvan V., Johnson A., Villanti A.C., Evans W.D., Turner M., “Patterns of social media use and their relationship to health risks among young adults”, *Journal of Adolescent Health*, **64:2** (2019), 158–164.
- [9] Inuwa-Dutse I., Liptrott M., Korkontzelos I., “Detection of spam-posting accounts on Twitter”, *Neurocomputing*, **315** (2018), 496–511.

- [10] Junco R., “Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance”, *Computers in human behavior*, **28**:1 (2012), 187–198.
- [11] Kachamas P., Akkaradamrongrat S., Sinthupinyo S., Chandrachai A., “Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis”, *International Journal of Machine Learning and Computing*, **9**:1 (2019), 91–97.
- [12] Kalimeri K., Beiro M.G., Bonanomi A., Rosina A., Cattuto C., “Traditional versus Facebook-based surveys”, *Demographic research*, **42** (2020), 133–148.
- [13] Khlobystova A.O., Abramov M.V., Tulupyeva T.V., “Application of the Alternatives Method Probabilities in Construction of Intensity of User Communications Estimates”, *2020 XXIII IEEE International Conference on Soft Computing and Measurements*, SCM (St. Petersburg, Russia. 2020), 22–24, <https://doi.org/10.1109/SCM50615.2020.9198751>.
- [14] Koller D., Friedman N., *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, 2009, 1231 pp.
- [15] Marino C., Gini G., Vieno A., Spada M.M., “A comprehensive meta-analysis on problematic Facebook use”, *Computers in Human Behavior*, **83** (2018), 262–277.
- [16] Marino C., Finos L., Vieno A., Lenzi M., Spada M.M., “Objective Facebook behaviour: Differences between problematic and non-problematic users”, *Computers in Human Behavior*, **73** (2017), 541–546.
- [17] Neapolitan R.E., *Learning Bayesian Networks*, Pearson Prentice Hall, 2003, 674 pp.
- [18] Orosz G., Toth-Kiraly I., Bothe B., “Four facets of Facebook intensity — The development of the Multidimensional Facebook Intensity Scale”, *Personality and Individual Differences*, **100** (2016), 95–104, <https://doi.org/10.1016/j.paid.2015.11.038>.
- [19] Rathore S., Loia V., Park J.H., “SpamSpotter: an efficient spammer detection framework based on intelligent decision support system on facebook”, *Applied Soft Computing*, **67** (2018), 920–932.
- [20] Schiaffino S., Amandi A., “Intelligent user profiling”, *Artificial Intelligence An International Perspective*. V.5640, Lecture Notes in Artificial Intelligence, ed. M. Bramer, Springer, Berlin, Heidelberg, 2009, 193–216, https://doi.org/10.1007/978-3-642-03226-4_11.
- [21] Scutari M., “Learning Bayesian Networks with the bnlearn R Package”, *Journal of Statistical Software*, **35**:3 (2010), 1–22, <https://doi.org/10.18637/jss.v035.i03>.
- [22] Stoliarova V.F., Tulupiyev A.L., “Regression Model for the Problem of Parameter Estimation in the Gamma Poisson Model of Behavior: an Application to the Online Social Media Posting Data”, *Conference proceedings of XXI International conference on Soft Computing and Measurements*, SCM-2021 (St. Petersburg) (to appear).

