Данные онлайновых соцсетей и масштабные сети

Николай Скворцов
nskv@mail.ru

План

- Сбор данных из сетей для анализа
- Подходы к анализу масштабных сетей
  - Репрезентативная выборка
  - Распараллеливание
- Доверие данным онлайновых соцсетей

Подходы к сбору данных социальных сетей

- Традиционные подходы сбора данных
  - Опрос (анкетирование, интервьюирование)
    - Заполнение форм опросов без участия или с участием интервьюера
    - Сеть формируется из фрагментов эгоцентрических сетей респондентов
  - Наблюдение
    - Часто наиболее точен, основан на фиксации фактов, а не на воспоминаниях о прошлых взаимодействиях
  - Эксперимент
    - Наблюдение взаимодействий в контролируемых условиях
  - Анализ архивов публикаций
- Цифровые способы сбора данных могут повторять традиционные
  - Формы веб-сайтов и приложений – опросы
  - Веб и логи – наблюдение

Данные онлайновых социальных сетей, форумов и веб-сайтов

- Личная информация
  - Данные, предоставляемая о себе пользователями – атрибуты агентов сети
- Данные публичной деятельности
  - Сообщения, заметки, отметки, обмен содержимым (перепубликации) – данные о предпочтении
- Непубличные предпочтения и действия
  - Информация об использовании сетей, время сессий – данные поведения
  - Корпоративные базы данных о действиях клиентов, транзакциях, коммуникациях, перемещениях
- Статические и динамические отношения
  - Статические связи друзей – неориентированные графы социальных связей
  - Статические связи следования – ориентированные графы подписки
  - Перепубликации – графы видимых взаимодействий
  - Посещения страниц других пользователей – графы скрытых взаимодействий

Подходы к работе с большими сетями

- Скачивание и обработка всей мастабной сети невозможна
- Работа с интересующими фрагментами
- Потоковая обработка обновлений
- Предобработка и агрегация
- Работа с репрезентативными выборками
- Распараллеливание

Репрезентативные выборки

- Репрезентативная выборка (sampling) – процесс (и результат процесса - sample) выборки фрагмента сети, повторяющего характеристики всей сети
- Рассмотрим сеть
  - Однотипные вершины и однотипные рёбра (или дуги)
  - Данные о вершине доступны, только когда она входит в выборку
  - Атрибуты вершины и непосредственные соседи

Имитируемые характеристики в выборках

- Распределение степеней
  - либо входящих/исходящих степеней
- Распределение размеров компонент
  - С учётом или без учёта направления дуг
- Распределение коэффициента кластеризации
- Распределение расстояния между вершинами
  - Свойство тесных миров
- Распределение сингулярных чисел матрицы смежности

Методы репрезентативной выборки

- Выбор вершин
  - RN (random node) – случайная вершина
  - RDN (random degree node) – выбор из распределения степеней вершин
    - Должны быть известны степени всех вершин
  - RPN (random PageRank node) – выбор из распределения PageRank
    - Должен быть рассчитан PageRank всех вершин
- Выбор рёбер
  - RE (random edge) – случайное ребро (с парой вершин)
  - RNE (random node edge) – случайная вершина и случайное ребро этой вершины
  - IE (induced edge) – случайное ребро и включение рёбер с уже вошедшими вершинами
- Блуждания
  - RW (random walk) – случайное блуждание
  - RWR (random walk with restart) – случайное блуждание со случайным переходом в начальную вершину
  - RJ (random jumps) – случайное блуждание со случайным переходом в случайную вершину
  - FF (forest fire) – метод лесного пожара
    - Выбор подмножества соседей (из не входящих в выборку), каждого с вероятностью p

Методы репрезентативной выборки

- MHRW (Metropolis-Hastings random walks) – случайное блуждание по Метрополису-Гастингсу
  - Переход зависит от степеней вершин: $P_{vw} = \begin{cases} \frac{1}{k_v} \min\left(1, \frac{k_v}{k_w}\right), \in E \\ 1 - \sum P, w = v \end{cases}$

