Метод автоматического определения возраста пользователей с помощью социальных связей

[1,2]А.Г. Гомзин <gomzin@ispras.ru>
[1,2,3]С.Д. Кузнецов <kuzloc@ispras.ru>
[1]Институт системного программирования РАН,
109004, Россия, г. Москва, ул. А. Солженицына, дом 25
[2]Московский государственный университет имени М.В. Ломоносова,
119991 ГСП-1 Москва, Ленинские горы, МГУ имени М.В. Ломоносова, 2-й учебный корпус, факультет ВМК
[3]Московский физико-технический институт,
141700, Московская область, г. Долгопрудный, Институтский пер., 9

Аннотация. Работа посвящена методам определения возраста пользователей социальных сетей. Социальные сети предоставляют пользователям возможность заполнять свои профили, которые могут включать в себя возраст. В связи с тем, что профили заполняются не полностью, возникает задача предсказания неуказанных значений атрибутов пользователей. Явно указанные и предсказанные значения возраста используются в рекомендательных и маркетинговых системах, они позволяют фильтровать целевую аудиторию рекомендуемых товаров и услуг. Кроме того, предсказанные значения могут использоваться для более точного определения демографического профиля интернет-сообществ, целевую аудиторию рекламных кампаний в Интернете. В данной работе предлагается метод, предсказывающий незаполненное значение возраста пользователя. Метод использует следующую доступную информацию из социальной сети: явно указанные пользователями значения возраста и социальные связи. Метод основан на распространении меток по графу друзей и подписок пользователей на сообщества.

Ключевые слова: социальные сети; демографические атрибуты; векторная модель; социальный граф; распространение меток

DOI: 10.15514/ISPRAS-2016-28(6)-12

Для цитирования: Гомзин А.Г., Кузнецов С.Д. Метод автоматического определения возраста пользователей с помощью социальных связей. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 171-184. DOI: 10.15514/ISPRAS-2016-28(6)-12

1. Введение

Профили пользователей социальных сетей как правило включают в себя демографические атрибуты, такие как пол, возраст, семейное положение, уровень образования, религиозные, политические взгляды и т.д. Значения демографических атрибутов используются в рекомендательных и маркетинговых системах. Они позволяют фильтровать целевую аудиторию рекомендуемых товаров и услуг.

По тем или иным причинам далеко не все атрибуты заполняются пользователями. Кроме того, иногда пользователи оставляют в профиле ложные сведения. Данная работа посвящена методам, предсказывающим незаполненные значения возраста, с использованием социальных связей, явно указанным другими пользователями значениями возраста. В качестве исходных данных используются открытые данные социальной сети Вконтакте[1], такие как профили пользователей, друзья пользователей, подписки пользователей на сообщества. Методы поиска ложно указанных атрибутов в данной статье не рассматриваются.

Представленный в работе метод определения демографических атрибутов основан на использовании социального графа. Пользователи и сообщества – узлы данного графа, отношения дружбы между пользователями, подписок на сообщества – ребра. Под сообществом понимается специальная страница в социальной сети, объединяющая пользователей по интересам: пользователи подписываются на интересующие их сообщества для получения релевантной информации. Значения атрибутов предсказываются путем распространения меток в этом графе. Метки представляют собой значения возраста.

Вначале будут описаны существующие методы решения задачи определения демографических атрибутов и смежных задач. Затем описан предлагаемый метод. В конце статьи представлены результаты экспериментального исследования разработанного метода.

2. Краткий обзор существующих решений

В этом разделе представлен обзор решений задач определения демографических атрибутов пользователей Интернета.

Наибольший интерес для исследователей методов определения демографических атрибутов представляют социальные сети, такие как Facebook[2], Twitter[3] и другие. Кроме данных ресурсов, в некоторых исследованиях анализируются комментарии на Youtube [5], новости и электронные письма [2].

