Использование мультимодальных данных из различных источников для обнаружения сообществ в социальных сетях

Механиков Д. Ю., студент каф. компьютерных технологий
Университет ИТМО, dmekhanikov@gmail.com

Фарсеев А. И. ассистент каф. вычислительной техники,
Национальный Университет Сингапура (НУС), farseev@u.nus.edu

Фильченков А. А. к. ф.-м. н., доцент каф. компьютерных технологий
Университет ИТМО, aaafil@mail.ru

Аннотация

В данной работе проводится анализ возможности использования мультимодальных данных о пользователях, полученных из нескольких социальных сетей, в том числе информации о социальном графе, для автоматического обнаружения сообществ пользователей. В статье представлены результаты, полученные при анализе данных о пользователях из нескольких популярных социальных сетей, проживающих в трёх крупных городах.

Введение

Интернет уже длительное время является пространством, где можно не только найти необходимую информацию для работы и развлечений, но и пообщаться с единомышленниками а также опубликовать то, что будет доступно другим пользователям сети. Люди высказывают своё мнение и размещают личную информацию в социальных сетях, делая их огромным источником данных, доступных для исследования. Больше половины активных пользователей размещают информацию о себе сразу в нескольких социальных сетях[1].

Так же, как и в реальной жизни, пользователи интернета стремятся находить собеседников со схожими мыслями. Анализируя профили пользователей в социальных сетях, можно автоматически находить группы индивидуумов, размещающих информацию на схожие темы, а значит, с большой вероятностью также имеющих общие интересы в реальной жизни. Также пользователи могут сами вступать в группы, если такая возможность поддерживается используемой социальной сетью. В итоге,

[1] www.pewinternet.org/fact-sheets/social-networking-fact-sheet

информацию о явных и неявных сообществах можно использовать для отображения эффективной рекламы, анализа аудитории, выявления скрытых атрибутов пользователей, рекомендации друзей и в других прикладных задачах.

Взаимодействия между пользователями социальной сети могут быть представлены в виде графа, в котором в качестве вершин выступают пользователи, а в качестве рёбер – социальные связи между ними. Выявление сообществ в ориентированных графах является гораздо более изученной областью [1,2], чем в неориентированных. В данной работе графы пользователей также считаются неориентированными. Благодаря алгоритмам кластеризации решаются такие прикладные задачи, как объединение веб-страниц по темам [3], рекомендация гео-точек для посещения [4], и даже выделение функциональных групп протеинов в раковых клетках [5]. Однако на данный момент существует мало исследований использования данных различной модальности из множества источников.

Основываясь на опыте предыдущих исследований [6,7], с помощью использования данных различной модальности и из различных социальных сетей одновременно, возможно получить более полную картину о пользователях, чем это возможно при рассмотрении только одного источника. Данная работа использует результаты, приведённые в работе [7], в которой рассматриваются вопросы использования текста, изображений и информации о геопозиции, которую пользователи размещают в своих профилях в социальных сетях. Однако в этой работе не приводится способов обнаружения сообществ при использовании мультимодальных данных, а также не принимается во внимание информация о социальных связях пользователей, таких как дружба, подписка и т.д.

Целью данной работы является нахождение эффективного способа обнаружения сообществ пользователей при использовании мульти модальных данных из различных социальных сетей, в том числе информации о социальных связях пользователей.

Источники данных

Как говорилось ранее, для анализа доступно два вида данных: та информация, которую пользователи сами публикуют, и информация о том, кто находится в списках их друзей и подписчиков. Эти два вида данных несут в себе информацию разного характера: профиль пользователя описывает человека, его интересы и манеру общения, но не его социальные взаимодействия. Так же и личные данные человека не всегда

могут быть выявлены по списку его друзей и подписок. Данные о наполнении профилей пользователей были собраны и проанализированы в рамках работы [7]. Для сбора информации о социальных связях пользователей Twitter, Instagram и Foursquare в рамках данной работы была написана специальная программа. Полученные данные были использованы для построения неориентированного графа, вершинами которого являются пользователи, а рёбрами — их социальные связи. Таким образом, были получены два набора данных, которые в совокупности можно использовать более эффективно, чем по отдельности.

Связь между характеристиками пользователей и их социальными связями

Для проверки того, могут ли несколько видов данных быть использованы совместно, был произведён анализ сообществ, получаемых при разбиении пользователей по похожести размещаемой ими информации, и по их социальных графам. Для выделения сообществ в графе был выбран метод Louvain [8], который жадным образом максимизирует метрику modularity [9], объединяя сообщества в более крупные. Для кластеризации пользователей по их характеристикам также были построены графы, в которых рёбрами соединялись пользователи, косинусное расстояние между которыми было меньше некоторого порога, который выбирался таким образом, чтобы количество рёбер в обоих графах было соразмерным.