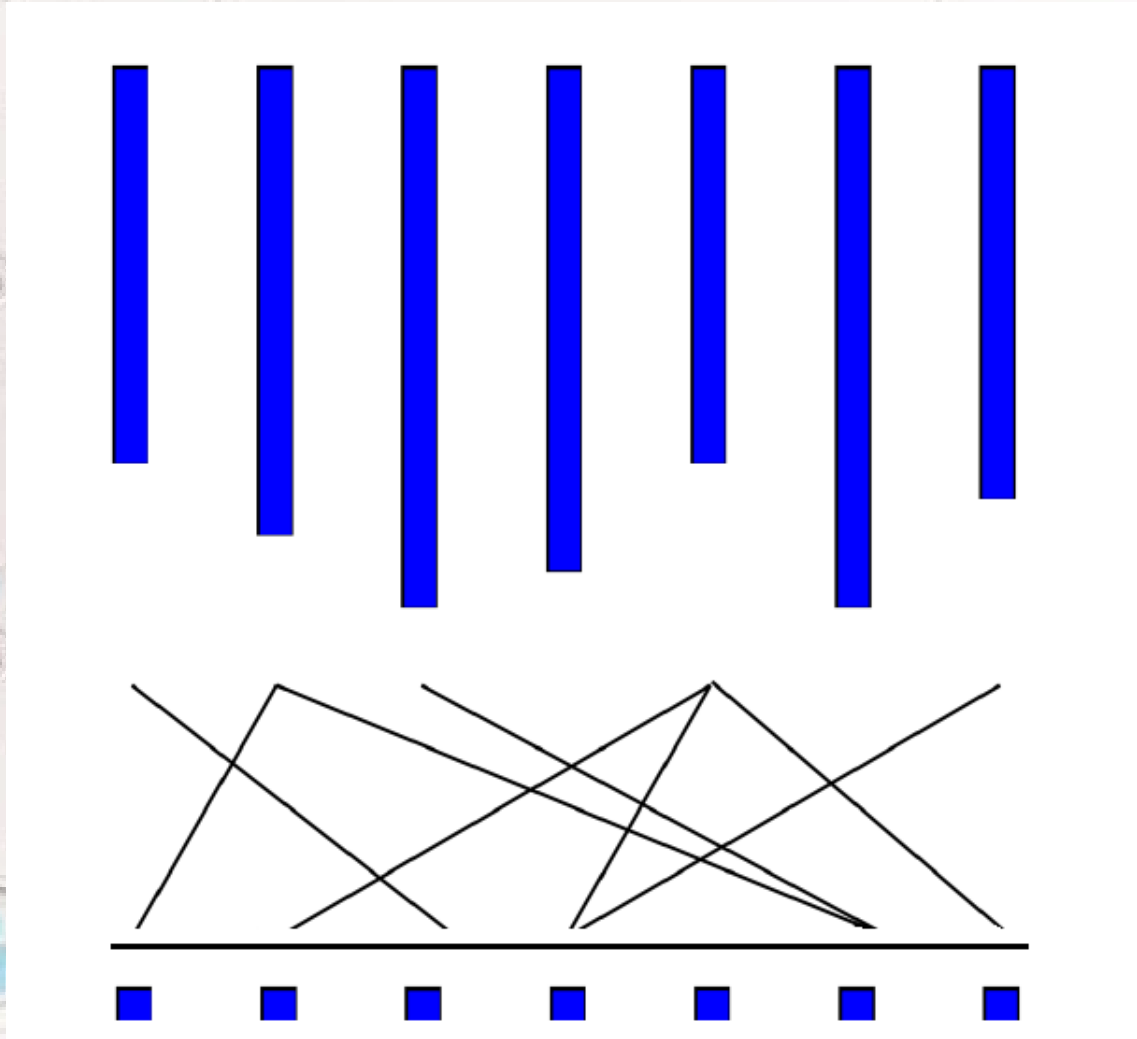
Распараллеливание обработки больших сетей

