ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Образовательная программа
«Прикладная математика и информатика»

КУРСОВАЯ РАБОТА

На тему: Различные способы построения графовых структур данных для поиска ближайшего соседа

Тема на английском: Several Ways to Form Graph Based Nearest Neighbour Search Structure

Студент / студентка 2-го курса
группы № БПМИ2110/21ПМИ-2:
Рябков Игорь Дмитриевич
(Ф.И.О.)
Научный руководитель:
Пономаренко Александр Александрович
(Ф.И.О.)
Доцент, НН Кафедра прикладной математики и информатики
(должность, звание)

Содержание

1 Введение

В последнее десятилетие наша жизнь стала тесно связана с удобными приложениями и сервисами. Многие из них базируются на рекомендательных системах, для предоставления целевого товара на основе наших интересов. Некоторые используют компьютерное зрение для решения огромного кол-ва задач (от масок в социальных сетях, до автопилота в электроавтомобилях). Поиск синонимов и системы автоматического дополнения текста(T9) также используются каждый день миллионами пользователей. Это лишь малая часть задач, которые можно решить используюя алгоритм поиска ближайшего соседа (или поиска K-ближайших соседей).

Вот ещё некоторые задачи, о которых хотелось бы упомянуть:

- Поиск дубликатов (определить являются ли 2 текстовых документа одинаковыми)
- Задача кластеризации (Определить, как какой группе относится выбранный объект)
- Поиск ближайших географических объектов (карты)
- Поиск схожих фрагментов в фильмах или музыке

Именно поэтому так важно искать новые подходы для улучшения скорости данного алгоритма. Чтобы достичь поставленную цель, необходимо разработать структуру данных, которая сможет наиболее эффективно осуществлять две операции добавления и поиска. Вариантов подходящих структур - огромное множество. Например, некоторые могут быть построены на базе вектора, списка, дерева, графа (в виде сети). Некоторые поддерживают точные поиск, а некоторые только приближённый. Некоторые формируются по средствам детерминированных алгоритмов, а некоторые используют рандомизированный подход.

Моё исследование буддет в основном основываться на изучении графовых структур (в виде сетей), так как они более современные и эффективные (подробнее ниже). Перед собой я ставлю следующие задачи:

- Изучить какие структуры для поиска ближайшего соседа существуют
- сравнить их асимптотику построения этих структур и поиска внутри них, выявить преимущества и недостатки
- Исследовать феномен тесного мира.
- Обосновать выбор именно графовых структур, а также подробнее исследовать те из них, которые имеют свойствр тесного мира.

- Выдвинуть несколько предположений о модификациях, которые смогли бы улучшить имеющиеся методы.
- Разработать необходимые абстракции для работы с подобными структурами
- На основе сравнительного анализа определить наилучшую конфигурацию параметров в разработанных классах для наибольшей эффективности алгоритма поиска ближайшего соседа
- А также проверить эффективность предложенные ранее модификации

2 Основная часть

2.1 История развития структур для поиска ближайшего соседа

2.2 Феномен тесного мира

Для того, чтобы понять, каким должен быть граф для оптимальной работы нашего алгоритма, я предлагаю обратиться к структурам, которые возникают само собой в природе, к графам, которые формирует общество.

2.2.1 Теория 6 рукопожатий

Венгерский писатель Karinthy Frigyes Ernő в 1928г в рассказе "Звенья цепи"впервые сформулировал данную проблему. Согласно ей, любые два человека на планете связаны через 5-6 общих знакомых.

Рассуждая над этой проблемой, Stanley Milgram - американский социальный психолог и педагог в своей статье "The Small World Problem"описал проведённый им эксперимент: Он выдал 300 писем жителям из разных городов и попросил доставить их одному человеку из Бостона (США). Важным условиям было то, что люди могли передавать письма только своим знакомым, которые по их мнению могли знать целовека-цель. По результатам исследовани даже не смотря на то, что до места назначения дошли далеко не все письма, те, которым это удалось, прошли в среднем через цепь из 5-6 человек.

Данные наблюдения кажутся очень полезными для нас. Ведь если нам удастся воссоздать граф, который с хорошей точностью моделируют общественные сети, мы сможем из любой вершины графа добираться до цели за сравнимо малое число посредников. Причём для этого нам даже не придётся использовать сложные алгоритмы поиска. Вспомним эксперимент Милгрема, в нём каждый человек не пытался искать оптимальный путь от него до цели, он лишь отправлял письма тем знакомым, которые казались ближе к месту назначения. Формально - это простой жадный поиск, который в эксперименте Милгрема прошёлся всего лишь по 5-6 промежуточным значениям и, что не мало важно, достиг цели.

