Компьютерные инструменты в образовании, 2019
№ 3: 29–43
УДК: 004.8
http://cte.eltech.ru
doi:10.32603/2071-2340-2019-3-29-43

алгоритмическая математика и математическое моделирование

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ АККАУНТОВ ПОЛЬЗОВАТЕЛЯ В ДВУХ СОЦИАЛЬНЫХ СЕТЯХ*

Корепанова А. А.[2,1], младший научный сотрудник, aak@dscs.pro
Олисеенко В. Д.[1,2], студент, subster3@gmail.com
Абрамов М. В.[2,1], кандидат технических наук, ✉ mva@dscs.pro
Тулупьев А. Л.[1,2], доктор физико-математических наук, alt@dscs.pro

[1]Санкт-Петербургский государственный университет,
Университетская наб., 7–9, 199034, Санкт-Петербург, Россия
[2]Санкт-Петербургский институт информатики и автоматизации Российской академии наук,
14 линия, 39, 199178, Санкт-Петербург, Россия

Аннотация

В статье описывается подход к решению задачи сопоставления профилей пользователей разных социальных сетей и идентификации тех из них, которые принадлежат одному человеку. Предложен соответствующий метод, основанный на сопоставлении социального окружения и значений атрибутов профиля аккаунтов в двух разных социальных сетях. Проведено сравнение результатов применения различных моделей машинного обучения к решению данной задачи. Новизна подхода заключается в предложенном новом комбинировании различных методов и приложении к новым социальным сетям. Практическая значимость исследования заключается в автоматизации процесса определения принадлежности профилей в различных социальных сетях одному пользователю. Данные результаты могут быть применены в задаче построения мета-профиля пользователя информационной системы для последующего построения профиля его уязвимостей, а также в других исследованиях, посвящённых социальным сетям.

Ключевые слова: *социальные сети, идентификация пользователя, социоинженерные атаки, машинное обучение, информационная безопасность, защита пользователя, профиль уязвимостей пользователя.*

Цитирование: Корепанова А. А., Олисеенко В. Д., Абрамов М. В., Тулупьев А. Л. Применение методов машинного обучения в задаче идентификации аккаунтов пользователя в двух социальных сетях // Компьютерные инструменты в образовании. 2019. № 3. С. 29–43. doi:10.32603/2071-2340-2019-3-29-43

1. ВВЕДЕНИЕ

Актуальность вопросов информационной безопасности только растет с течением времени: согласно отчету IBM, за последние пять лет потери от утечек данных воз-

*Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2019-0003 и при финансовой поддержке РФФИ (гранты № 18-01-00626, № 18-37-00323, № 20-07-00839).

росли на 12 % и составляют в среднем 3.92 миллионов долларов [1]. Стоит отметить, что программно-технические аспекты обеспечения безопасности информационных систем достаточно хорошо изучены и продолжают совершенствоваться. Тем не менее, убытки компаний от инцидентов информационной безопасности продолжают расти [2]. Причина продолжающегося роста видится, в том числе, в прогрессирующем объеме проводимых социоинженерных атак, который отмечается экспертами. Например, согласно отчету Proofpoint по информационной безопасности в сфере здравоохранения, только число полученных компаниями здравоохранения электронных писем от злоумышленников увеличилось на 300 % по сравнению с предыдущим кварталом [3]. Социоинженерные атаки в данной статье рассматриваются как набор прикладных психологических и аналитических приемов, которые злоумышленники применяют для скрытой мотивации пользователей публичной или корпоративной сети к нарушениям устоявшихся правил и политик в области информационной безопасности [4]. Такие атаки направлены на использование уязвимостей пользователя информационной системы, а не программно-технического комплекса и, согласно исследованиям компаний, занимающихся информационной безопасностью, являются трендом в мире современных киберугроз [5]. Это определяет важность исследований в области защиты пользователей информационных систем от социоинженерных атак.

Одна из задач в этом срезе, лежащем на пересечении областей исследований информационной безопасности, искусственного интеллекта и теоретических основ информатики, — автоматизированный анализ защищенности пользователей информационных систем от социоинженерных атак. В рамках этого направления исследований уже существуют наработки для оценки защищенности пользователей информационных систем от социоинженерных атак [6–11]. В соответствии с предложенными подходами, указанные оценки в существенной степени зависят от профиля уязвимостей пользователя [12]. Одним из источников информации для построения профиля уязвимостей пользователя опосредованно является его профиль в социальных сетях (термины «профиль» и «аккаунт» в данной статье используются как синонимы) [13]. Зачастую люди имеют несколько профилей в разных социальных сетях [14]. Чем больше профилей одного человека в разных социальных сетях нам доступно, тем больше информации для оценки выраженности его уязвимостей можно извлечь. В России на данный момент пользуются наибольшей популярностью несколько социальных сетей, каждая из которых имеет свою направленность и специфику размещаемого контента, а значит, может содержать различные фрагменты информации, необходимые для построения профиля уязвимостей пользователя. К сожалению, не всегда легко идентифицировать все страницы пользователя в разных социальных сетях. Вследствие чего актуальной видится задача автоматизированной идентификации профилей пользователя в разных социальных сетях. Одним из этапов решения этой задачи является сопоставление двух профилей. Данная статья посвящена подходу к решению этой подзадачи, которая состоит в сопоставлении двух профилей пользователей социальных сетей «ВКонтакте» и «Одноклассники» и в определении, принадлежат ли они одному человеку. Эти социальные сети были выбраны, поскольку пересечение их аудитории по некоторым оценкам составляет 19 млн пользователей [14].

Статья имеет дидактическую и научную цели, что влияет на её организацию и подход к изложению результатов. Дидактическая цель статьи заключается в том, чтобы продемонстрировать пример задачи, которая может быть решена методами компьютерных наук и анализа данных (классификации данных, корреляционного анализа и т. д.). На-

учная цель состоит в том, чтобы разработать подход сопоставления профилей пользователей разных социальных сетей и идентификации тех из них, которые принадлежат одному человеку.