- Необходимость распараллеливания обработки в кластерах компьютеров
  - Интернет (WWW)
    - По оценке Google – более 1 триллиона страниц
  - Социальные медиа
    - Facebook: 2010 – $500*10^6$, 2013 – 10^9 ($650*10^6$ акт.польз./день),
    - LinkedIn: 2013 – $8*10^6$, $60*10^6$ связей
    - Twitter: 2011 – $140*10^6$ сообщений/день
    - россияне создают 30 млн постов в социальных медиа каждый день
  - Транспортные сети
  - Банковская деятельность
  - Биоинформатика

- Примеры моделей параллельной обработки данных социальных сетей
  - Bulk Synchronous Parallel (BSP)
  - Gather-Apply-Scatter (GAS)
  - Map/Reduce

Bulk Synchronous Parallel (BSP)

- Программная модель
  - Параллельное выполнение подзадач на узлах в начале шага
  - Посылка сообщений необходимым узлам по результатам подзадач
  - Глобальная синхронизация начала следующего шага
- Реализации
  - Google Pregel, Apache Giraph, Apache Hama и другие



Pregel

- Google, 2010 год
- Программная модель
  - Bulk Synchronous Parallel
  - Параллельная обработка вершин
  - Коммуникации с помощью обмена сообщениями между вершинами
  - C++
- Плюсы Pregel:
  - отсутствие блокировок (deadlock) и гонок по данным (data race)
  - отказоустойчивость (естественный способ снятия контрольных точек)
- Минусы Pregel:
  - Простаивание процессора/узла из-за синхронизации
  - Требует много памяти

Giraph

- Наиболее популярный open-source клон Google Pregel
- Разрабатывается с 2012 года
- Java + Hadoop/Yarn (map only) + (HDFS + Zookeper)
- Поддержка мультитрединга внутри узла (multithreaded mappers)
- Коммуникации – TCP/IP (Netty)
- Распределение вершин: Hash(Vertexid) mod N
- Используется в Facebook (с лета 2013), LinkedIn, Yahoo!
- Вычисления в памяти (in-memory)
- Удобные средства отладки и мониторинга (JobTracker, TaskTracker),
- возможность работать через http

Gather-Apply-Scatter (GAS)

- Программная модель
  - Init – для данной вершины при необходимости проводится обработка сообщений, полученных от смежных вершин
  - Gather – входящие, исходящие дуги или неориентированные рёбра собираются и суммируются. Семантика оператор сложения (+ =) переопределяется программой
  - Apply – собранные данные применяются для изменения данных вершины и/или программы вершины.
  - Scatter – рассылка сообщений смежным вершинам по входным, выходным дугам или неориентированным рёбрам, несколько сообщений одной вершине объединяются посредством переопределённого оператора сложения (+ =)
- Реализации
  - GraphLab

GraphLab

- Carnegie Mellon University, 2009
  - Модель Gather–Apply–Scatter (GAS)
  - Поддержка разных режимов выполнения программы
    - Asynchronous, Synchronous (BSP)
  - Поддержка транзакций
    - vertex consistency, edge consistency, full consistency
  - Распределение вершин по узлам системы
    - random, grid, oblivious (heuristic)
  - Коммуникации TCP/IP sockets
  - Вычисления в памяти (in-memory)
- Преимущества
  - Алгоритмы ориентированы на обработку вершин
  - Естественность для графовой модели
- Недостатки
  - Не все алгоритмы ориентированы на вершины
  - Большие накладные расходы, если обработка на вершине мала
  - Возможны конфликты

Map-Reduce (MR)

- Hadoop
  - Дешёвые узлы
  - Java
  - Возможность работы в готовом облаке
- Проблемы Hadoop
  - Двухфазная обработка (особенность реализации алгоритмов)
  - Невозможно запоминать состояние и возвращаться к началу обработки (сложность для реализации итеративности)
  - Сложность оптимизации
- Модификация Haloop

Средства обработки данных социальных сетей

- BSP
  - Google Pregel, Apache Giraph, Apache Hama
- GAS
  - GraphLab
- Модель Map-Reduce
  - Yarn, Haloop, Pegasus
- Библиотеки в языках программирования
  - R (sna)
  - C++, Python (SNAP)
- Графовые СУБД
  - Neo4j
- Среда анализа
  - Gephi
- Моделирование
  - NetLogo

