Знакомы ли «друзья» в сети Vkontakte в личной жизни?♣

А.В. Якушев, А.В. Бухановский

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
andrew.yakushev@yandex.ru, avb_mail@mail.ru

Аннотация

Все чаще для исследований процессов в реальном мире используются данные о структуре и свойствах социальных сетей. Это обосновывается утверждением, что социальные сети повторяют структуру социума, однако подробно этот вопрос никем не изучался. В данной работе изучается вопрос о доле связей в социальной сети Vkontakte, которые могут быть объяснены знакомством пользователей в реальном мире.

1. Введение

В последние годы анализ социальных сетей привлек к себе большое внимание, как представителей науки, так и представителей бизнеса. Для представителей бизнеса анализ огромного объема извлекаемой из социальных сетей информации позволяет качественней оценивать направление развития и создавать более точные рекомендательные системы [1]. Для академических исследований социальные сети интересны в первую очередь тем, что раскрывают ранее ненаблюдаемую информацию о структуре сообщества и различных аспектах взаимодействия индивидуумов в обществе [2] [3]. Считается, что некоторые социальные сети достаточно точно повторяют структуру реального сообщества и поэтому исследование их свойств может пролить свет и на процессы, происходящие в реальном сообществе.

Однако вопрос о том, как соотносятся друг с другом структуры социальных сетей и реального мира, никто не исследовал. Полный ответ на обозначенный вопрос должен, в свою очередь, содержать ответы на ряд других вопросов. Насколько социальная сеть репрезентативна с точки зрения состава сообщества? Насколько полно структура сообщества представлена в социальной сети? Как много в социальных сетях связей, обусловленных связями в реальном мире и обусловленных исключительно виртуальным знакомством пользователей? На некоторых из этих вопросов достаточно сложно получить ответ, поскольку их исследование неизбежно требует знания состава или структуры сообщества, а эти данные достаточно сложно получить.

В данной работе рассматривается вопрос о связях между пользователями социальной сети, обусловленных связями в реальном сообществе. Как много связей в социальной сети, которые так же присутствуют и в реальном мире? Как много с социальной сети «виртуальных» связей, отсутствующих в сообществе? Чем обусловлены эти связи в социальной сети?

Чтобы формализовать решаемую задачу необходимо в первую очередь определить, что такое структура сообщества. Как и в социальных сетях, в которых присутствуют различные классы связей между пользователями, так и в социуме можно поразному определять связь между людьми. Два человека могут быть родственниками, друзьями, коллегами или просто знакомыми. В данной работе мы рассматриваем отношение между двумя люди - «знакомые», означающее личное знакомство в жизни. Данный тип связи является наиболее общим и включает в себя другие типы связей между людьми.

Чтобы достоверно и точно ответить на поставленные вопросы, необходимы данные о структуре сообществе - для каждого человека необходимо знать список всех его знакомых. При наличии этих данных, ответ на поставленные вопросы получается тривиальными методами. Однако на практике получить данные о структуре сообщества затруднительно. И большинство методов для получения такой структуры позволят собрать малый объем данных, которые будут иметь невысокую достоверность.

Для ответа на поставленные вопросы, возможен и другой подход, основанный на построении классификатора связей в социальной сети. Классификатор на основе данных о двух пользователях, между которыми есть связь в социальной сети, будет предсказывать, имеется ли между ними связь в реальном мире. В основе построения классификатора будет лежать ряд критериев или признаков того, что при наличии связи между пользователями в сети, между ними так же есть связь и в реальном мире. Критерии в свою очередь, базируются на основе простых и рациональных предположений о связях между людьми в сообществе. Отметим, что предположения так или иначе будут иметь вероятностный характер,

Интернет и современное общество: сборник научных статей XVI Всероссийской объединенной конференции IMS-2013, Санкт-Петербург, 9 - 11 октября 2013 г.

и поэтому точный ответ на обозначенный вопрос получить невозможно. Однако мы считаем, что используемый поход позволит получить разумное первое приближение.