[1] https://vk.com
[2] https://www.facebook.com/
[3] https://twitter.com/

Наиболее распространенным подходом, используемым при решении задач определения демографических атрибутов, является извлечение признаков из текстов пользователей и применение к ним методов машинного обучения. Вначале будет описаны признаки, которые использовали авторы работ, затем перечислены используемые алгоритмы.

В работе [5] определяется пол пользователей Youtube. Сначала авторы используют метод распространения пола в графе пользователи-видео, где ребро между пользователем и видео означает факт просмотра видео пользователем. Затем в качестве признаков рассматриваются статистические признаки, такие как средняя длина комментария в символах/словах/предложениях, словесные n-граммы, возраст пользователя, а также распределение пола, полученное с помощью модели распространения атрибута «пол» в графе пользователи-видео. В работе [1] пол пользователей Twitter определяется по текстам их сообщений (твитов). Используются символьные и словесные n-граммы. В работе [8] рассматривается задача определения возраста пользователей, пишущих на голландском языке. Возраст пользователей разбивается на интервалы. В качестве признаков используются символьные и словесные 1,2 и 3-граммы. Помимо решений, в которых множество значений возраста пользователей разбивается на несколько интервалов, существуют методы, которые предсказывают числовое значение возраста [7]. Авторы [3] определяют политические предпочтения пользователей социальной сети Twitter. Рассматриваются три класса: демократы, республиканцы, неявная политическая позиция. В качестве признаков используются словесные юниграммы, хэштеги, сообщества пользователей (полученные с помощью алгоритма, основанного на распространении меток в социальном графе пользователей).

Одним из самых простых используемых алгоритмов является Наивный байесовский классификатор. Этот метод используется в работе [1]. При классификации на два класса часто используется линейный классификатор. Одним из популярных алгоритмов обучения линейного классификатора является метод опорных векторов. Его используют авторы работ [2], [1], [8], [7]. В работах также встречаются такие алгоритмы, как решающие деревья и логистическая регрессия [2]. Для определения числового значения возраста используется линейная регрессия [7].

Более полный обзор методов определения демографических атрибутов пользователей по текстам их сообщений представлен в работе [10].

Помимо текстовой информации в качестве исходных данных для определения демографических атрибутов используются социальные связи. В работе [6] анализируется университетская социальная сеть. Атрибуты определяются с использованием алгоритма кластеризации социального графа методом распространения меток. В работе [4] в качестве исходных данных рассматривается мобильная социальная сеть, в которой связи между

пользователями составляются на основе звонков и коротких сообщений между ними.

Примером использования одновременно двух видов данных (текстов пользователей и социальных связей) является работа [9]. Авторы определяют тональность сообщений пользователей Twitter. При этом строится граф, в котором присутствуют пользователи, сообщения, слова, эмотиконы, используется метод распространения меток в этом графе.

В этой работе будет предложен метод, определяющий возраст пользователей путем распространения меток в графе, включающем в себя пользователей, сообщества и связи между ними.

3. Описание предлагаемого метода

Для работы метода определения демографических атрибутов необходимы следующие данные:

- Профили пользователей, содержащие значения демографических атрибутов
- Социальные связи (достаточно одного из нижеперечисленных видов данных):
    - Списки друзей пользователей
    - Списки подписчиков сообществ

Сначала из профилей пользователей извлекаются значения возраста. Назовем данный процесс разметкой.

Затем неуказанные значения возраста определяются для всех пользователей с использованием социальных связей.

В данном разделе статьи сначала идет описание краулера, т.е. сборщика данных, затем описание разметки атрибутов, затем алгоритм определения неуказанных значений атрибутов.

3.1 Сбор данных

Сбор данных осуществлялся из социальной сети ВКонтакте.

Для сбора использовались методы VK API для разработчиков приложений (https://vk.com/dev/methods).