На обоих графах запускался алгоритм кластеризации Louvain, затем полученные результаты сравнивались между собой. Мерой схожести получаемых разбиений служил коэффициент корреляции Пирсона между векторами пользователей сообществ, которые были получены при кластеризации вершин в двух разных графах. Для этого строилась матрица, в которой столбцам соответствуют сообщества, а строкам – пользователи. В пересечении ставилась единица, если пользователь принадлежит соответствующему сообществу.

На Рисунке 1 изображена диаграмма корреляций между сообществами, где столбцы соответствуют сообществам, построенным по характеристикам, а строки – по социальным связям. На пересечении строки под номером i и столбца под номером j находится индикатор корреляции между сообществами по характеристикам и по социальным связям с номерами i и j соответственно. В таблице опущены статистически незначимые значения. Размеры строк и столбцов пропорциональны количеству пользователей в соответствующих сообществах.

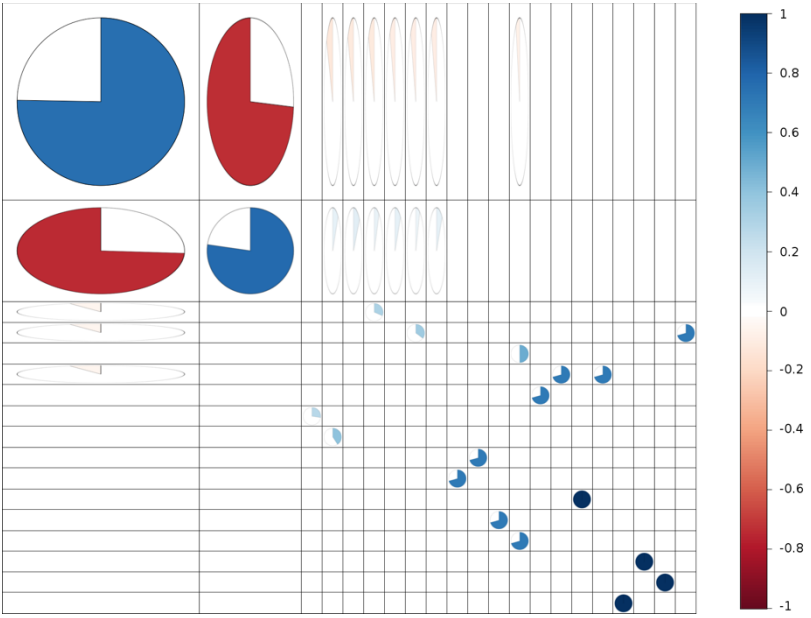

Рисунок 1: Диаграмма корреляций между сообществами, полученными при кластеризации вершин двух разных графов

Как видно из Рисунка 1, наблюдается связь между разбиениями, полученными двумя данными способами. Из этого можно сделать вывод, что пользователи социальных сетей взаимодействуют с людьми, которые похожи на них самих по размещаемому контенту.

Заключение

В работе были представлены аргументы в пользу использования мультимодальных кроссплатформенных данных для обнаружения сообществ пользователей в социальных сетях. Также была показана связь между личными данными пользователя и его социальным графом, что может служить основанием для разработки метода кластеризации, использующего совокупную информацию.

Литература

1. Girvan M.,Newman M. E. J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99 (12): 7821–7826. 2002.
2. Fortunato S. Community detection in graphs. Phys. Rep. 486 (3-5): 75–174. 2010.

3. Dourisboure, Y., F. Geraci, and M. Pellegrini, 2007, in WWW '07: Proceedings of the 16th international conference on the World Wide Web (ACM, New York, NY, USA), с. 461 - 470.
4. А. Фарсеев, Н. Жуков, И. Государев, и Ю. Заричняк. Разработка Кросплатформенной Рекомендательной Системы на Основе Извлечения Данных из Социальных Сетей Компьютерные Инструменты в Образовании. Июнь 2014.
5. P. F. Jonsson, T. Cavanna, D. Zicha and P. A. Bates, Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics, 7. 2006.
6. A. Farseev, D. Kotkov, A. Semenov, J. Veijalainen, and T.-S. Chua. Cross-Social Network Collaborative Recommendation. ACM International Conference on Web Science (WebSci) 2015.
7. Farseev A., Liqiang N., Akbari M., Chua T.-S. Harvesting multiple sources for user profile learning: a Big data study // ACM International Conference on Multimedia Retrieval (ICMR). China. June 23-26, 2015.
8. Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008 (10): P10008. 2008.
9. Newman M. E. J. Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA 103 (23): 8577–8696. 2006.