Используемые в классификации связей признаки позволяют объяснить связь в социальной сети знакомством пользователей в реальном мире, однако если критерий для связи не выполнен, то невозможно сказать, знакомы ли пользователи в реальном мире. Поэтому фактически решается задача о классификации связей в социальной сети на два класса: пользователи знакомы в реальном мире и неизвестно, знакомы ли пользователи. Таким образом, проверка насколько структура социальной сети соответствует структуре реального мира, сводится к определению доли связей, которые с высокой долей вероятности означают физическое знакомство пользователей.

По своей функциональности социальные сети сильно отличаются друг от друга, а их большое количество покрывает весь спектр возможных интернет-услуг. И именно предоставляемые сетью услуги оказывают непосредственное влияние на принцип формирования связей в сети. В узкоспециализированных сетях, посвященных например музыке, пользователи будут образовывать связи исходя из своих музыкальных предпочтений, тогда как в сетях упрощающих общение с друзьями, граф дружбы будут более похож на структуру социального сообщества.

Другим важным критерием при выборе социальной сети для анализа, является количество доступных о пользователе данных, на основе которых могут быть составлены признаки знакомства пользователей в реальном мире. Большая часть социальных сетей не предоставляет пользователям возможность указать подробные данные о себе.

При анализе данных так же имеется проблема недостоверности введенных данных, возникающая из-за пользователей умышленно указывающих неверную о себе информацию. Ряд социальных сетей ведет политику, направленную на то, чтобы пользователям было выгодно указывать свои реальные данные, поэтому число пользователей с недостоверными данными невелико и, из-за большого объема анализируемых данных, они не должны вносить большую погрешность в результаты.

Под описанные требования, в русскоязычном сегменте интернета, подходят две социальные сети Одноклассники и Vkontakte. Нами была выбрана вторая сеть, как наиболее посещаемая и большая сеть в России и как сеть, предоставляющая более гибкие и удобные средства для получения данных.

2. Данные пользователей Vkontakte

Для ответа на поставленные в работе вопросы мы использовали данные собранные из социальной сети Vkontakte. Для сбора данных был написан краулер [7], который путем обращения к API Vkontakte (http://vk.com/dev/) получил расширенную информацию о пользователях (метод users.get), связях между ними (метод friends.get), а так же о группах (метод users.getSubscriptions), в которых они состоят. Данные были получены в удобном для парсирования формате JSON и были загружены в не реляционную базу данных, предназначенную для работы с большим объемом информации, Apache HIVE, где уже были преобразованы в удобный для дальнейшего анализа формат.

Полученные данные являются открытыми, не требующими авторизации в системе, и доступны всем желающим, в том числе и поисковым системам. В настройках своего профиля, пользователь может указать, какие именно данные видны снаружи системы, и только эти данные Vkontakte выдает через свои методы API. Таким образом, никакая частная информация о пользователях не собиралась.

Среди собранной расширенной информации о пользователях, особенный интерес представляют сведения о местах его обучения и местах его проживания. На основании данных о продолжительно пребывании пользователя в том или ином месте мы будем строить признаки связей, говорящие о физическом знакомстве пользователей. На момент сбора данных (июль 2013), информацию о карьере пользователя, местах его службы или местах отдыха, методы API не предоставляют.

При анализе любых данных возникает задача обработки пропущенных данных. В социальных сетях она особенно актуальна, поскольку содержат большое количество пропущенных данных, возникающих, когда пользователь не указал часть данных или пожелал их скрыть в настройках приватности. Существует несколько подходов для работы с пропущенными данными, как вносящие искажение в данные (интерполяция, присвоение средней величине), так и не вносящие искажений. В данной работе, чтобы избежать лишнего искажения, анализируются только те связи, для которых известны все признаки. Несмотря на то, что доля связей, для которых неизвестно значение хотя бы одного признака велика, объем анализируемой выборки по-прежнему остается большим. Так же мы считаем, что отбрасывание связей с пропущенными признаками не вносит никакое смещение в рассматриваемую выборку. Т.е. в социальных сетях, вероятность того что пользователь не укажет часть своих данных зависит не от его данных, а от каких-то его внутренних «установках», на которые не оказывает влияние ни его образование, ни место жительства. Таким образом, анализируемая выборка по-прежнему является репрезентативной и обладает свойствами всей сети.