Сбор информации затрагивает всех пользователей, но не все сообщества. При сборе профилей пользователей, а также скачивании графа дружбы, краулер предварительно получает список идентификаторов всех пользователей из каталога (https://vk.com/catalog.php). Скачивание графа подписок на сообщества осуществляется для 1 миллиона «наиболее активных» сообществ VK. Список таких сообществ был составлен предварительно (до начала сбора данных) путем ранжирования всех доступных на тот момент сообществ по дате наиболее позднего публичного сообщения.

Для сбора **профилей пользователей** используются методы API users.get и groups.getById. Методы принимают на вход списки идентификаторов пользователей или сообществ и возвращают списки их профилей в формате JSON. За один запрос к каждому из методов скачивается 200 профилей.

Для сбора **графов дружбы и подписки** используются методы API friends.get и groups.getMembers. Методы принимают идентификатор одного пользователя или сообщества и возвращают списки идентификаторов его друзей или подписчиков.

Все используемые методы сбора данных используют версию API 5.52.

Реализация краулера данных выполнена на основе фреймворка MODIS Crawler. Данный фреймворк позволяет параллельно осуществлять множество запросов к методам VK API.

3.2 Разметка возраста

Алгоритм определения возраста пользователей использует указанные другими пользователями значения возраста.

Возраст пользователя извлекается из даты его рождения, указанной в профиле. Поле «дата рождения» может быть представлено в трех вариантах:

1. DD-MM-YYYY - доступна полная дата
2. YYYY - доступен год рождения
3. DD-MM - доступна дата без года

Здесь YYYY - год, MM - месяц, DD - день месяца.

В первых двух вариантах известен год рождения. Возраст определяется как:

$$Y_c - Y_u \tag{1}$$

Здесь Y_u - указанный в профиле год, Y_c - текущий год.

3.3 Определение возраста

Система определения возраста авторов определяет неуказанные значения атрибутов пользователей на основе информации о размеченных значениях и социальных связей (граф друзей, граф подписок на сообщества).

Социальный граф состоит из узлов и связей между ними. Узлы бывают двух видов: пользователи и сообщества. Граф включает в себя следующие связи между узлами:

- граф друзей: связи между пользователями (отношение дружбы).
- граф подписчиков: связи между пользователем и сообществом (подписка пользователя на сообщество).

Каждому узлу (пользователю или сообществу) в графе ставится в соответствие набор меток. Каждая метка соответствует определенному значению атрибута (например, «возраст=23»).

Схема алгоритма:

1. Инициализация
2. Построение векторной модели
3. Вычисление весов пользователей и сообществ, распространение меток на узлы-сообщества
4. Построение векторной модели с учетом весов
5. Распространение меток на узлы-пользователей с учетом весов

На этапе инициализации узлы-пользователи получают метки. Далее метки распространяются на узлы-сообщества. Затем метки узлов пользователей и узлов-сообществ распространяются на узлы-пользователей, у которых отсутствует метка (не указан явно возраст).

Распространение меток представляет собой вычисление метки узла на основе меток его соседей в социальном графе. Алгоритм вычисления метки представлен далее.

Инициализация

Изначально узлы-пользователи инициализируются метками в соответствии с разметкой.

Построение векторной модели

Для каждого пользователя строится распределение значений атрибута среди его соседей. На рисунке 1 приведен пример. Затем все распределения группируются по значению атрибута пользователя. После этого для каждого значения возраста вычисляется среднее распределение значений возраста соседей. В итоге получается так называемая векторная модель для возраста. Например, на рисунке 2 изображена векторная модель, в которой для каждого значения возраста задано распределение возрастов соседей.

Обозначим эту модель $Model_{avg}$.

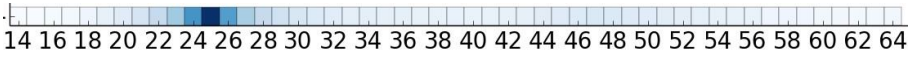

Рис. 1. Распределение значений возраста соседей пользователя. Вероятность обозначена интенсивностью цвета (чем темней, тем больше).

Fig. 1. The distribution of neighbors' ages for the user. The probability is indicated by color strength (the darker - the more).