- [23] Toropova A., Tulupyeva T., “Comparison of Behavior Rate Models Based on Bayesian Belief Network”, *Recent Research in Control Engineering and Decision Making. ICIT 2020*. V. 337, Studies in Systems, Decision and Control, eds. Dolinina O. et al., Springer, Cham, 2021, 510–521, https://doi.org/10.1007/978-3-030-65283-8_42.
- [24] Tulupyeu A., Suvorova A., Sousa J., Zelterman D., “Beta prime regression with application to risky behavior frequency screening”, *Statistics in medicine*, **32**:23 (2013), 4044–4056.
- [25] Wilson R.E., Gosling S.D., Graham L.T., “A review of Facebook research in the social sciences”, *Perspectives on psychological science*, **7**:3 (2012), 203–220.
- [26] Wu T., Wen S., Xiang Y., Zhou W., “Twitter spam detection: Survey of new approaches and comparative study”, *Computers & Security*, **76** (2018), 265–284.
- [27] Abramov M.V., “Automation of social network analysis to assess security against socioengineering attacks”, *Avtomatizatsiya protsessov upravleniya [Automation of management processes]*, **1**:51 (2018), 34–40 (in Russian).
- [28] Abramov M.V., Tulupeva T.V., Tulupev A.L., *Sotsioinzhenernye ataki: sotsialnye seti i otsenki zashchishchennosti polzovatelej [Socioengineering attacks: Social networks and user security assessments]*, GUAP, SPb., 2018 (in Russian), 266 pp.
- [29] Azarov A.A., Tulupeva T.V., Suvorova A.V., Tulupev A.L., Abramov M.V., Yusupov R.M., *Sotsioinzhenernye ataki: problemy analiza [Socioengineering attacks: problems of analysis]*, Nauka Publ., SPb., 2016 (in Russian), 349 pp.
- [30] VKontakte summed up the results of the first quarter of 2021: the company’s revenue grew by 21%, and the audience - by 6%, <https://vk.com/press/q1-2021-results> (in Russian).
- [31] Korepanova A.A., Abramov M.V., Tulupeva T.V., “Identification of user accounts in social networks Vkontakte and Odnoklassniki”, *Sbornik nauchnykh trudov semnadtsatoj natsionalnoj konferentsii po iskusstvennomu intellektu s mezhdunarodnym uchastiem, KII-2019 (21–25 oktyabrya 2019 g., g. Ulyanovsk, Rossiya)*, UISTU Publ., Ulyanovsk, 2019, 153–163 (in Russian).
- [32] Kornienko D.S., Gorbushina E.A., “Features of user profiles, intensity and obsession in using the social network and perfectionist self-presentation”, *Vestnik Permskogo gosudarstvennogo humanitarno-pedagogicheskogo universiteta. Seriya № 1. Psikhologicheskie i pedagogicheskie nauki [Bulletin of the Perm State Humanitarian Pedagogical University. Series No. 1. Psychological and pedagogical sciences]*, 2020, № 1, 14–25 (in Russian).
- [33] Pashchenko A.E., Tulupev A.L., Tulupeva T.V., Krasnoselskikh T.V., Sokolovskij E.V., “Indirect assessment of the probability of HIV infection based on data on recent episodes of risky behavior”, *Zdravookhranenie Rossijskoj Federatsii [Health-care of the Russian Federation]*, 2010, № 2, 32–35 (in Russian).
- [34] Stepanov D.V., Musina V.F., Suvorova A.V., Tulupev A.L., Sirotkin A.V., Tulupeva T.V., “Likelihood function with heterogeneous arguments in the identification of the Poisson model of risky behavior in the case of information deficiency”,

- Informatika i avtomatizatsiya [Computer Science and automation]*, 2012, № 23, 157–184.
- [35] Suvorova A.V., *Models and algorithms for the analysis of ultrashort granular time series based on Bayesian trust networks*, PhD Thesis, 2013 (in Russian).
- [36] “Bayesian belief networks for risky behavior rate estimates”, *Nechetkie Sistemy i Myagkie Vychisleniya [Fuzzy Systems and Soft Computing]*, **9:2** (2014), 115–129 (in Russian).
- [37] Tulupev A.L., Nikolenko S.I., Sirotkin A.V., *Osnovy teorii bajesovskikh setej [Fundamentals of Bayesian network theory]*, Handbook, Saint-Petersburg State University Publ., SPb., 2019 (in Russian), 399 pp.
- [38] Tulupeva T.V., Pashchenko A.E., Tulupev A.L., Krasnoselskikh T.V., Kazakova O.S., *Modeli VICH-riskovannogo povedeniya v kontekste psikhologicheskoy zashchity i drugih adaptivnykh stilej [Models of HIV-risky behavior in the context of psychological protection and other adaptive styles]*, Monograph, 2008 (in Russian), 346 pp.