Round of the social count for definition of people on the social networks possessing the general meaning of any attribute with the set selection of users

Shompolov I.[1], Sidorets R.[2]

Обход социального графа для определения людей в социальных сетях, обладающих общим значением произвольного атрибута с заданной выборкой пользователей

Шомполов И. Г.[1], Сидорец Р. А.[2]

[1]Шомполов Игорь Григорьевич / Shompolov Igor – доктор педагогических наук, кандидат физико-математических наук, преподаватель,
кафедра высшей математики;
[2]Сидорец Роман Андреевич / Sidorets Roman - бакалавр прикладных математики и физики,
соискатель степени магистра,
Московский физико-технический институт, г. Москва

Аннотация: *в работе описан метод поиска множества пользователей в социальной сети «Вконтакте», обладающих общим значением произвольного атрибута. Обладая информацией о заданной выборке пользователей заранее, посредством найденного объединения или же списка группы лиц, обладающих общим значением произвольного атрибута, можно найти большее (отличное от выбранного) множество с той же характеристикой. Данная задача имеет практическое бизнес-применение и изучена недостаточно подробно. В частности, задача поиска потенциальных покупателей того или иного товара, потенциальных пользователей того или иного ресурса, интересантов в тех или иных услугах – у всех этих групп есть общий атрибут (интерес). В работе не рассматривается задача выявления этого атрибута, а рассматривается задача нахождения подмножества пользователей, наделенных данных атрибутом на основе произвольной выборки объединенных данным атрибутом людей. Кроме того, рассмотренный метод сравнен с аналогами, проведены эксперименты, подтверждающие корректность и эффективность метода. Рассмотрена возможность применения метода в области работы с абитуриентами МФТИ или же поиск объединений студентов по тем или иным интересам.*

Abstract: *in work the method of search of a great number of the users on social network possessing a general meaning of any attribute is described. Possessing information about the set selection of the users possessing a general meaning of any attribute it is possible to find a bigger set with the same characteristic. This task has practical business application and is studied insufficiently in detail. In particular, the task of search of potential buyers of these or those goods, potential users of this or that resource, interested parties in these or those services – at all these groups is general attribute (interest). In work the task of identification of this attribute isn't considered, and the task of finding of a subset of users of the allocated data on the basis of any selection of the people united by this attribute is considered by attribute. Besides, the considered method is compared to analogs, the experiments confirming a correctness and efficiency of a method are made. The possibility of application of a method in the field of work from the entrant of MIPT is considered.*

Ключевые слова: *атрибут, социальные сети, выявление, большие числа групп, Вконтакте, анализ, пользователи, покупатели.*

Keywords: *attribute, social networks, identification, large numbers of groups, VKontakte, analysis, users, buyers.*

Анализ социальных данных стремительно набирает популярность во всём мире. В 2016 году у каждого пользователя интернета множество аккаунтов в тех или иных сетях (YouTube, VK, Facebook, Twitter и другие) [1, 2]. Сети включают в себя не только свойства пользователей, такие как имя, пол, дата рождения, но и их принадлежность к тем или иным социальным группам [3, 4]. Группы могут быть основаны на общих интересах, дружественных связях, месторасположении или же месте учебы/работы. Таким образом, социальные сети являются уникальным источником данных о личной жизни и интересах реальных людей [5].

В нашей работе внимание сфокусировано на поиске группы (подмножества) пользователей социальной сети, обладающих общим значением атрибута. Данная задача имеет практическое бизнес-применение [8] и изучена недостаточно подробно. В частности, задача поиска потенциальных покупателей того или иного товара, потенциальные пользователи того или иного ресурса, интересанты в тех или иных услугах – у всех этих групп есть общий атрибут. В работе не рассматривается задача выявления этого атрибута, а рассматривается задача нахождения подмножества пользователей,

наделенных данных атрибутом на основе произвольной выборки объединенных данным атрибутом людей.

В частности, в работе исследовались пользователи социальной сети ВКонтакте (http://vk.com/). В качестве связей, определяющих общий атрибут пользователей, была выбрана информация о принадлежности пользователя к тем или иным группам Вконтакте, пабликам, встречам (далее – группа).

1. Начальными (входными) данными служит произвольное подмножество пользователей социальной сети (начальное). В нашем случае пользователи (их уникальные идентификаторы) в сети Вконтакте. Нам известно, что данные пользователи имеют схожий атрибут, природа которого, сам факт его наличия и детерминированность в данной работе не обсуждаются.
2. Данный атрибут может быть интересом, увлечением, желанием купить что-то или же территориальной принадлежностью. Мы определяем это как некоторое общее значение атрибута для данной группы пользователей.
3. Мы предполагаем, что информация об атрибуте (о его значении/наличии) для всех пользователей социальной сети заложена в социальных связях принадлежности к группе. Как для начального подмножества, так и для искомого
4. Формализуя задачу, мы имеем ненаблюдаемый социальный граф пользователей и их связей с множеством групп. Однако информация о связях является доступной и упирается лишь в производительность вычислительных машин и ограничения API
5. Задача ставится в нахождении подмножества пользователей (искомое) с тем же значением атрибута.

