МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра социальной информатики

МОДЕЛИ СЛУЧАЙНЫХ ГРАФОВ ИНТЕРНЕТ СООБЩЕСТВ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

(автореферат бакалаврской работы)

студентки 5 курса 531 группы
направления 09.03.03 – Прикладная информатика
профиль Прикладная информатика в социологии
социологического факультета
Ивановой Анны Сергеевны

Научный руководитель
кандидат физико-математических наук, доцент_____ Л.Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент_____ И.Г.Малинский
подпись, дата

Саратов, 2020

ВВЕДЕНИЕ

Актуальность проблемы исследования. В современном мире существует большое количество информации, которую можно представить в виде объектов и отношений между ними. Например, объектами могут являться научные статьи, тогда, если одна из них ссылается на вторую, то между этими статьями есть связь. Таким образом, все существующие научные работы могут быть представлены в виде графа. Такое же представление возможно для многих других структур из самых разных областей знания: люди и их социальные взаимоотношения, интернет-ресурсы, ссылки.

Графы - это простая, мощная абстракция, применяемая во многих областях науки и техники. Графы позволяют моделировать произвольные системы, представимые в виде набора объектов и связей между ними. За последние десятилетия популярность графов значительно возросла. Объясняется это их универсализмом и независимостью от предметной области, что позволяет обрабатывать и анализировать системы произвольной сложности. Информационный взрыв, вызванный развитием Всемирной Паутины, и развитие компьютерных технологий сделали доступным большое количество информации. Сегодня теория графов используется в биологии при анализе ветвящихся процессов, а именно, размножении бактерий, в радиоэлектронике - при проектировании печатных схем, в химии - для описания кинетики сложных реакций, в экономике - при оптимизации маршрутов и грузоперевозок, в социологии - для изучения связей и закономерностей в обществе и т.п. Графы являются центральным инструментом при разработке программного обеспечения, структурировании бизнес-процессов, проектировании баз данных и используются во многих областях математики и компьютерных наук.

Степень изученности темы исследования. Для всех подобных структур является естественным их представление в виде графа. Все описанные примеры являются так называемыми социальными графами: они обладают неким набором свойств, специфичных только для такого типа графов. С каждым днем размеры графов увеличиваются, их сложность растет, и даже

такая важная задача, как визуализация, становится проблематичной. Один из возможных подходов к ее решению заключается в более высокоуровневом представлении исходного графа: изображать вместо вершин графа их группы. Однако, важно, чтобы при группировке вершин не потерялась глобальная структура графа. Такие группы можно назвать сообществами в графе.

Методы выделения сообществ в социальных графах активно исследуется в последнее время, и визуализация графа — всего лишь один из практических аспектов этой задачи. В данной работе анализ графов будет идти именно с этой точки зрения.

Целью выпускной квалификационной работы является анализ случайных графов крупных сетях.

Для решения поставленной цели необходимо выполнить ряд **задач**:

1) Изучить существующие модели случайных графов
2) Исследовать применение графов при анализе крупных сетей
3) Проанализировать пример применения кластеризации на крупных сообществах.

Структура выпускной квалификационной работы. Данная работа состоит из введения, 3 разделов, заключения, списка использованных источников и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Основные модели случайных графов» описывает популярные модели случайных графов. Раскрывается понятие безмасштабных сетей.

В настоящее время для построения различных моделей социальных сетей применяют теорию случайных графов. Существует много видов моделей, генерирующих случайные графы, близкие по свойствам к реальным сетям. Их можно разделить по генерируемым графам на несколько основных классов:

1) Модели случайных графов (модель Эрдёша-Реньи);
2) Простейшие модели безмасштабных сетей (модель Боллобаша-Риордана , модель копирования и др.);

3) Более гибкие модели безмасштабных сетей (модель Чунг-Лу, модель Янсона- Лучака);
4) Модель стохастических графов Кронекера;

Теория случайных графов стала активно развиваться после публикации в конце 1950-х статей Эрдёша-Реньи об эволюции случайных графов[1]. Это модель неориентированного случайного графа без кратных ребер и петель. На данный момент эта модель является самой изученной среди случайных графов. Но, в начале 2000-х, выяснилось, что она плохо подходит для описания графов, возникающих в реальных социальных сетях.