2. ФОРМУЛИРОВКА ЗАДАЧИ

Задачу определения принадлежности профилей в разных социальных сетях одному пользователю можно свести к задаче бинарной классификации со следующим условием: пусть X — множество пар профилей пользователей социальных сетей «ВКонтакте» и «Одноклассники», а Y — множество классов $\{0; 1\}$, где 0 означает, что пара профилей не принадлежит одному пользователю, а 1 — что принадлежит (таблица 1). Необходимо построить алгоритм $a : X \to Y$, который будет способен классифицировать любой $x \in X$, основываясь на анализе информационных следов пользователей профилей. Пользователь социальной сети, как правило, оставляет в ней довольно много информации о себе: подписки на группы, посты на стене, фотографии, аудиозаписи — все эти данные так или иначе могут его охарактеризовать, но не все можно достаточно просто обрабатывать посредством машинного анализа. В этой работе рассматриваются анкетные данные пользователя и его социальное окружение, представленное списком друзей, родственников и семейным положением, в качестве признаков классификации. Отметим, что решение задачи сопоставления профилей является этапом в задаче поиска профилей одного человека в различных социальных сетях. Вследствие этого, вероятно, будет возникать необходимость работать на множестве профилей, полученных в результате поиска в социальной сети по определённым атрибутам и, соответственно, имеющих близкие значения этих атрибутов профиля. Таким образом, задача усложняется необходимостью находить менее очевидные различия между в целом похожими профилями.

Таблица 1. Пример множества

X (пары профилей)		**Y** (принадлежат одному человеку)
Профиль 1	Профиль 2	0
Профиль 1	Профиль 3	1
Профиль 2	Профиль 4	0
…	…	…

Кроме того, задача автоматического сопоставления профилей характеризуется неполнотой, нечисловым характером и нечёткостью информации, указанной в профилях пользователей. Данные, которые пользователь оставляет о себе в социальных сетях, могут быть неполными, устаревшими, заведомо ложными, содержать ошибки и опечатки и т. д. Заполнение одних и тех же атрибутов профиля пользователя в разных социальных сетях может совпадать семантически, но значительно отличаться синтаксически и по морфемному составу. В одной социальной сети человек может не обновить место работы после смены или вовсе не указывать его, если, например, использует социальную сеть только для общения с родственниками. Также пользователи иногда нарочно публикуют ложные данные, например, уменьшают свой возраст. Все эти факторы обуславливают сложность в выборе методов сопоставления различных полей анкеты пользователя, а также в выборе модели определения вероятности того, что профили принадлежат одному и тому же пользователю.

3. РЕЛЕВАНТНЫЕ РАБОТЫ

Задача идентификации пользователей в различных социальных сетях не нова. За последние годы были опубликованы результаты множества исследований, посвященных этой и смежным проблемам. Так, были предложены методы сопоставления и восстановления пропущенных значений атрибутов, нахождения профилей одного и того же пользователя в разных социальных сетях [15–21].

Большинство подходов по восстановлению пропущенных данных основывается на алгоритмах извлечения информации из социального окружения, то есть друзей пользователя, а также групп и сообществ, в которых он состоит. Так, например, в публикациях [15, 16], посвящённых восстановлению основного места проживания пользователя, проводится обзор и сравнение методов восстановления по социальному окружению — по случайному другу, по распространению меток, на основе машинного обучения (сравнения векторов узлов социального графа). В [17] показана взаимосвязь значений атрибутов пользователя и его друзей с группами, в которых он состоит. Работы [18, 19] посвящены методам автоматического определения возраста пользователей с помощью социальных связей.

В зарубежной литературе представлено множество работ по данным тематикам. Так, например, в статьях [20, 21] были предложены подходы для нахождения профилей пользователя в различных социальных сетях и также разработан фреймворк для создания единого профиля (FOAM). Стоит отметить, что в большинстве указанных работ методы реализованы для таких социальных сетей, как «Facebook», «Twitter», «Instagram» и др. Обозначенные социальные сети имеют свою специфику, которая накладывает ограничения на применимость и адаптацию предложенных методов к социальным сетям «Одноклассники» и «ВКонтакте».

Таким образом, вопрос автоматизированной идентификации профилей пользователей в социальных сетях «ВКонтакте» и «Одноклассники» является открытым. Решение данного вопроса в рамках тематики социоинженерных атак позволило бы создавать единый мета-профиль пользователя, аккумулирующий большой объем информации о нём. Единый мета-профиль пользователя будет опосредованно способствовать построению профиля уязвимостей пользователя [22] и выработке рекомендаций по выбору более безопасных политик поведения пользователя в социальных сетях.

4. МЕТОДИКА РЕШЕНИЯ ЗАДАЧИ

В данной работе профиль пользователя представляется как набор значений атрибутов (фамилия, имя, город, возраст и т. д). Оценка подобия пары профилей основывается на результатах сопоставления значений соответствующих атрибутов каждого профиля. То есть в качестве признаков классификации выступают числовые характеристики схожести значений соответствующих атрибутов. Таким образом, задача разбивается на несколько крупных этапов: формирование датасета, выбор ключевых атрибутов, предобработка данных, восстановление неполной информации, сопоставление профилей через сопоставление значений атрибутов, применение методов бинарной классификации. Рассмотрим каждый из этих этапов подробнее.

4.1. Формирование датасета

Для решения задачи идентификации аккаунтов пользователей были собраны данные 800 различных открытых профилей социальных сетей «Вконтакте» и «Однокласс-

ники». Из этих данных сформировано множество пар X следующим образом: 30 % пар содержат профили, принадлежащие одному пользователю (относящиеся к классу 1), профили из остальных 70 % принадлежат разным пользователям (относятся к классу 0), но значения атрибутов «фамилия» пользователей в паре одинаковы, а значения атрибута «имя» схожи с точностью до словоформы (например Катя, Катерина, Екатерина и т. п.). Итоговый датасет включал в себя набор из 500 пар.

4.2. Выбор атрибутов

Профиль пользователя в социальной сети можно представить как набор значений атрибутов. В этой работе в качестве атрибутов рассмотрим поля публичной анкеты и социальное окружение пользователя, представленное списком друзей. Эти наборы различны для «ВКонтакте» и «Одноклассники», и не все их элементы просты для машинного анализа.

Для дальнейшей работы выделим множество ключевых атрибутов, по значениям которых будет производиться сопоставление профилей. Основными критериями отбора ключевых атрибутов является доступность информации и ее согласованность. Под доступностью информации подразумевается, что значение атрибута должно присутствовать в большинстве профилей. А под согласованностью — соответствие значений атрибута в различных социальных сетях.