3. Определение знакомства пользователей

Для проверки гипотезы о похожести структуры социальной сети и сообщества строится бинарный классификатор, принимающий значения $\{0, 1\}$, который классифицирует связи между пользователями

в социальной сети на классы «пользователи лично знакомы» (значение 1) и «неизвестно, знакомы ли пользователи» (значение 0). Для построения классификатора используется признаковое описание связи между пользователями. В качестве признаков используется набор «сильных» бинарных признаков, которые с высокой вероятностью говорят о том, что пользователи знакомы в реальном мире.

Признаки знакомства пользователей имеют некоторый интуитивный характер и говорят о личном знакомстве, только при наличии между ними связи в социальной сети. При отсутствии связи между пользователями, наличие признака вовсе не означает личное знакомство пользователей. Например, признак «учились в одном университете», но отсутствие связи в социальной сети скорее говорит о том, что пользователи лично не знакомы. Признаки упорядочиваются в порядке своей «силы», соответствующей насколько сильно он говорит о физическом знакомстве пользователей. Данный порядок признаком является некоторой интуитивной экспертной оценкой и вытекает из простых практических соображений. Например, если пользователи дружат в сети, учились в одной школе и в одном университете, то, скорее всего, они знакомы еще со школы.

В качестве алгоритма классификации связей используется функция «ИЛИ», которая принимает значение 1, если хотя бы один из признаков, описывающих связь, равен 1. Так же представляется интересным знать, по какому именно признаку была образована связь. Будем считать, что связь образована самым «сильным» признаком, имеющим значение 1.

Могут быть выделены две группы признаков знакомства пользователей в реальном мире: на основании атрибутов пользователей и на основе топологической структуры сети.

3.1. Признаки на основе аттрибутов пользователей

Признак Обучения в одной школе. Для каждого пользователя составляется список школ, в которых он обучался. Затем для каждой пары пользователей, между которыми есть связь в графе дружбы, проверяется, учились ли они в одной школе. Для чего достаточно проверить, не пустое ли пересечение списка их школ. Чтобы уникально идентифицировать школу необходимо использовать три величины: страну и город, в которых она находится, а так же свой номер, уникальный в черте города. Данный признак является бинарным.

Признак Обучения в одном университете. Данный признак является бинарным и аналогичен предыдущему, за исключением того, что в данной ситуации рассматриваются высшие учебные заведения, которые указал пользователь.

Признак Проживания в одном городе. Данный признак не является сильным индикатором того, что пользователи знакомы лично, но он используется для оценки числа связей между пользователями, проживающими в разных городах.

Признак Общих интересов. Данный признак базируется на предположении что люди в реальном мире, знакомятся исходя из наличия единого увлечения или хобби. Например, люди, увлеченные различными видами спорта, особенно экстремальными, знакомы со многими другими людьми, обладающими такими же интересами. Однако это справедливо не для всех видов интересов. Для пользователей в социальной сети, можно посчитать похожесть их интересов, и для пользователей с похожими интересами можно сделать предположение о том, что они знакомы в реальном мире. Для описания интересов пользователя, можно использовать список групп, в которых он состоит. А для определения похожести между ними, хорошо зарекомендовавшие себя в анализе текстов метрики. Например, метрика косинусной похожести между двумя векторами [6]. Однако правдивость этого признака должна быть проверена.