Для описания сети интернет введем понятие веб-графа – это граф, вершинами которого являются какие-либо конкретные структурные единицы в интернете: страницы, сайты, хосты. Для определенности принято считать вершинами веб-графа сайты, а ребрами соединять те вершины, между которыми есть ссылки. При этом число ребер между вершинами равно числу ссылок между соответствующими сайтами, возможны ссылки сайта на себя (т.е. петли), ребра разумно полагать направленными. Таким образом, веб-граф ориентирован и он может иметь кратные ребра и петли.

В начале 2000-х Барабаши, Альберт и Х. Джеонг описали следующие свойства веб-графа:

1. Веб-граф формируется добавлением к нему новых вершин, соединяемых ребрами со старыми вершинами.

2. Диаметр веб-графа мал, в 1999 году он имел величину 5-7.

3. Веб-граф имеет степенное распределение вершин.

Графы со степенным распределением степеней вершин называются безмасштабными. Исходя из своих наблюдений, Барабаши и Альберт ввели понятие предпочтительного присоединения, но никак не указали, какую именно модель генерации графа, они рассматривают. Тогда Боллобаш и Риордан

[1] Erdős P., Rényi A. On random graphs I // Publ. Math. Debrecen.1959. V. 6. P. 290-297.

предложили свою спецификацию[2], в которой имеются следующие недостатки – нельзя менять показатель степенного распределения вершин графа, этот показатель не соответствует экспериментальным данным для сети Интернет. Также модели Боллобаша-Риордана плохо распараллеливаются.

В модели копирования также призвана объяснить феномен степенного закона в реальных сетях. Она принадлежит Р. Кумару, П. Рагхавану, С. Раджагопалану, Д. Сивакумару, А. Томкинсу и Э. Упфалу [3]. Основным недостатком модели является то, что она плохо подходит для моделирования социальных сетей. Показатель распределения в этой модели также больше 2, значит, модель плохо подходит для моделирования социальных сетей.

К более гибким моделям безмасштабных сетей относятся модели Чунг-Лу[4] и Янсона-Лучака[5]. Основным преимуществом модели Чанг-Лу является возможность выбрать показатель степенного распределения вершин. Существенным недостатком является плохая распараллеливаемость задачи построения случайного паросочетания на множестве вершин.

В модели Янсона-Лучака преимущество состоит в возможности выбрать показатель степенного распределения.

Основной недостаток модели в том, что ничего не известно об эффективном применении данной модели для больших графов.

К модели стохастических графов относится модель Кронекера. Она основанная на умножении матриц Кронекера, была предложена, как модель, отвечающая многим свойствам реальных сетей. Было установлено, что степени вершин графа, порожденного данной моделью, имеют распределение с тяжелым хвостом, и средняя степень растет с размером графа по степенному

[2] Bollobás B., Riordan O. Mathematical results on scale-free random graphs // Handbook of graphs and networks. Weinheim, Germany: Wiley-VCH, 2003. P. 1–34

[3] Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E. Stochastic models for the web graph // Proc. 41st Symposium on Foundations of Computer Science. Washington, DC, USA: IEEE Computer Society, 2000. P. 57

[4] W. Aiello, F. Chung, L. Lu. A Random Graph Model for Massive Graphs // Experimental Mathematics. 2001. V.10. P. 53-56.

[5] S. Janson, T. Łuczak, I. Norros. Large cliques in a power-law random graph // Journal of Applied Probability. V.47. N.4. P.1124-1135 URL:http://arxiv.org/abs/0905.0561. Дата обращения (1.05.2020).