Собранный датасет был проанализирован, в нем были выделены некоторые особенности. Ниже представлена статистика заполненности значениями атрибутов, присутствующих в обеих социальных сетях, по собранному датасету (таблица 2).

Таблица 2. Статистика заполненности значений атрибутов

Атрибут	Процент заполнивших
Имя	100,0%
Фамилия	100,0%
Друзья	99,2%
Фотография	88,9%
Город проживания	79,0%
Дата рождения (всего)	60,5%
Дата рождения (без указания года)	33,3%
Родной город	31,1%
Образование	24,3%
Карьера	19,6%
Семейное положение	18,8%
Девичья фамилия	11,2%
Любимая музыка	9,3%
Любимые фильмы	8,4%
Любимые книги	7,2%
Любимые игры	6,1%

В соответствии с описанными ранее критериями к выбору признаков для сопоставления профилей пользователей используются следующие атрибуты, в наибольшей сте-

пени удовлетворяющие описанным критериям: «имя», «фамилия», «друзья», «город проживания», «дата рождения».

4.3. Предобработка данных

Предобработка данных необходима для минимизации влияния факторов, снижающих их качество и мешающих работе алгоритмов сопоставления. Подготовка данных в текущей работе проводилась для атрибутов «имя», «фамилия» и «город проживания». Предобработка включала в себя перевод букв в нижний регистр во всех полях, обратную транслитерацию там, где это было необходимо, исключение специальных символов. Например, значения атрибутов «имя» и «фамилия» могут быть заданы как «IVANOV Vasya!», «ИВАНОВ Vasya» и т. п. Данное значение после предобработки будет модернизировано в «иванов вася».

Также необходимо унифицировать форму представления данных, которая может быть различной в зависимости от социальной сети. Так, например, в значении атрибута «город проживания» в «Одноклассниках» может быть указана административная единица, к которой относится населённый пункт, в «ВКонтакте» такая информация не указывается, что негативно может повлиять на сопоставление этого поля, поэтому она удаляется.

4.4. Восстановление неполной информации

В социальной сети «Вконтакте» из выбранных атрибутов обязательными для заполнения являются «имя», «фамилия», в «Одноклассниках» — «имя», «фамилия», «дата рождения», «город проживания». Но даже они могут быть скрыты гибкими настройками безопасности и анонимности. Как видно в таблице выше, в собранном датасете у 21 % пользователей в публичной анкете не заполнено поле «город», у 39.5 % — «дата рождения», что нарушает первый критерий отбора ключевых атрибутов, усложняет задачу идентификации и ведёт за собой проблему применимости моделей машинного обучения, так как не все они работают с неполной информацией. Хорошая новость состоит в том, что отсутствующая информация может быть в некоторых случаях восстановлена по цифровым следам пользователя через анализ его социального окружения, подписок и т. д. [23].

Вместо атрибута «дата рождения» в дальнейшем используется атрибут «возраст». Необходимость в этом возникает в связи с тем, что день и месяц рождения в случае их отсутствия в профиле восстановить достаточно сложно. Значение атрибута «возраст» формируется посредством значения атрибута «дата рождения», если он задан. В противном случае значение данного атрибута восстанавливается посредством анализа информации о социальном окружении пользователя.

В данной работе восстановление значений атрибутов «город» и «возраст» с помощью расчета моды городов проживания и возрастов друзей пользователя. То есть у пользователя с пропущенным значением атрибута «город» или «возраст» анализируются их значения у друзей [24]. Вместо пропущенного значения соответствующего атрибута ставится значение моды. Оценка точности данного подхода проводилась при помощи метрики accuracy Accuracy = $\frac{T}{N}$, где T — количество правильно восстановленных значений, а N — общее количество элементов. Данный метод показал 76 % и 69 % точности соответственно на собранном датасете. При оценке точности восстановления значения атрибута «возраст» делалось допущение отклонения, абсолютная величина которого не превысила 3. Если восстановленное значение попадало в данный интервал, то оно считалось восстановленным верно.

4.5. Сопоставление атрибутов

Сопоставление профилей производится через сопоставление соответствующих значений атрибутов. Сопоставление различных значений атрибутов профиля требует применения индивидуального подхода. Например, атрибуты «фамилия» и «город проживания» являются строковыми, при их заполнении пользователи могут допускать ошибки и опечатки, поэтому для сопоставления значений этих атрибутов используется метод нечёткого сопоставления — метрика строковой схожести Джаро–Винклера [25] $d_j = \frac{1}{3}(\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m})$, где m — число совпадающих слов, s_1 и s_2 — длины сравниваемых строк, а t — число перестановок. Два символа считаются совпадающими, если расстояние между ними не превышает $L = [\max \frac{(s_1,s_2)}{2}] - 1$. Данная метрика хорошо зарекомендовала себя для аналогичных задач в работах [26, 27]. Результат сопоставления — число в интервале [0; 1].

Значения атрибута «возраст» являются числовыми. В случае, когда дата рождения указана пользователем, применяется точное сопоставление возрастов. Если возраст был восстановлен посредством расчета моды возрастов друзей, два возраста считаются совпадающими, если абсолютная разность меньше или равна трём.

Атрибут «имя» является строковым, но две формы одного и того же имени (Евгений — Женя) могут быть слишком далеки друг от друга при посимвольном сопоставлении, чтобы целиком полагаться на метрику Джаро — Винклера. Для сопоставления значения этого атрибута используется словарь имён, в котором полным именам поставлены в соответствие их краткие формы. Если, согласно словарю, имена совпадают, результатом сопоставления будет 1, если нет — 0, если какое-то из имён отсутствует в словаре, применяем метрику Джаро — Винклера.

Для сопоставления значений атрибута «друзья» в качестве меры социальной схожести был выбран коэффициент Браун-Бланке. Данный выбор коэффициента обоснован высокой корреляцией между ним и целевым значением $y \in Y$. $K_{0,1} = \frac{P_1 \cap P_2}{\max(|P_1|,|P_2|)}$, где P_1, P_2 — множества всех друзей профиля 1 и 2 соответственно, а $|P_1|, |P_2|$ — их мощность. Пересечение друзей находится через их поимённое сопоставление.