3.2. Признаки основанные на топологии сети

Идея использования топологических признаков базируется на предположении, что если у пользователей есть много общих знакомых, то, скорее всего, они лично знакомы. Топологические признаки хорошо зарекомендовали себя во многих задачах [4] [5]. Существуют различные математические функции для оценки числа общих знакомых, в данной работе были использованы признаки:

Коэффициент Жаккара. Для двух пользователей x, y коэффициент Жаккара вычисляется как $J(x,y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$, где $N(x)$ – множество вершин, инцидентных вершине x. Данный количественный признак принимает значения из диапазона $[0,1]$ и показывает какая доля из вершин инцидентных вершинам x, y, инцидентна одновременно им обоим.

Коэффициент Адамика\Адара. Данный коэффициент похож на предыдущий, однако каждому узлу приписывается вес, зависящий от числа его связей (от степени узла). Интуитивно, если у пользователя немного связей, то связь с ним является более сильным сигналом о физическом знакомстве. Для двух пользователей x, y коэффициент Адамика\Адара вычисляется как $A(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{log|N(z)|}$.

3.3. Построение бинарных признаков по количественным

Функция «ИЛИ», используемая для классификации связей в графе, принимает в качестве аргументов только бинарные признаки. Поэтому все количественные признаки, чье множество значений отлично от множества $\{0,1\}$, должны быть преобразованы в бинарные признаки.

Таблица 1. Гистограмма значений бинарных признаков, означающих что пользователи «знакомы» лично (в т.ч. статистика по числу связей между пользователями, проживающими в одном городе)

Значение признака	Проживали в одном городе	Учились в одной Школе	Учились в одном университете	«Много» общих друзей
Отсутствует	19237	78990	59614	36345
Присутствует	72918	13165	32541	55810
Доля связей, %	72%	14%	35%	61%

Поскольку количественные признаки обладают свойством монотонности – чем больше общих связей между пользователями, тем больше коэффициент Жаккара, то для преобразования их в бинарный признак можно использовать подход, основанный на сравнении с пороговым значением.

Бинарный признак $b_i(x,y)$, соответствующий количественному признаку $f_i(x,y)$, определяется как $b_i(x,y) = f_i(x,y) \geq F_i$, где F_i пороговое значение признака $f_i(x,y)$. Оптимальное значение F_i определяется на основе обучающей выборки, состоящей из признакового описания связей знакомых пользователей. На основе обучающей выборки, строится распределение значений величины $f_i(x,y)$, которое аппроксимируется некоторым типовым распределением. В качестве значения F_i используется 50% перцентиль восстановленного распределения.

Стоит заметить, что если рассмотреть вторую группу пользователей, для которых не установлено, знакомы они в жизни или нет, то распределение хорошего числового признака должно в нем отличаться от распределения в обучающей выборке, в которой пользователи знакомы. Для группы знакомых пользователей распределение должно иметь большой хвост.

3.4. Ограничения описанного метода

Описанный подход позволяет определить долю связей в социальной сети, которая может быть обусловлена физическим знакомством пользователей. Однако он не позволяет определить, насколько полно представлена структура сообщества в социальной сети. В процессе анализа будет определена доля связей в сети, для которых не выполняется ни один из используемых признаков личного знакомства пользователей. Среди этих связей есть как связи вида «физически знакомы», так и «виртуально знакомы», однако определить долю тех или иных описанный подход не позволяет.

Несмотря на интуитивный подход, имеющий вероятностный характер, который используется для определения личного знакомства пользователей, он строится на рациональных и простых предположениях. Поэтому, как нам кажется, он может быть использован для поиска «первого приближения» для ответа на вопрос о доле связей в сети, обусловленных реальным знакомством пользователей.

4. Результаты

Для анализа были использованы данные о не удаленных пользователях из сети Vkontakte со значением уникальных пользовательских идентификаторов меньше 100000. При анализе графа дружбы пользователей рассматриваются только ребра соединяющие вершины из этого списка. Самая большая компонента связности полученного графа дружбы пользователей содержит 80205 вершин и 1381671 ребер. На основании данных из профилей пользователей, для каждого ребра были посчитаны признаки «Учились в оной школе/университете», Посчитаны коэффициенты Жакарра, Адамика/Адара, а так определены тематические близости между пользователями, на основе числа общих групп, в которых они состоят. После удаления пользователей, с пропущенными значениями хотя бы одного признака, в анализируемом графе осталось 92155 ребер, которые были классифицированы на два класса. Статистика по значениям признаков на исходном графе, до удаления пропущенных значений представлена в таблице 1.