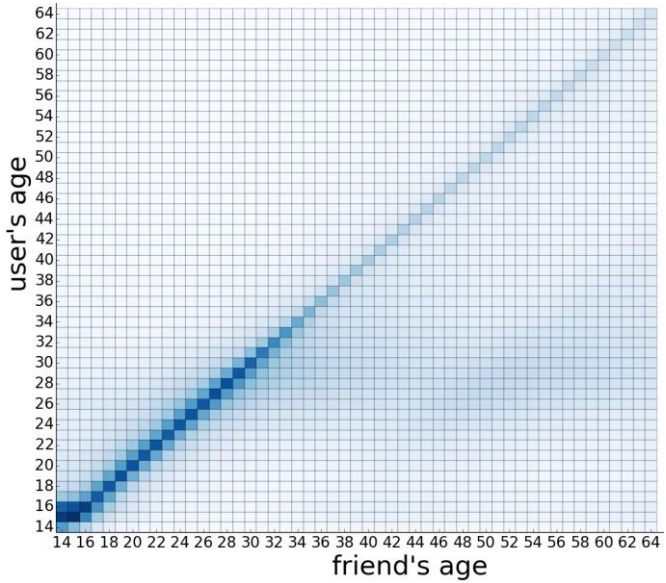

Рис. 2: Векторная модель для атрибута «возраст». В каждой строке – усредненные распределения возраста соседей.

Fig. 2. Vector model for the "age" attribute. Each line is average distribution of neighbors' ages.

Данная векторная модель, в которой производится усреднение распределений по всем данным, применяется при распространении меток в сторону пользователей. Для распространения меток в сторону узлов-сообществ используется модель, в которой распределение значений атрибутов соседей для каждого значения атрибута имеет следующий вид:

$$p(val_n|val_c) = \begin{cases} 1 & \text{если } val_n = val_c \\ 0 & \text{если } val_n \neq val_c \end{cases} \quad (2)$$

Здесь $p(val_n|val_c)$ – вероятность того, что значения атрибута соседа равно val_n, при условии что свое значение атрибута равно val_c.

Обозначим эту модель $Model_{max}$.

Векторные модели используются для оценки близости распределения соседей узла, для которого вычисляется метка, к соответствующему распределению из модели. При использовании модели $Model_{max}$ максимальная близость достигается, когда все метки соседей узла принимают одинаковое значение.

Вычисление весов сообществ, распространение меток на сообщества

На этом этапе алгоритма моделируется распространение меток по социальным связям. При этом для каждого сообщества на основе значений атрибута соседей вычисляются метка (значение атрибута) и вес.

Вес – вещественное число, определяющее, насколько метка (значение атрибута) данного пользователя или сообщества соответствует векторной модели. Его можно интерпретировать как уверенность алгоритма в своем решении.

При определении метки (значения атрибута) узла строится распределение значений данного атрибута у соседей *Distr* (см. рисунок 1), затем для каждого значения атрибута вычисляется близость данного распределения к соответствующему распределению из векторной модели. В качестве меры близости используется косинусная мера. Таким образом, значение метки вычисляется по формуле:

$$L = \arg\max_{val}(sim(Model_*(val), Distr)) \quad (3)$$

$$S = \max_{val}(sim(Model_*(val), Distr)) = sim(Model_*(L), Distr) \quad (4)$$

где:

$$sim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \quad (5)$$

Здесь L — значение атрибута (метка), S — близость, соответствующая максимально близкому значению атрибута *val*, *Distr* — распределение значений атрибута соседей, $Model_*(val)$ — распределение из модели для значения атрибута *val* (для пользователей используется модель $Model_{avg}$, для сообществ – $Model_{max}$).

Значения L и S вычисляются для узлов, у которых множество узлов-соседей с указанным значением атрибута непустое.