В результате поатрибутного сопоставления профилей каждой пары из X получается набор оценок схожести для каждого из сопоставляемых значений — число, лежащее в интервале от 0 до 1.

4.6. Классификация

Для прикидочной оценки зависимости числовых характеристик схожести атрибутов («имя», «фамилия», «город», «друзья») к классам из множества Y (целевая переменная) применяется корреляция Спирмана. Выбор данной корреляции обусловлен тем, что целевая переменная (Y) является дихотомической, а остальные признаки — интервальные. Для оценки корреляции между признаками «возраст» и целевой переменной используется коэффициент корреляции φ, так как он является дихотомическими. Результаты представлены в таблице 3.

Таблица 3. Корреляции признаков

	Возраст	**Имя**	**Фамилия**	**Город**	**Браун-Бланке**	**ИФГ**
Один пользователь	0.287	0.252	0.191	0.643	0.799	0.654

Из таблицы 3 видно, что корреляция между признаками коэффициентов подобия имени, фамилии и целевой величиной довольно низкая, поэтому имеет смысл объединить их с признаком города («ИФГ»), посчитав новый признак как среднее арифметическое от них. Между новым признаком («ИФГ») и целевой велечиной Y выявлена высокая корреляция.

Для построения модели классификации в рамках данной работы применяются три метода бинарной классификации: логистическая регрессия, метод опорных векторов, решающее дерево CART [28].

Для оценки качества классификатора используется ROC-кривая, которая показывает, как зависит TPR (True Positive Rate — число объектов, правильно отнесенных к классу одинаковых профилей) от FPR (False Positive Rate — число объектов, не одинаковых профилей, которые были неправильно отнесены алгоритмом к классу одинаковых профилей).

Для получения числовой характеристики ROC-кривой используется площадь под графиком AUC (Area Under Curve): чем ближе ее значение к 1, тем лучше получится классификатор [29]. В примере на рисунке 1 представлена ROC-кривая и её значение для предложенных алгоритмов. Стоит отметить, что при построении ROC-кривой и её AUC была использована процедура 4-fold скользящего контроля, которая позволяет получить оценку обобщающей способности выбранных моделей. На рисунке 1 представлены результаты построения ROC-кривой и её AUC.

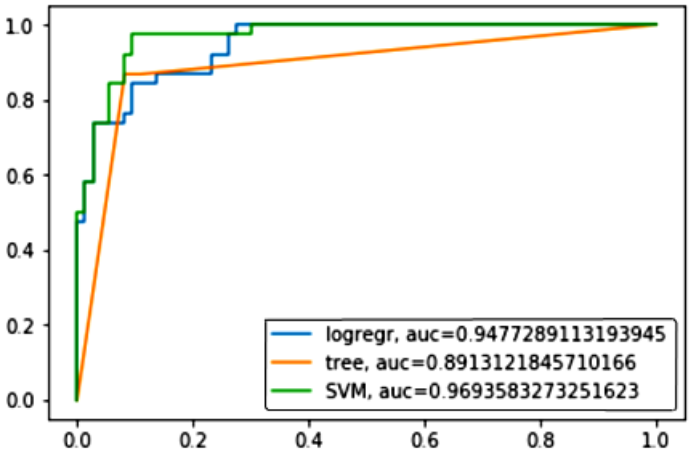

Рис. 1. Сравнение моделей

4.7. Использованные инструменты

Работа была выполнена при помощи языка Python 3.7 и ряда пакетов для него. Список средств и библиотек, используемых при разработке:

- Язык Python 3.7 [30];
- Среда разработки Jupyter Notebook [31];
- Библиотека Pandas 0.25.3 и NumPy 1.17 — для удобного анализа данных [32, 33];
- Библиотека vk-api 11.6.0 и python-odnoklassniki — для получения данных из социальных сетей [34, 35];
- Библиотека Matplot 3.1.1 — для создания графиков [36].

4.8. Выводы

Согласно анализу собранного датасета (таблица 2), можно сделать вывод что профили пользователей характеризуются слабой заполненностью. При восстановлении значений атрибутов (город, возраст) основными причинами ошибок являются два фактора: во-первых, пользователи могут установить себе ложный город проживания, выбирая, например, административный центр своего района, если живут в близлежащей деревне; во-вторых, малое количество друзей приводит к сильному снижению вероятности верного восстановления. Также стоит отметить, что задачу определения принадлежности профиля в различных социальных сетях одному пользователю действительно можно свести к задаче бинарной классификации и применить методы машинного обучения для её решения. По полученным результатам эксперимента можно сделать вывод, что все модели показывают хорошие результаты на текущем наборе данных, но модель на основе метода опорных векторов выигрывает за счёт большей устойчивости к выбросам. Также стоит отметить полученный алгоритм для обучения модели машинного обучения на рисунке 2.

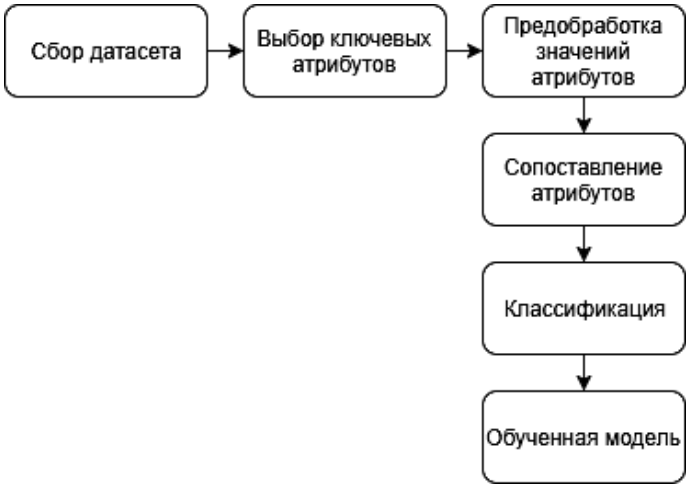

Рис. 2. Алгоритм построения модели

5. ЗАКЛЮЧЕНИЕ

В данной работе предложен метод для определения профилей, принадлежащих одному и тому же пользователю в социальных сетях «ВКонтакте» и «Одноклассники». Выработанный подход основан на сопоставлении социального окружения и значений атрибутов профиля. Рассмотрен ряд частных подзадач: отобраны ключевые атрибуты для сопоставления, применены методы восстановления отсутствующих данных, проведено сравнение моделей классификации. Обучены модели машинного обучения, которые могут быть использованы для дальнейшей автоматизации процесса нахождения профилей одного пользователя в различных социальных сетях. Теоретическая значимость работы заключается в получении обученной модели машинного обучения (рис. 2) для определения принадлежности профилей в различных социальных сетях одному пользователю. Практическая значимость заключается в автоматизации процесса определения принадлежности профилей в различных социальных сетях одному пользователю. Данные результаты могут быть применены в задаче построения мета-профиля пользователя ин-

формационной системы для последующего составления профиля его уязвимостей, а также в других исследованиях, посвящённых социальным сетям.