Для определения порогового значения для коэффициента Жаккара, при превышении которого пользователи считаются знакомыми лично, использовалась обучающая выборка, состоящая из анализируемых данных. Гистограмма значений коэффициента для двух классов ребер, приведена на рисунке 1. Распределение для класса знакомых пользователей хорошо аппроксимируется LogN(-2.50, 0.81) распределением, с 50% перцентилем равным 0.08.

Напомним, что мы считаем пользователей лично знакомыми, если посчитанный для них коэффициент Жаккара больше 0.08. Точное пороговое число общих друзей зависит от степени вершин, поэтому оно различное для разных вершин. Но для двух вершин, имеющих среднее для сети число друзей, равное 35, пользователи будут классифицированы как лично знакомые, если у них больше 5 общих связей.

При анализе значений коэффициента Адамика\Адара было обнаружено, что он сильно коррелирует с коэффициентом Жаккара, при этом распределение значений плохо аппроксимируется каким-либо из типовых распределений, поэтому он не был использован для классификации связей.

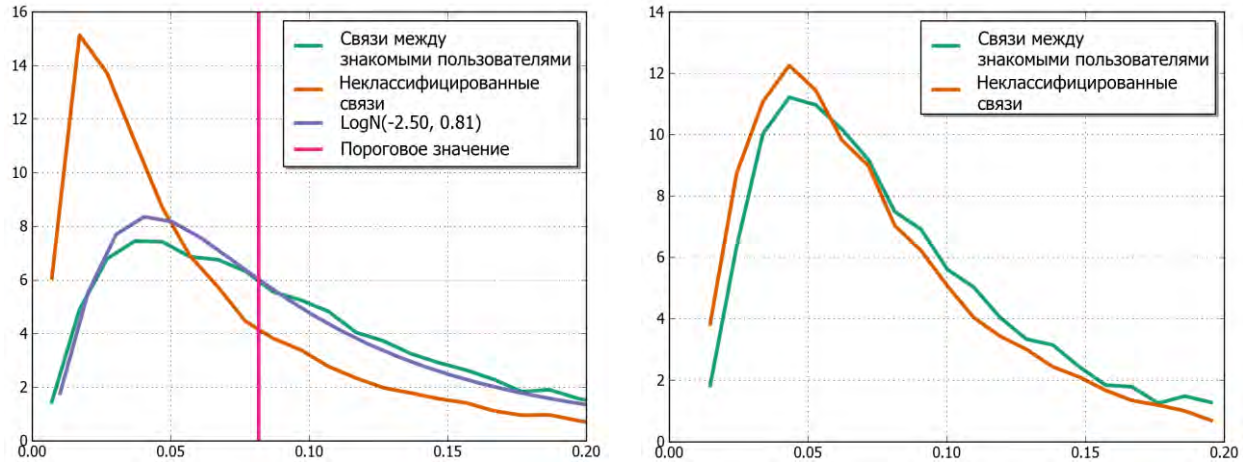

Рис. 1. Гистограмма значений числовых признаков, для класса пользователей знакомых друг с другом и класса неизвестных связей. а) коэффициент Жаккара, который аппроксимирован распределением LogN(-2.50, 0.81). 50% перцентиль равен 0.08. б) Распределение косинусного коэффициента похожести между пользователями, посчитанного на основании групп, в которых состоят пользователи

Для проверки выполнения признака того, что пользователи могут быть лично знакомы, при наличии схожих интересов, была построены гистограммы значений косинусной похожести для двух групп пользователей: знакомых между собой пользователей и группы с неизвестными связями (рисунок 1.б). Было обнаружено, что в обозначенных двух группах распределения величин совпадают, поэтому данный признак не может быть использован для классификации связей.