Проблема доверия данным социальных сетей

- Традиционные социальные опросы
  - Затратные: вербальный контакт с каждым человеком, труд интервьюеров, транспорт, обработка данных
- Альтернатива – анализ данных, полученных из сети
  - Дешёвые данные
  - Большая выборка данных
  - Данные разной тематики
  - Выделение данных по времени, регионам, тематике, упоминанию объектов и событий
- Специалисты часто отказываются от данных, доступных в интернет, в пользу традиционного сбора и анализа данных опросов
  - Методы сбора и анализа отличаются от методов в соцопросах
  - Результаты анализа часто существенно отличаются от результатов соцопросов
  - Данные соцсетей содержат большой объём сторонних данных
- Неизбежно встаёт вопрос о качестве данных социальных сетей и доверии им
  - Какие факторы влияют на качество данных социальных сетей?
  - Как увеличить доверие данным социальных сетей?

Оценки качества данных соцсетей

- Точность измерений (accuracy)
  - Сравнение социальной сети, полученной в результате сбора данных, с реальной сетью
    - Сравнение результатов опросов с результатами наблюдения
    - Данные эгоцентрических сетей (на основе прямых отношений агентов) обычно точнее данных когнитивных сетей (отражающих мнение об отношениях других людей)
- Надёжность измерений (reliability)
  - Разброс результатов при повторных измерениях
    - Повтор во времени имеет смысл только для достаточно статичных отношений
    - Для динамических отношений можно сравнивать данные, полученные в одно время из разных источников или разными методами
- Достоверность измерений (validity)
  - корректность метода
    - Постановка вопросов при сборе данных
    - Применяемые методы анализа данных
    - Теоретические прогнозы в моделях
- Точность и надёжность – случайная погрешность измерений
- Достоверность –систематическая погрешность

Вопросы применимости данных онлайновых соцсетей

- Не могут охватить часть населения, не использующую социальные сети
  - ограничение, невосполнимое без использования традиционных подходов
  - невозможно обобщить результаты исследований на всех жителей территории
  - Не влияет на целенаправленные исследования пользователей социальных сетей
- Ограничение правилами онлайновой социальной сети
  - Принимаемые пользователями при регистрации
  - Установленный уровень конфиденциальности данных
- Необязательность ответа на вопросы
  - При опросах человек вынужден отвечать
  - Пользователь социальной сети не напишет о том, чем не хотел бы делиться со своей аудиторией в сети
- Ограничения влияют на точность и достоверность измерений

Сложности интерпретации данных

- Данные онлайновых соцсетей не имеют прямого отношения к исследованию
  - в отличие от опросов, они не отвечают на чётко сформулированные вопросы
- Сложности интерпретации текста
  - Цитирования других людей, суждения в иронической форме, в переносном значении
- Количество учитываемых мнений
  - В опросах анкета соответствует мнению одного респондента
  - В соцсети человек не один раз обозначает свою позицию по одной теме
  - Неодинаковая активность агентов
- Время описываемых событий выбирается пользователей сети произвольно
  - Сообщения могут описывать текущее состояние или воспоминания о взаимодействии
- Различия терминологии разных людей
  - В опросах респонденты ограничены рамками контролируемого словаря опроса
  - Пользователи онлайновых соцсетей ничем не связаны в выражении своего мнения
- Влияние функциональности онлайновых соцсетей
  - возможности перепубликации, публикация фотографий и видео существенно меняют поведение людей в сети
- Вывод: страдает достоверность измерений

Психологические факторы различий в социальных данных

- Выход из обычной обстановки
  - Респонденты соцопросов вырываются из обычной деятельности специально для опроса
  - Общение в соцсети в комфортной обстановке, по собственной инициативе, без ограничений по времени, выбор обсуждаемой тематики
- Целевая аудитория
  - Респондент опроса отвечает так, чтобы это понравилось интервьюеру или исследователю
  - Пользователь соцсети создаёт посредством сообщений определённый имидж перед другими пользователями: знакомыми с установленными связями или сторонними читателями
  - Пользователь соцсети рассчитывает на реакцию и ответ
  - Пользователь соцсети может выражать мнение сообщества
- Анонимность
  - Респондент знает об анонимности своих ответов
  - Раньше в форумах степень анонимности была высокой
  - Пользователь соцсети в большинстве случаев идентифицируем
  - Пользователь соцсети выражает мнение с учётом публичности
  - Учёт всеобъемлющего слежения ряда правительств за информационными коммуникациями по всему миру
  - Личные данные пользователи соцсетей часто подделывают
- Вывод: страдает точность измерений