Список литературы

1. *Cassy L.* IBM Study Shows Data Breach Costs on the Rise; Financial Impact Felt for Years. URL: https://newsroom.ibm.com/2019-07-23-IBM-Study-Shows-Data-Breach-Costs-on-the-Rise-Financial-Impact-Felt-for-Years (дата обращения: 30.08.2019).
2. *Хайрук С.* Мировой объем утечек информации в 2017 году увеличился более чем в четыре раза по данным аналитического центра компании. URL: https://www.infowatch.ru/company/presscenter/news/20235 (дата обращения: 30.08.2019).
3. 2019 Healthcare Threat Report Protecting Patients, Providers and Payers. URL: https://www.proofpoint.com/us/resources/threat-reports/healthcare-threat-report (дата обращения: 30.08.2019).
4. *Азаров А. А., Тулупьева Т. В., Суворова А. В., Тулупьев А. Л., Абрамов М. В., Юсупов Р. М.* Социо-инженерные атаки. Проблемы анализа. М.: Наука, 2016. 352 с.
5. 2019 Phishing Trends & Intelligence Report: The Growing Social Engineering Threat. URL: https://securityboulevard.com/2019/04/2019-phishing-trends-intelligence-report-the-growing-social-engineering-threat/ (дата обращения: 30.08.2019).
6. *Абрамов М. В., Азаров А. А., Фильченков А. А.* Распространение социоинженерной атаки злоумышленника на пользователей информационной системы, представленных в виде графа социальных связей пользователей // Сборник докладов Международной конференции по мягким вычислениям и измерениям (SCM–2015). 2015. С. 329–332.
7. *Khlobystova A. O., Abramov M. V., Tulupyev A. L.* An approach to estimating of criticality of social engineering attacks traces // Recent Research in Control Engineering and Decision Making. ICIT 2019. Vol. 199. Studies in Systems, Decision and Control. С. 446–456. doi: 10.1007/978-3-030-12072-6_36
8. *Тулупьев А. Л., Пащенко А. Е., Азаров А. А., Тулупьева Т. В.* Визуальный инструментарий для построения информационных моделей комплекса «Инфомационная система — персонал», использующихся в имитации социоинженерных атак // Труды СПИИРАН: SPIIRAS Proceedings. 2010. С. 231–245.
9. *Азаров А. А., Тулупьев А. Л., Соловцов Н. Б., Тулупьева Т. В.* SQL-представление реляционно-вероятностных моделей социо-инженерных атак в задачах расчета агрегированных оценок защищенности персонала информационной системы с учетом весов связей между пользователями // Труды СПИИРАН: SPIIRAS Proceedings. 2013. С. 41–53.
10. *Абрамов М. В.* Автоматизация анализа социальных сетей для оценивания защищённости от социоинженерных атак // Автоматизация процессов управления. 2018. №1(51). С. 34–40.
11. *Suleimanov A., Abramov M. V., Tulupyev A. L.* Modelling of the social engineering attacks based on social graph of employees communications analysis // Proceedings — 2018 IEEE Industrial Cyber-Physical Systems, ICPS 2018. Institute of Electrical and Electronics Engineers Inc., 2018. P. 801–805. doi: 10.1109/ICPHYS.2018.8390809
12. *Азаров А. А., Абрамов М. В., Тулупьева Т. В., Тулупьев А. Л.* Анализ защищённости групп пользователей информационной системы от социоинженерных атак: принцип и программная реализация // Компьютерные инструменты в образовании. 2015. № 4. С. 52–60.
13. *Абрамов М. В., Тулупьев А. Л., Тулупьева Т. В.* Агрегирование данных из социальных сетей для восстановления фрагмента мета-профиля пользователя // Шестнадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ-2018 Труды конференции: в 2-х томах. 2018. С. 189–197.
14. Статистика социальных сетей в России на 2018 год. URL: https://hiconversion.ru/blog/statistika-socialnyh-setej-v-rossii-na-2018-god/ (дата обращения: 30.08.2019).
15. *Трофимович Ю. С., Козлов И. С., Турдаков Д. Ю.* Подходы к определению основного места проживания пользователей социальных сетей на основе социального графа // Труды ИСП РАН. 2016. № 6. URL:https://cyberleninka.ru/article/n/podhody-k-opredeleniyu-osnovnogo-