Таким образом, в качестве критериев для определения, знакомы ли пользователи сети лично, были использованы три признака: учились в одной школе, учились в одном университете, а так же признак «много общих друзей», определяемый через коэффициент Жаккара. Встает вопрос, сколько связей в сети может объяснить тот или иной признак. Однако для некоторых связей могут быть выполнены сразу несколько признаков.

Таблица 2. Статистика классификации связей между пользователями в сети

Признак знакомства пользователей	**Число связей**	**Доля связей**
Учились в школе	13165	14%
Учились в университет	29407	32%
Много общих друзей	35508	39%
Общая статистика связей		
Связей между знакомыми пользователями	78080	85%
Неклассифицированных связей	14075	15%

Поэтому чтобы снять неопределенность в выборе признака для определения связи, признаки были упорядочены в порядке своей достоверности. Результаты приведены в таблице 2. Тем самым, можно говорить о том, что 85% связей в сети, могут быть объяснены знакомством пользователей в реальном мире. 15% связей не могут быть объяснены используемыми признаками.

Заключение

В работе представлен анализ связей между пользователями социальной сети Vkontakte. Основываясь на простых и интуитивных предположениях, выделены критерии, по которым определятся знакомство пользователей в реальном мире. Было обнаружено, что 85% связей в сети соответствуют лично знакомым пользователям и природа образования 15% связей не может быть установлена на базе используемых признаков.

В работе представлена оценка числа связей в социальной сети, обусловленных реальным знакомством пользователей. Однако вопрос о репрезентативности социальной сети и ее смещения по сравнению со связями присутствующими в обществе не рассматривается. Поиск ответа на него представляет большой научный интерес.

Литература

[1] Swamynathan G. et al. Do social networks improve e-commerce?: a study on social marketplaces //Proceedings of the first workshop on Online social networks. – ACM, 2008. – С. 1-6.

[2] McPherson M., Smith-Lovin L., Cook J. M. Birds of a feather: Homophily in social networks //Annual review of sociology. – 2001. – С. 415-444.

[3] Liben-Nowell D., Kleinberg J. The link-prediction problem for social networks //Journal of the American society for information science and technology. – 2007. – Т. 58. – №. 7. – С. 1019-1031.

[4] Al Hasan M. et al. Link prediction using supervised learning //SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. – 2006.
[5] Shalizi C. R., Thomas A. C. Homophily and contagion are generically confounded in observational social network studies //Sociological Methods & Research. – 2011. – Т. 40. – №. 2. – С. 211-239.
[6] Singhal A. Modern information retrieval: A brief overview //IEEE Data Eng. Bull. – 2001. – Т. 24. – №. 4. – С. 35-43.
[7] Якушев А.В., Дейкстра Л.Й., Митягин С.А.. Распределенный краулер для социальных сетей на основе модели Map/Reduce // Информационно-измерительные и управляющие системы. – 2012. – Т. 10. – №. 11. – С. 47-53.

Are friends in social network Vkontakte know each other in real world?

A.V. Yakushev, A.V. Boukhanovsky

It becomes popular to use data from Social Networks Sites (SNS) to model processes in real world. The main assumption in these researches is that structure of complex network in SNS is similar to the structure of real society, but details of these issues have not been studied. This paper studies the question of proportion of links in popular Russian SNS Vkontakte that can be explained by the familiarity of users in real world.

Three probabilistic but strong features used to determine the proportion of links between familiar users in SNS. Together features "studied in same school", "studied in same university" and "have a lot of common friends in SNS" were able to explain 85% of links in SNS.

♣ Работа выполнена в рамках реализации постановления №220 Правительство РФ (договор №11.G34.31.0019) при поддержке Минобрнауки Российской Федерации, ФЦП «Научные и научно-педагогические кадры инновационной России на 2009 - 2013 годы», соглашение 14.B37.21.1886 от 04.10.2012 г. и 14.B37.21.0596 от 17.08.2012 г.