СОПОСТАВЛЕНИЕ АЛГОРИТМОВ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ ПРОФИЛЯ[1]

Корепанова А.А., м.н.с. даб ТиМПИ СПИИРАН, студентка СПбГУ,
aak@dscs.pro

Аннотация

Восстановление пропущенных значений в профиле пользователя, построенном на основе обработки аккаунта в социальной сети, имеет множество применений в рамках исследований, опирающихся на анализ социальных сетей. Решение этой задачи актуально в том числе в рамках автоматизации анализа защищённости пользователей информационной системы от социоинженерных атак. В данной работе представлен новый алгоритм восстановления города проживания пользователя, а также приведено сравнение качества его работы с известными алгоритмами.

Введение

На данный момент задача автоматизированной идентификации пользователей социальных сетях «ВКонтакте» и «Одноклассники» не решена. Её решение поспособствовало бы сбору данных для построения профиля уязвимостей пользователя в рамках исследования, посвящённому автоматизации построения оценок защищённости пользователей от социоинженерных атак [1]. В работе [2] предложен подход к решению данной задачи, основанный на сопоставлении значений атрибутов профилей в социальных сетях. Одной из проблем при работе с данными профилей является возможная нечеткость содержащейся в них информации.

Данные могут быть неполные, неточные. Согласно статистике по собранному в рамках исследования [2] датасету у 100% пользователей профиле содержатся значения только атрибутов «имя» и «фамилия», остальные атрибуты заполнены не у всех: у 21% отсутствует значение атрибута «город», у 39.5% — «дата рождения», у 68,9% — «родной город».

Это усложняет задачу идентификации пользователей в различных социальных сетях, так как многие модели машинного обучения,

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

теоретически применимые в данной задаче, не работают с неполными данными. Существуют различные способы решить проблему неполноты информации при работе с моделями машинного обучения: например, исключение из выборки всех профилей с неполными данными, или подстановка фиксированного значения вместо отсутствующего. Однако, мы не всегда можем довольствоваться ими: чем больше полезной информации мы сможем извлечь из профиля, тем более точное предположение о принадлежности профилей одному пользователю мы сможем сделать, так что возникает необходимость восстанавливать пропущенные значения атрибутов профиля посредством анализа информационных следов пользователя. Существует несколько алгоритмов восстановления пропущенных значений атрибутов профиля пользователя, но ни один не даёт 100% точность, так что в данной работе предложен ещё один алгоритм и проведено сравнение его точности с некоторыми существующими.

Постановка задачи

Данная работа продолжает исследование [3], в рамках которого были рассмотрены алгоритмы восстановления значений атрибутов «город проживания» и «возраст» посредством анализа социального окружения и информации в профиле об образовании. Текущая задача состоит в том, чтобы провести сравнение нескольких методов восстановления пропущенных значений атрибута «город» и определить, возможно ли улучшить результат, полученный в [3], с их помощью.

Рассматриваемые алгоритмы

Предложенный в работе [3] метод был сопоставлен с двумя методами восстановления города проживания: посредством выделения сообществ в социальном графе и посредством анализа подписок аккаунта на группы в социальной сети.

Был рассмотрен алгоритм выделения сообществ, представленный в [4]. Он базируется на анализе структуры графа ближайшего окружения пользователя и атрибутов вершин этого графа

Второй рассматриваемый метод основывается на анализе сообществ, групп и публичных страниц (далее называется обобщенно “группы”), на которые подписан аккаунт. Основная идея этого алгоритма исходит из следующего эвристического предположения: пользователи социальных сетей, проживающие в одном городе, подписываются на одни те же группы: группы с местными новостями, событиями и заведениями. Алгоритм состоит в следующем:

1. Из множества групп, на которые подписан аккаунт, выбираются те, чье количество подписчиков не превышает 100000 аккаунтов.

2. Для каждой такой группы подсчитывается число подписчиков, для которых совпадает значение атрибута “город”. Если число пользователей из одного города больше, чем 20 % от общего числа пользователей, подписанных на группу, то город заносится в потенциальные города проживания пользователя.

3. Предполагается, что город, добавленный в множество потенциальных городов проживания наибольшее число раз, является городом проживания пользователя.

Результаты

Было проведено тестирование алгоритмов восстановления города проживания пользователя на выборке из 1000 аккаунтов с заполненными значениями атрибутов “город”.

	Точность
Анализ социального окружения	0.762
Выделение сообществ	0.560
Подписки	0.731

Таблица 1

Наилучший результат показал исходный алгоритм анализа социального окружения.

Заключение

В работе предложен алгоритм восстановления города проживания пользователя посредством анализа групп, на которые подписан аккаунт. Было проведено тестирование предложенного алгоритма, а также сравнение его с известными решениями. Предложенный алгоритм показал хорошие результаты, хотя и не лучшие по сравнению с другими рассмотренными алгоритмами. Результаты данного исследования планируется применить в дальнейшей работе над алгоритмами идентификации пользователей в различных социальных сетях с помощью алгебраический байесовских сетей [5].

Литература

1. Хлобыстова А.О., Абрамов М.В., Тулупьев А.Л., Золотин А.А. Поиск кратчайшей траектории социоинженерной атаки между парой пользователей в графе с вероятностями переходов // Информационно-управляющие системы. 2018. №6. С. 74-81. doi: 10.31799/1684-8853-2018-6-74-81
2. 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
3. Слёзкин Н.Е., Абрамов М.В., Тулупьева Т.В. Подход к восстановлению мета-профиля пользователя информационной системы на основании данных из социальных сетей // Нечеткие системы и мягкие вычисления. Промышленные применения материалы Первой всероссийской научно-практической конференции. 2017. С. 399-404.
4. Чесноков В.О. Предсказание атрибутов профиля пользователя социальной сети путем анализа сообществ графа его ближайшего окружения // Вестник МГТУ им. Н.Э. Баумана. Серия «Приборостроение». 2017. №2 (113).
5. Kharitonov N.A., Maximov A.G., Tulupyev A.L. Algebraic Bayesian Networks: The Use of Parallel Computing While Maintaining Various Degrees of Consistency // Studies in Systems, Decision and Control, 2019. vol. 199. pp. 696–704. doi: 10.1007/978-3-030-12072-6_56