Целью данной работы является исследование и разработка метода поиска подмножества пользователей (далее – *искомое подмножество*) социальной сети, обладающих общим значением произвольного атрибута, природу которого, вообще говоря, не обсуждаем. Тестирование метода будет проводится путем применения метода к заданной выборкой пользователей (далее – *исходное множество*) с заведомо заданным атрибутом. Мерой точности метода будут выступать стандартные метрики Precision, Recall и F1-мера. Для достижения цели необходимо решить следующие задачи:

1. Исследовать предметную область, изучить существующие методы кластеризации пользователей и вычисления значения того или иного конкретного атрибута.
2. Разработать и реализовать по меньшей мере 2 метода поиска людей с заданным значением атрибутов, на основе связей типа «Друзья» и информации о принадлежности группам
3. Провести экспериментальное исследование и сравнение разработанных алгоритмов.

Основной задачей, поставленной в данной работе, является исследование и разработка метода поиска *искомого подмножества* на основе информации о членстве пользователей в группах. Соответственно решение данной задачи включает в себя следующие пункты:

1. Реализация метода поиска пользователей, обладающих общим значением произвольного атрибута на основе информации о членстве пользователей в группах и *исходном множестве.*
2. Подбор параметров фильтров для получения оптимального результата.
3. Сравнение результатов с «наивным» методом и методом анализа дружественных связей.

Как и говорилось ранее, для реализации задачи используется программа, написанная на Python 3. Написана библиотека для работы с VK.API, не имеющая подобных удобных аналогов. В качестве среды для разработки выбран PyCharm, имеющий консоль откладки. Также, стоит отметить, что используемые в Python структуры данных (list, dict, set) идеально подходят, как для анализа, так и для взаимодействия с VK.API.

Для хранения данных в рамках одного instance используются локальные и глобальные переменные в самой программе. Все полученные с помощью API данные помещаются в базу данных SQLite, расположенную на SSD-диске.

Использование SQLite и БД в целом обусловлены: быстрой установкой, необходимостью кешировать результаты запросов ВК, наличием встроенных функций сортировок и фильтров, табличной структурой данных, SQL-возможностями JOIN, COUNT, SORT.

При использовании VK.API применяется метод VK.excute, позволяющий ускорить процесс получения данных с серверов Вконтакте в 25 раз (до 75 запросов в секунду, до 1000 значений-результатов в рамках одного запроса). Кроме того, встроенный JavaScript-подобный язык VK.excute позволяет перенести часть вычислительных нагрузок на сервера ВКонтакте.

В данной работе исследовались и разрабатывались методы поиска подмножества пользователей социальной сети, обладающих общим значением произвольного атрибута с заданной выборкой пользователей путем обхода социального графа. Все поставленные задачи были выполнены, в частности:

1. Исследована предметная область, изучены существующие методы кластеризации пользователей и вычисления значения того или иного конкретного атрибута.

2. Разработаны и реализованы 2 метода поиска людей с заданным значением атрибутов, на основе связей типа «Друзья» и информации о принадлежности группам.

3. Проведено экспериментальное исследование и сравнение разработанных алгоритмов.

Перспектива исследования и улучшение качества поиска возможна по следующим направлениям:

- Улучшение алгоритмов фильтрации групп и пользователей, более тщательная система ранжирования.
- Учет информации и о дружественных, и о связях типа группа для повышения точности результата.

Литература

1. *Boyd D. M., Ellison N. B.* Social network sites: Definition, history, and scholarship // Journal of Computer-Mediated Communication, 2007. 13 (1), article 11.
2. *Pallis G., Zeinalipour-Yazti D., Dikaiakos Marios D.* Online Social Networks: Status and Trends // New Directions in Web Data Management 1, Studies in Computational Intelligence Volume 331, 2011. Pp. 213-234.
3. *Najork M., Wiener J. L.* Breadth-first crawling yields high-quality pages // Proceedings of the 10th international conference on World Wide Web. ACM, 2001. С. 114-118.
4. *Leskovec J., Faloutsos C.* Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006. С. 631-636.
5. *Buzun N., Korshunov A.* Innovative Methods and Measures in Overlapping Community Detection // Proceedings of the International Workshop on Experimental Economics and Machine Learning (EEML 2012), Brussel, Belgium.
6. *Бузун Н., Коршунов А.* Выявление пересекающихся сообществ в социальных сетях // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов» АИСТ'2012. Екатеринбург, 16-18 марта 2012 г.
7. [Электронный ресурс]: Facebook Open Graph. Режим доступа: https://developers.facebook.com/docs/opengraph/.
8. Social Network Data Analytics. Editors: Charu C. Aggarwal // Springer, 2011
9. *Бартунов С., Коршунов А.* Идентификация пользователей социальных сетей в Интернет на основе социальных связей // Доклады Всероссийской научной конференции «Анализ изображений, сетей и текстов» (АИСТ’2012). Екатеринбург, 16-18 марта 2012 г
10. *Коршунов А.* Задачи и методы определения атрибутов пользователей социальных сетей // Труды. – 2013.
11. *Коршунов А. и др.* Анализ социальных сетей: методы и приложения // Труды Института системного программирования РАН, 2014. Т. 26. №. 1.
12. *Шомполов И. Г.* Новые образовательные технологии научно-педагогической системы выявления, отбора и методического сопровождения одаренных школьников в рамках межвузовской системы образования в московском физико-техническом институте в 2014/2015 учебном году. // МФТИ, 2015.
13. *Коршунов А.* Определение демографических атрибутов пользователей микроблогов // Труды Института системного программирования РАН. Том 25, 2013 г. С. 179-194.