mesta-prozhivaniya-polzovateley-sotsialnyh-setey-na-osnove-sotsialnogo-grafa (дата обращения: 30.08.2019).
16. *Кавеева А. Д., Гурин К. Е.* Локальные сети дружбы «ВКонтакте»: восстановление пропущенных данных о городе проживания пользователей // Мониторинг общественного мнения: общественные и социальные преремены. 2018. № 3. С. 78–90. URL: https://cyberleninka.ru/article/n/lokalnye-seti-druzhby-vkontakte-vosstanovlenie-propuschennyh-dannyh-o-gorode-prozhivaniya-polzovateley (дата обращения: 30.08.2019).
17. *Гурин К. Е.* Структурирование сетей дружбы в онлайн-сообществах СМИ // Дискуссия. 2016. № 6. С. 64–71. URL: https://cyberleninka.ru/article/n/strukturirovanie-setey-druzhby-v-onlayn-soobschestvah-smi (дата обращения: 30.08.2019).
18. *Гомзин А. Г., Кузнецов С. Д.* Метод автоматического определения возраста пользователей с помощью социальных связей // Труды ИСП РАН. 2016. Т. 28. Вып. 6. С. 171–184. doi: 10.15514/ISPRAS-2016-28(6)-12
19. *Грезин В. С., Новосядлый В. А* О проблеме определения возраста участника социальной сети // Известия вузов. Северо-Кавказский регион. Серия: Естественные науки. 2015. № 1 (185). С. 12–18. URL: https://cyberleninka.ru/article/n/o-probleme-opredeleniya-vozrasta-uchastnika-sotsialnoy-seti (дата обращения: 30.08.2019).
20. *Paridhi J., Ponnurangam K., Anupam J.* @I seek 'fb.me': Identifying Users Across Multiple Online Social // 2013 Companion: Proceedings of the 22nd International Conference on World Wide Web. NY, USA: ACM, 2013. pp. 1259–1268. doi: 10.1145/2487788.2488160
21. *Raad E., Chbeir R., Dipanda A.* User profile matching in social networks // Network-Based Information Systems (NBiS). (Sep. 2010). Japan. 2010. P. 297–304. doi: 10.1109/NBiS.2010.35
22. *Абрамов М. В., Азаров А. А., Тулупьева Т. В., Тулупьев А. Л.* Модель профиля компетенций злоумышленника в задаче анализа защищённости персонала информационных систем от социоинженерных атак // Информационно-управляющие системы. 2016. № 4. С. 77–84
23. *Слёзкин Н. Е., Абрамов М. В., Тулупьева Т. В.* Подход к восстановлению мета-профиля пользователя информационной системы на основании данных из социальных сетей // Сборник научных трудов Первой Всероссийской научно-практической конференции «Нечёткие системы и мягкие вычисления. Промышленные применения» (14–15 ноября, 2017). Ульяновск: УлГТУ, 2017. С. 394–399.
24. *Абрамов М. В., Слезкин Н. Е., Тулупьева Т. В.* Агрегация данных из социальных сетей для определения наиболее вероятной конфигурации пропущенных значений параметров мета-профиля пользователя // Сборник докладов Международной конференции по мягким вычислениям и измерениям (SCM-2018). СПб, 2018. Т. 1. С. 118–121.
25. *Winkler W. E.* String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage // Proceedings of the Section on Survey Research Methods (American Statistical Association). 1990. P. 354–359.
26. *Коршунов А., Белобородов И. К., Бузун Н.* Анализ социальных сетей: методы и приложения // Труды ИСП РАН, 2014. Т. 26. Вып. 1. С. 439–456.
27. Патент РФ № 2011145077/08, 08.11.2011. Способ интеграции профилей пользователей онлайновых социальных сетей // Патент России № 2011145077. 2011. Бюл. № 8 / Бартунов С. О., Коршунов А. В., Турдаков Д. Ю. и др.
28. *Breiman L., Friedman J. H., Olshen R. A., Stone C.T.* Classification and Regression Trees. Wadsworth. Belmont. California. 1984.
29. *Zweig M. H., Campbell G.* Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine // Clinical chemistry. 1993. Vol. 39. № 4. P. 561–577.
30. *Pranskevichus E.* What's New In Python 3.7. URL: https://docs.python.org/3.7/whatsnew/3.7.html (дата обращения: 30.08.2019).
31. Project Jupyter. URL: https://jupyter.org/about (дата обращения: 30.08.2019).
32. Easy-to-use data structures and data analysis tools for the Python programming language. URL: https://pandas.pydata.org/index.html (дата обращения: 30.08.2019).
33. NumPy — fundamental package for scientific computing with Python. URL: https://numpy.org/ (дата обращения: 30.08.2019).

34. Python модуль для написания скриптов для социальной сети ВКонтакте (API wrapper). URL: https://pypi.org/project/vk-api/ (дата обращения: 30.08.2019).
35. Odnoklassniki.ru python API wrapper. URL: https://github.com/alternativshik/python-odnoklassniki (дата обращения: 30.08.2019).
36. Python 2D plotting library Matplotlib. URL: https://matplotlib.org/3.1.1/index.html (дата обращения: 30.08.2019).

Поступила в редакцию 26.07.2019, окончательный вариант — 30.08.2019.

Корепанова Анастасия Андреевна, младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН; студент, бакалавр, СПбГУ, aak@dscs.pro

Олисеенко Валерий Дмитриевич, студент, магистр СПбГУ; стажёр лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН, subster3@gmail.com

Абрамов Максим Викторович, кандидат технических наук, заведующий лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН; доцент кафедры информатики СПбГУ, ✉ mva@dscs.pro

Тулупьев Александр Львович, доктор физико-математических наук, профессор кафедры информатики, СПбГУ; главный научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН, alt@dscs.pro

Computer tools in education, 2019
№ 3: 29–43
http://cte.eltech.ru
doi:10.32603/2071-2340-2019-3-29-43

Application of Machine Learning Methods in the Task of Identifying User Accounts in Two Social Networks

Korepanova A. A.[2,1], junior researcher, aak@dscs.pro
Oliseenko V. D.[1,2], student, subster3@gmail.com
Abramov M. V.[2,1], PhD, Senior Researcher, ✉ mva@dscs.pro
Tulupyev A. L.[1,2], PhD, Dc. Sci., Professor, alt@dscs.pro

[1]Saint Petersburg State University, Universitetskaya nab., 7-9, 199034, Saint Petersburg, Russia
[2]Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 39, 14 Line, 199178, Saint Petersburg, Russia

Abstract

The article describes the approach to solving the problem of comparing user profiles of different social networks and identifying those that belong to one person. An appropriate method is proposed based on a comparison of the social environment and the values of account profile attributes in two different social networks. The results of applying various machine learning models to solving this problem are compared. The novelty of the approach lies in the proposed new combination of various methods and application to

new social networks. The practical significance of the study is to automate the process of determining the ownership of profiles in various social networks to one user. These results can be applied in the task of constructing a meta-profile of a user of an information system for the subsequent construction of a profile of his vulnerabilities, as well as in other studies devoted to social networks.

Keywords: *social networks, user identification, social engineering attacks, machine learning, information security, user protection, user vulnerability profile.*

Citation: A. A. Korepanova, V. D. Oliseenko, M. V. Abramov, and A. L. Tulupyev, "Application of Machine Learning Methods in the Task of Identifying User Accounts in Two Social Networks," *Computer tools in education*, no. 3, pp. 29–43, 2019 (in Russian); doi:10.32603/2071-2340-2019-3-29-43.