закону, оставляя диаметр графа ограниченным константой. Также было показано, что графы Кронекера удовлетворяют свойствам графов реальных сетей, таких как Интернет. Достоинством модели является то, что все равбра можно вставлять параллельно, т.е. алгоритм хорошо распараллеливается. Основным недостатком данной модели является ограниченность в задании числа вершин n(степеней двойки), а также то, что графы Кронекера не являются безмасштабными, хотя и имеют малый диаметр[6].

Итак, в данном разделе были рассмотрены основные типы моделей генерации случайных безмасштабных графов. Из них наиболее изученная модель Эрдёша-Реньи была показана Барабаши и Альберт не соответствующей реальным графам, модели Боллобаша-Риордана и модель копирования плохо подходят для моделирования социальных сетей т.к. имеют показатель распределения > 2. Модель графов Кронекера хорошо распараллеливается, но имеет ряд недостатков, приведенных выше. Модели Чунг-Лу, Янсона-Лучака являются достаточно гибкими, позволяя задавать показатель распределения < 2, при этом модель Чунг-Лу обладает хорошо распараллеливаемой аппроксимационной моделью.

Во втором разделе «Модели кластеризации сообществ социальных сетях» рассказывается о возможности кластеризации сообществ в крупных сетях с применением теории графов. Представлены примеры кластеризации с помощью графов которые можно применять в социальных сетях. В данном случае связь – это какие-либо социальные отношения между двумя индивидами. Для этого нужно ввести понятие социальный граф. Социальный граф – это граф, узлы которого представлены социальными объектами, такими как пользовательские профили с различными атрибутами, сообщества и так далее, ребрами в социальном графе являются социальные связи между объектами.

[6] A. Pinar, C. Seshadhri, T. G. Kolda. The Similarity between Stochastic Kronecker and Chung-Lu Graph Models // SDM12: Proceedings of the Twelfth SIAM International Conference on Data Mining. Philadelphia,USA:SIAM, 2012. P.1071-1082 URL: http://arxiv.org/abs/1110.4925. Дата обращения (5.05.2020).

Социальным графам присущ следующий набор свойств:

1) Одна большая общая компонента связности. В большинстве социальных графов присутствует одна большая компонента связности, которая захватывает большинство вершин. Остальные компоненты гораздо меньшего размера.

2) Распределение на степенях вершин. Социальные сети относятся к так называемым безмасштабным сетям.

Эмпирически было установлено, что многие естественно возникающие сети – социальные, коммуникационные, биологические, графы цитирований ссылок в Интернете и другие системы – хорошо моделируются безмасштабными графами.

3) Среднее расстояние. Под расстоянием между вершинами понимают минимальную длину цепи в графе, соединяющие эти вершины. Социальные сети в среднем имеют очень маленькое расстояние между двумя случайными вершинами. Существует множество коэффициентов для графа, которые показывают некоторую его качественную характеристику.

4) Коэффициент кластеризации. Существует множество коэффициентов для графа, которые показывают некоторую его качественную характеристику.

В работе мы рассматриваем алгоритм, основанный на поиске оптимального значения функции модулярности. Методы кластеризации, основанные на максимизации модулярности являются одними из самых популярных среди алгоритмов, позволяющих автоматически определять количество кластеров. Задача таких методов – поиск оптимального разбиения, максимизирующего значение функции модулярности.

В третьем разделе «Выделение сообществ в больших сетях» приводится пример применения кластеризации и анализа крупного сообщества на основе данных представленных в работе Д. Блонделья, Жан-Лу Гийома, Рено Ламбиотти и Этьен Лефевра.

В данном разделе мы рассмотрим метод развертывания сообществ в больших сетях. Это эвристический метод, основанный на оптимизации

показателя модулярности. Показано, что он превосходит любой другой известный метод обнаружения сообществ в терминах времени вычисления. Модулярность используется как целевая функция для оптимизации. Авторы статьи «Быстрое развертывание сообществ в крупных сетях» Винсент Д. Блондель, Жан-Лу Гийом, Рено Ламбиотт и Этьен Лефевр [7] внедрили этот алгоритм оптимизации модулярности, который позволяет изучать сети большого размера.