Значения S используются для определения весов узлов $W(node)$. Веса вычисляются отдельно для каждого типа узла. Вычисленные значения S сортируются по возрастанию и помещаются в массив. Затем вес узла определяется по формуле:

$$W(node) = \left(\frac{pos(S_{node})}{N}\right)^2 \quad (6)$$

Здесь $pos(S_{node})$ – порядковый номер значения S (от 1 до N) в отсортированном массиве, N – количество узлов, для которых вычислено значение S.

Для узлов, у которых не вычислено значение S вес равен 0.

Построение векторной модели и распространение меток на пользователей с учетом весов

Для каждого типа соседей (пользователи, сообщества) вручную задается вес, задающий вклад каждого источника данных (W_{User}, W_{Comm}, соответственно). Данные веса подбираются при тестировании для каждого атрибута.

После того, как метки распространены на сообщества и определены веса сообществ, запускается распространение меток для пользователей с учетом вычисленных на предыдущем шаге весов $W(node)$ (т.е. вклад каждого сообщества-соседа пользователя в распределение $Distr$ равен его весу). Для каждого типа соседей (пользователь, сообщество) отдельно вычисляется распределение значений атрибута соседей данного типа. Затем это распределение домножается на соответствующий вес (W_{User}, W_{Comm}). Полученная сумма распределений нормализуется.

При распространении меток на данном этапе алгоритма используется модель $Model_{avg}$. После распространения меток незаполненные атрибуты заполняются в соответствие с распространенными метками.

4. Тестирование

Для оценки качества определения демографических атрибутов используется кросс-валидация с разбиением данных на 10 частей. Кросс-валидация запускается при различных значениях параметров W_{User}, W_{Comm}. Каждый из этих параметров при тестировании принимает значения 1, 10 или 100.

В данном разделе сначала описывается выборка, затем метрики качества, затем результаты.

4.1 Выборка

Сначала среди всех сообществ выбираются сообщества, у которых имеется хотя бы K подписчиков с явно указанным значением возраста.

Затем в выборку попадают пользователи, у которых:

- имеется хотя бы K социальных связей: друзей с явно указанным значением атрибута, отобранных сообществ и
- размечено значение возраста

В проведенных экспериментах $K = 10$.

Количество пользователей в выборке: 28940134

4.2 Метрики

Для атрибута *возраст* определяется точность. С увеличением возраста пользователя абсолютная ошибка в предсказании его возраста становится менее критичной, поэтому при оценке точности используется величина относительной ошибки. Считается, что значение *возраста* предсказано верно, если:

$$|age_u - age_p| \leq 0,15 \cdot age_u$$

Здесь age_u – значение *возраста* пользователя из разметки, age_p – предсказанное значение *возраста*.

Для атрибута возраст также вычисляется средняя абсолютная ошибка (MAE):

$$\frac{\sum |age_u - age_p|}{N}$$

Здесь N – количество предсказанных значений.

4.3 Результаты

Тестирование проводилось при различных значениях параметров: W_{User}, W_{Comm}. Рассматриваются конфигурации данных параметров, когда они равны и когда один из них преобладает на порядок. В таблице 1 представлены значения точности и средней абсолютной ошибки.

Таблица 1. Результаты тестирования.

Table 1. Test results.

Значения весов	Метрика	Значение
$W_{User} = 1$, $W_{Comm} = 1$	точность	81,3 %
	MAE	2,79 года
$W_{User} = 1$, $W_{Comm} = 10$	точность	77,6 %
	MAE	3,28 года
$W_{User} = 10$, $W_{Comm} = 1$	точность	81,1 %
	MAE	2,81 года

Из полученной таблицы видно, что наибольший вклад в качество определения возраста приносит граф друзей.

5. Выводы

В данной работе рассматривается задача определения возраста пользователей социальных сетей. Предложен метод, позволяющий определять значения возраста пользователей, у которых имеется хотя бы один из видов данных: список друзей, список подписок на сообщества. Алгоритм основан на распространения меток в социальном графе. Пользователи и сообщества – узлы данного графа, отношения дружбы между пользователями, подписок на сообщества – ребра. Метки представляют собой значения возраста.