Acknowledgements: *This work was partially supported by the by RFBR according to the research projects No. 18-01-00626 No. № 18-37-00323 and Governmental contract (SPII-RAS) No. 0073-2019-0003*

References

1. L. Cassy, *IBM Study Shows Data Breach Costs on the Rise; Financial Impact Felt for Years*. [Online]. Available: https://newsroom.ibm.com/2019-07-23-IBM-Study-Shows-Data-Breach-Costs-on-the-Rise-Financial-Impact-Felt-for-Years
2. C. Hajruk, *According to the analytical center of the company, the volume of information leaks in 2017 increased.* [Online]. Available: https://www.infowatch.ru/company/presscenter/news/20235
3. *2019 Healthcare Threat Report Protecting Patients, Providers and Payers.* [Online]. Available: https://www.proofpoint.com/us/resources/threat-reports/healthcare-threat-report
4. A. A. Azarov, T. V. Tulupyeva, A. V. Suvorova, A. L. Tulupyev, M. V. Abramov, and R. M. Jusupov, *Socioengineering attacks. Problems of analysis*, St Petersburg: Nauka Publ., 2016 (in Russian).
5. *2019 Phishing Trends & Intelligence Report: The Growing Social Engineering Threat*. [Online]. Available: https://securityboulevard.com/2019/04/2019-phishing-trends-intelligence-report-the-growing-social-engineering-threat/
6. M. V. Abramov, A. A. Azarov, and A. A. Filchenkov, "Rasprostranenie socioinzhenernoj ataki zloumyshlennika na pol'zovatelej informacionnoj sistemy, predstavlennyh v vide grafa social'nyh svjazej pol'zovatelej" [Distribution of a social engineering attack by an attacker on users of an information system, presented in the form of a graph of user social connections], in *Sbornik dokladov Mezhdunarodnoj konferencii po mjagkim vychislenijam i izmerenijam* (SCM-2015), 2015, pp. 329–332 (in Russian).
7. A. O. Khlobystova, M. V. Abramov, and A. L. Tulupyev, "An approach to estimating of criticality of social engineering attacks traces," in *Recent Research in Control Engineering and Decision Making. ICIT 2019* (Studies in Systems, Decision and Control, vol. 199), O. Dolinina, A. Brovko, V. Pechenkin, A. Lvov, V. Zhmud, and V. Kreinovich, eds, Springer, 2019, pp. 446–456; doi: 10.1007/978-3-030-12072-6_36
8. A. L. Tulupyev, A. E. Pashhenko, A. A. Azarov, T. V. Tulupyeva, "Vizual'nyj instrumentarij dlja postrienija infofrmacionnjyh modelij kompleksa 'Infomacionnaja sistema — Personal', ispol'zujushhihsja v imitacii socioinzhenernyh atak" [Visual tools for building information models of the complex "information system — personnel" used in simulation of socio-engineering attacks], in *SPIIRAS Proc.*, 2010, pp. 231–245 (in Russian).
9. A. A. Azarov, A. L. Tulupyev, N. B. Solovcov, and T. V. Tulupyeva, "SQL-predstavlenie reljacionno-verojatnostnyh modelej socio-inzhenernyh atak v zadachah rascheta agregirovannyh ocenok zashhishhennosti personala informacionnoj sistemy s uchetom vesov svjazej mezhdu pol'zovateljami" [SQL representation of relational-probabilistic models of socio-engineering attacks in the tasks of calculating aggregate assessments of the security of information system personnel, taking into account the weights of connections between users], in *SPIIRAS Proc.*, 2013, pp. 41–53 (in Russian).