Перспективный подход заключается в декомпозиции сетей на подгруппы или сообщества, которые являются наборами узлов с высокой степенью взаимосвязи. Суть этапов алгоритма: на первом этапе модулярность оптимизируется за счет допуска только локальных изменений сообществ; на втором этапе найденные сообщества агрегируются, чтобы построить новую сеть сообщества. Шаги повторяются итеративно до тех пор, пока не произойдет допустимо возможное увеличение модулярности. Самый быстрый аппроксимационный алгоритм оптимизации модулярности на больших сетях был предложен Аароном Клаусетом, Дж. Ньюманом и Кристофером Муром в работе «Finding community structure in very large networks»[8].Этот метод состоит в периодическом объединении сообществ, которые оптимизируют модулярность. Аарон Клаусет, Дж. Ньюман и Кристофер Мур применили данный алгоритм для большой сети, построенной по записям бельгийской компании мобильной связи, подсчитали общее количество телефонных звонков в течение 6-месячного периода. Сеть состояла из 2,6 млн. клиентов, между которыми есть связи, каждый клиент идентифицировался ключом, с которым связаны несколько записей, например, его возраст, пол, язык и почтовый индекс места проживания. Применив алгоритм к данной большой сети, было выявлено, что в Бельгии два основных языковых сообществ: французский и

[7] Fast unfolding of communities in large networks, Vincent D. Blondel;JeanLoup Guillaume, Renaud Lambiotte and Etienne Lefebvre URL: https://arxiv.org/abs/0803.0476 Дата обращения (10.05.2020)

[8] Clauset A., Newman M., Moore C. Finding community structure in very large networks // Physical Review. 2004. E 70. 066111

голландский. С социологической точки зрения, возможность выделить лингвистическую, религиозную или этническую однородность сообществ открывает перспективы для описания социальной сплоченности и потенциальной хрупкости страны. Также авторы заметили одно интересное наблюдение, связанное с наличием других языков. На самом деле существуют четыре языка для объединения клиентов этого конкретного оператора мобильной связи: французский, голландский, английский или немецкий. В то время как англоговорящие клиенты распределяются достаточно равномерно во всех 44 сообщетсвах, более 60% немецкоязычных клиентов сконцентрированы в одном сообществе. Это, вероятно, связано с тем, что говорящие по-немецки люди в основном сосредоточены в небольшом регионе недалеко от Германии, в то время как англоговорящие люди распространены по всей стране. Применение выше описанного метода может сильно помочь при проведении социологических исследований. Число пользователей социальных сетей ежедневно увеличивается и изучение алгоритмов кластеризации больших сообществ становится одной из самых актуальных проблем современной социологии.

ЗАКЛЮЧЕНИЕ

В представленной работе рассмотрена задача кластеризации в применении к выделению связанных структур (кластеров) в социальных сетях. Это может свидетельствовать о том, что исследуемые метрики и методы в меньшей мере приспособлены к извлечению неявных знаний о внутренних закономерностях и структуре данных в социальных сетей. Выявленные структурные характеристики достаточно сложно продемонстрировать с помощью формальных математических параметров. Визуализация же системы отношений (дружба, подписка) демонстрирует не только наличие плотно взаимодействующих (сплоченных) подгрупп (ядер в сети), но и связь между характеристиками реального и сетевого информационного пространства. Самый тесный контакт (близкое расположение узлов при визуализации) в

виртуальном пространстве наблюдается между пользователями, которые зарегистрированы в одном географическом регионе.

Неочевидным, но визуально представленным выводом стала географическая обусловленность плотности взаимодействия пользователей в группах. Этот вывод, к сожалению, можно ярко продемонстрировать только при наличии цветного изображения социальной сети, где цвет вершины соответствует месту жительства пользователя.