В результате тестирования были достигнуты приемлемые показатели качества определения возраста. В дальнейшем планируется применить данный метод к другим атрибутам пользователей, использовать генерируемый текстовый контент и зависимости между значениями различными атрибутами (например, между возрастом и уровнем образования).

Список литературы

[1]. John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[2]. Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[3]. Michael D Conover, Bruno Gon¸calves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 192–199. IEEE, 2011.

[4]. Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 15–24. ACM, 2014.

[5]. Katja Filippova. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488. Association for Computational Linguistics, 2012.

[6]. Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[7]. Dong Nguyen, Noah A Smith, and Carolyn P Ros´e. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[8]. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[9]. Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics, 2011.

[10]. Гомзин А.Г., Кузнецов С.Д. Методы построения социо-демографических профилей пользователей сети Интернет. *Труды ИСП РАН*, том 27, вып. 4, 2015, стр. 129-144. DOI: 10.15514/ISPRAS-2015-27(4)-7

A method of automatically estimating user age using social connections

[1,2]*A.G. Gomzin <gomzin@ispras.ru>*
[1,2,3]*S.D. Kuznetsov <kuzloc@ispras.ru>*
[1] *Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*
[2]*Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia.*
[3]*Moscow Institute of Physics and Technology (State University) 9 Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russia*

Abstract. The work is devoted to methods of social network users' age detection. Social networks allow users to fill their profiles that may contain an age. Profiles are not fully filled, so the task of unknown attributes detection arises. Explicit and predicted values are used in recommender and marketing systems. Moreover, the predicted values can be used for determining online communities' demographic profiles and for inferring the target audience of marketing campaigns in the Internet. In this paper a method for predicting unfilled age values is proposed. The method uses the following information available from the social network: explicit users' ages and social graph. The graph contains nodes representing users and communities. Community is the special page in the Internet that unites users on interests. Friendship relations between users and subscriptions of users on communities represented as edges of the social graph. The method is based on the label propagation in the friendship and subscription graphs. Ages of the users are representd by labels that are propagated in the graph. The scheme of the algorithm is following: initialize user labels according to explicit profiles; build vector model that contains distributions of the neighbours' ages grouped by user age; compute weights of users and communities, propagate labels to communities; build vector model considering calculated weights; propagate labels to users that have not filled their age in the profile. The paper describes the algorithm and contains experimantal results showing that friendship relations are more useful for age prediction in the social network than communities.

Keywords: social media; demographic attributes; vector model; social graph; label propagation

DOI: 10.15514/ISPRAS-2016-28(6)-12

For citation: Gomzin A., Kuznetsov S. A method of automatically estimating user age using social connections. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016, pp. 171-184 (in Russian). DOI: 10.15514/ISPRAS-2016-28(6)-12

References

[1]. John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1301–1309. Association for Computational Linguistics, 2011.

[2]. Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. Digital Investigation, 8(1):78–88, 2011.

[3]. Michael D Conover, Bruno Gon¸calves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pages 192–199. IEEE, 2011.

[4]. Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 15–24. ACM, 2014.

[5]. Katja Filippova. User demographics and language in an implicit social network. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1478–1488. Association for Computational Linguistics, 2012.

[6]. Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In Proceedings of the third ACM international conference on Web search and data mining, pages 251–260. ACM, 2010.

[7]. Dong Nguyen, Noah A Smith, and Carolyn P Ros´e. Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 115–123. Association for Computational Linguistics, 2011.

[8]. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pages 37–44. ACM, 2011.

[9]. Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP, pages 53–63. Association for Computational Linguistics, 2011.

[10]. Gomzin, A.G., Kuznetsov, S.D. Methods for Construction of Socio-Demographic Profile of Internet Users. Trudy ISP RAN/Proc. ISP RAS, vol 27, issue 4, 2015, pp. 129-144 (in Russian). DOI: 10.15514/ISPRAS-2015-27(4)-7.