10. M. V. Abramov, “Avtomatizacija analiza social’nyh setej dlja ocenivanija zashhishhjonnosti ot socioinzhenernyh atak” [Automation of the analysis of social networks for assessing security against social engineering attacks], *Avtomatizacija processov upravlenija*, no. 1, pp. 34–40, 2018 (in Russian).
11. A. Suleimanov, M. V. Abramov, and A. L. Tulupyev, “Modelling of the social engineering attacks based on social graph of employees communications analysis,” in *Proc. 2018 IEEE Industrial Cyber-Physical Systems* (ICPS 2018), 2018, pp. 801–805; doi: 10.1109/ICPHYS.2018.8390809
12. A. A. Azarov, M. V. Abramov, T. V. Tulupyeva, and A. L. Tulupyev, “Analiz zashhishhjonnosti grupp pol’zovatelej informacionnoj sistemy ot socioinzhenernyh atak: princip i programmnaja realizacija” [Analysis of the security of user groups of the information system from social engineering attacks: principle and software implementation], *Computer tools in education*, no. 4, pp. 52–60, 2015 (in Russian).
13. M. V. Abramov, A. L. Tulupyev, and T. V. Tulupyeva, “Agregirovanie dannyh iz social’nyh setej dlja vosstanovlenija fragmenta meta-profilja pol’zovatelja” [Social data aggregation to restore a fragment of a user’s meta-profile], in *Shestnadcataja Nacional’naja konferencija po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2018 Trudy konferencii*, 2018, pp. 189–197 (in Russian).
14. *Statistics of social networks in Russia for 2018*. [Online]. Available: https://hiconversion.ru/blog/statistika-socialnyh-setej-v-rossii-na-2018-god/ (in Russian).
15. Ju. S. Trofimovich, I. S. Kozlov, and D. Ju. Turdakov, “Podhody k opredeleniju osnovnogo mesta prozhivanija pol’zovatelej social’nyh setej na osnove social’nogo grafa” [Approaches to determining the main place of residence of users of social networks based on a social graph], in *Proc. ISP RAS*, no. 6, 2016. [Online]. Available: https://cyberleninka.ru/article/n/podhody-k-opredeleniyu-osnovnogo-mesta-prozhivaniya-polzovateley-sotsialnyh-setey-na-osnove-sotsialnogo-grafa (in Russian).
16. A. D. Kaveeva and K. E. Gurin, “Lokal’nye seti druzhby ‘VKontakte’: vosstanovlenie propushhennyh dannyh o gorode prozhivanija pol’zovatelej” [Local networks of friendship ‘VKontakte’: restoration of missing data on the city of residence of users], *Monitoring*, no. 3, 2018. [Online]. Available: https://cyberleninka.ru/article/n/lokalnye-seti-druzhby-vkontakte-vosstanovlenie-propuschennyh-dannyh-o-gorode-prozhivaniya-polzovateley (in Russian).
17. K. E. Gurin, “Strukturirovanie setej druzhby v onlajn-soobshhestvah SMI” [Structuring friendship networks in online media communities], *Diskussija*, no. 6, 2016. [Online]. Available: https://cyberleninka.ru/article/n/strukturirovanie-setey-druzhby-v-onlayn-soobschestvah-smi (in Russian).
18. A. G. Gomzin and S. D. Kuznecov, “Metod avtomaticheskogo opredelenija vozrasta pol’zovatelej s pomoshh’ju social’nyh svjazej” [A method for automatically determining the age of users using social connections], in *Proc. ISP*, 2016, vol. 28, no. 6, pp. 171–184 (in Russian); doi: 10.15514/ISPRAS-2016-28(6)-12
19. V. S. Grezin and V. A. Novosyadly, “O probleme opredelenija vozrasta uchastnika social’noj seti” [About the problem of determining the age of a member of a social network], *Izvestija vuzov. Severo-Kavkazskij region. Serija: Estestvennye nauki*, no. 1, pp. 12–18, 2016. [Online]. Available: https://cyberleninka.ru/article/n/o-probleme-opredeleniya-vozrasta-uchastnika-sotsialnoy-seti (in Russian).
20. J. Paridhi, K. Ponnurangam, and J. Anupam, “@I seek ‘fb.me’: Identifying Users Across Multiple Online Social,” *2013 Companion: proc. of the 22nd Int. Conf. on World Wide Web*, NY, USA: ACM, 2013, pp. 1259–1268; doi: 10.1145/2487788.2488160
21. E. Raad, R. Chbeir, and A. Dipanda, “User profile matching in social networks,” in *Network-Based Information Systems* (NBiS), Japan, Sep. 2010, pp. 297–304; doi: 10.1109/NBiS.2010.35
22. M. V. Abramov, A. A. Azarov, T. V. Tulupyeva, and A. L. Tulupyev, “Model’ profilja kompetencij zloumyshlennika v zadache analiza zashhishhjonnosti personala informacionnyh sistem ot socioinzhenernyh atak” [The model of the competence profile of an attacker in the task of analyzing the security of information systems personnel from social engineering attacks], *Informacionno-upravljajushhie sistemy*, no. 4. pp. 77–84, 2016 (in Russian).
23. N. E. Sljozkin, M. V. Abramov, and T. V. Tulupyeva, “Podhod k vosstanovleniju meta-profilja pol’zovatelja informacionnoj sistemy na osnovanii dannyh iz social’nyh setej” [An approach to

recovering a meta-profile of a user of an information system based on data from social networks], *Sbornik nauchnyh trudov Pervoj Vserossijskoj nauchno-prakticheskoj konferencii «Nechjotkie sistemy i mjagkie vychislenija. Promyshlennye primenenija*, Ul'janovsk, Russia: UlGTU, 2017, vol. 1, pp. 394–399 (in Russian).
24. M. V. Abramov, N. E. Sljozkin, and T. V. Tulupyeva, "Agregacija dannyh iz social'nyh setej dlja opredelenija naibolee verojatnoj konfiguracii propushhennyh znachenij parametrov meta-profilja pol'zovatelja" [Aggregation of data from social networks to determine the most likely configuration of missing values for user meta-profile parameters], in *Sbornik dokladov Mezhdunarodnoj konferencii po mjagkim vychislenijam i izmerenijam* (SCM-2018), Sankt Peterburg, 2018, pp. 118–121 (in Russian).
25. W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," in *Proc. of the Section on Survey Research Methods (American Statistical Association)*, 1990, pp. 354–359.
26. A. V. Korshunov, I. K. Beloborodov, and N. Buzun, "Analiz social'nyh setej: metody i prilozhenija" [Analysis of social networks: methods and applications], in *Proc. of ISP RAS*, 2014, pp. 439–456 (in Russian).
27. RF patent No. 2011145077/08, 08/08/2011. Method for integrating profiles of online social network users. Russian Patent No. 2011145077. 2011. Bull. No. 8. Bartunov S. O., Korshunov A. V., Turdakov D. Yu. et al. (in Russian).
28. L. Breiman, J. H. Friedman, R. A. Olshen, and C. T. Ston, *Classification and Regression Trees*, Belmont, California: Wadsworth, 1984.
29. M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.
30. E. Pranskevichus, *What's New In Python 3.7*. [Online]. Available: https://docs.python.org/3.7/whatsnew/3.7.html
31. *Project Jupyter*. [Online]. Available: https://jupyter.org/about
32. *Easy-to-use data structures and data analysis tools for the Python programming language*. [Online]. Available: https://pandas.pydata.org/index.html
33. *NumPy — fundamental package for scientific computing with Python*. [Online]. Available: https://numpy.org/
34. *Python modul' dlya napisaniya skriptov dlya sotsial'noi seti Vkontakte*, (API wrapper). [Online]. Available: https://pypi.org/project/vk-api/
35. *Odnoklassniki.ru python API wrapper*. [Online]. Available: https://github.com/\alternativshik/python-odnoklassniki
36. *Python 2D plotting library Matplotlib*. [Online]. Available: https://matplotlib.org/\3.1.1/index.html

Received 26.07.2019, The final version — 30.08.2019.

Anastasiya A. Korepanova, junior researcher, Laboratory of Theoretical and Interdisciplinary Problems of Informatics, SPIIRAS; student, SPbU, aak@dscs.pro

Valerii D. Oliseenko, student, SPbU; Intern, Laboratory of Theoretical and Interdisciplinary Problems of Informatics, SPIIRAS, subster3@gmail.com

Maxim V. Abramov, PhD, Senior Researcher, Laboratory of Theoretical and Interdisciplinary Problems of Informatics, SPIIRAS; Associate Professor, Computer Science Department, SPbU, ✉ mva@dscs.pro

Alexander L. Tulupyev, PhD, Dc. Sci., Professor, Computer Science Department, SPbU; Principal Researcher, Laboratory of Theoretical and Interdisciplinary Problems of Informatics, SPIIRAS, alt@dscs.pro