Рекламные технологии и вредоносная деятельность

- Массовое манипулирование результатами опросов в соцсети
  - Целенаправленная деятельность заинтересованных сообществ
- Искусственно навязанные обсуждения определённых тематик
  - Алгоритмы регулирования появления сообщений в лентах новостей и оповещениях пользователей
  - Алгоритмы не раскрываются соцсетями и меняются со временем без предупреждений
- Фиктивные учётные записи
  - Для рекламы и спама
  - Для формирования общественного мнения
  - Для доступа к данным профилей пользователей
  - Для раскрутки определённых страниц
- Ведутся дискуссии на тему этики использования открытых, а тем более полученных несанкционированным доступом данных для исследований

Повышение доверия к данным на этапе их сбора

- Выборка представительных данных
  - Сбор данных должен начинаться не с одной вершины, а с некоторого множества доступных вершин
  - Выборочный сбор данных (sampling)
    - формирование подграфа соцсети с характеристиками, приближенным к характеристикам полной сети
- Аннотирование происхождением данных
  - Идентификация авторов, их социально-демографические характеристики, принадлежность данных
  - Актуальности во времени и географические метки
  - Изначальные источники данных
  - Пути миграции данных
  - Способы получения вычислимых параметров
  - и другие
- Происхождение данных используется при очистке данных на всех этапах их использования для оценки достоверности, подлинности, актуальности, точности данных

Совместное использование данных различного происхождения

- Рекомендуется совмещать данные, полученные разными способами, в одном исследовании
  - Сравнение данных из разных источников
  - Обогащение данных, полученных разными способами
- Сопровождение данные онлайновых сетей данными анкет о деятельности пользователей в сети
  - После взаимодействия в сети незамедлительно предложить анкету с дополнительными вопросами о взаимодействии
- Данные опросов целесообразно подкреплять их данными наблюдений в сети
  - Проверка объективности ответов в опросах реальной картиной взаимодействий
- Важен не только анализ данных, которые сознательно предоставляются аудитории пользователями, но и анализ их поведения в сети
  - Наблюдаемое поведение пользователей сети чаще представляет объективные факты о взаимодействиях
  - Описание собственных взаимодействий пользователями сети представляет субъективное отношение к ним

Наблюдаемые и скрытые взаимодействия

- Наблюдаемые взаимодействия открыты для внешних наблюдателей и для исследования
  - Перепубликации, отметки, сообщения на стенах друзей и другие
  - Корректировка граф социальных связей
    - Агенты взаимодействуют только с небольшой частью связанных с ними агентов
- скрытые взаимодействия не имеют видимого следа в соответствии с правилами сети или выбранным уровнем конфиденциальности
  - Количество скрытых взаимодействий значительно больше, чем наблюдаемых
- Регистрация трафика
  - Сбор данных о сессиях
    - Частота и продолжительность
  - Сбор данные о практически всех типах взаимодействий
  - Идентификация пользователей
  - Мониторинг корпоративных сетей
  - Пользователи сети, установившие специальное приложение для регистрации трафика

Обнаружение вредоносной деятельности

- Фильтрация возможных фиктивных данных и навязанной деятельности
  - Фиктивные агенты связаны в основном друг с другом
    - Обнаружение сообществ из вредоносных агентов и исключение их из анализа
  - Генерируемое содержимое
    - Методы обнаружения спама
  - Компрометируемые учётные записи
    - Изменение характеристик взаимодействий агента с определённого момента

Выводы

- Для повышения доверия к данным онлайновых социальных сетей рекомендуется
- совместно использовать данные опросов, данные соцсетей, данные наблюдения,
- оценивать их качество, предпринимать усилия по их очистке
- снабжать метаданными происхождения