2.2.2 Случайные графы Erdős–Rényi

Попробуем предположить, что общественные сети можно представить, как случайный граф $G(n,p) = V, E$ состоящий из n вершин, с вероятностью проведения ребра p. Логично предположить, что $np = const = \lambda$. Можно объяснить это так: структура графа не должна зависить от кол-ва вершин в нём. Он всегда должен выглядеть примерно одинаково ($deg(v) \approx$

const). Ну или ещё одно описание: кол-во моих друзей не увеличится, если население планеты увеличится в несколько раз. Обоснуем это формально:

$$P(deg(v) = k) = \binom{n}{k} p^k (1-p)^{1-k} \Rightarrow deg(v) \sim Bin(p) \tag{1}$$

Так как $np = const = \lambda : \mathbb{E}[deg(v)] = np = \lambda$

Так мы описали граф, изучением которого занимались два великих Венгерских математика Paul Erdős и Alfréd Rényi. (Причём λ обычно выбирается много меньше чем n, ведь наш круг общения никчёмно мал по сравнению с 8 млрд. людей планете)

Разультаты, к которым пршли данные математики показали, что случайные графы имеют сравнительно малый средний диаметр графа $d(G)$. Где

$$d(G) = \frac{1}{|V|} \sum_{u,v \in V} d(u,v) \tag{2}$$

$$d(u,v) = \min_{k \in \mathbb{R}} \{\exists a_i,\ i = \overline{1k} : a_1 = u, a_k = v, (a_i, a_{i+1}) \in E\} \tag{3}$$

Казалось бы, малый диаметр - всё, что нам нужно. И это было бы так, если бы мы не использовали жадный алгоритм для поиска. Случайные рёбра в графе помогают нам легко и быстро приблизиться к месту назначения, однако отсутствие локальных рёбер приводит к большому кол-ву локальных минимумов, что ведёт за собой большую погрешность полученном результате, пример:

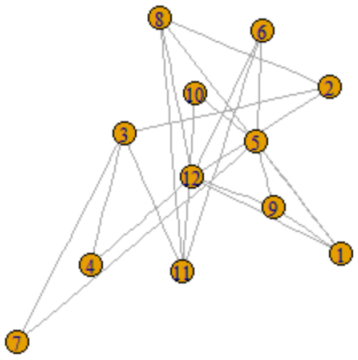

Рис. 1: Случайный граф

Взглянув на рисунок 1, можно заметить, что если мы хотим добраться до вершины 3, используя жадный поиск, то это нам удастся только из непосредственных соседей вершины номер 3. Это говорит о том, что такой граф хорош в навигации на больших масштабах, но плох при локальном поиске. Что не состыкуется с теорией 6 рукопожатий

Ошибка в наших рассуждениях была в том, что мы предположили, что связи между людьми абсолютно случайны, но ведь это не так. Милгрэм, описывая результаты своего

эксперимента, делал акцент на то, что общественные графы состоят из кластеров. Это можно описать так: Вероятность знакомства с человеком тем больше, чем мы ближе; А также, вероятность, того, что люди дружат прямо пропорциональна кол-ву их общих друзей. Наличие данной концепции образовало бы больше локальных рёбер, что решило бы проблему локальных минимумов. Формально это можно описать, через коэффицент кластеризации вершины:

$$cc(v) = \frac{\#\{(x,y) \in E,\ x,y \in V : (x,v),(y,v) \in E\}}{\#\{(x,y),\ x,y \in V : (x,v),(y,v) \in E\}}$$

По аналогии коэффицент кластеризации графа:

$$cc(G) = \frac{1}{|V|}\sum_{v \in V} cc(v)$$

Чем он больше, тем сильнее кластеризован наш граф. Давайте докажем, что в случайном графе он стремится к 0 при больших n.
Для начала посчитаем $\mathbb{E}[cc(v)]$. Обозначим $N = \binom{deg(v)}{t},\ t \leq deg(v)$. Тогда:

$$P(N * cc(v) = t) = \binom{N}{t} p^t (1-p)^{N-t} \Rightarrow N * cc(v) \sim Bin(p)$$

Тогда $\mathbb{E}[N * cc(v)] = Np \Rightarrow \mathbb{E}[cc(v)] = p : \forall N \in \mathbb{N}$ Заметим, что $cc(v)$ хоть и распределены одинаково, однако не независимы. Я утверждаю, что мы этим можем пренебречь, так как $k \ll n$. Данная гипотеза будет обоснована позднее.
Если считать, что $cc(v)$ - i.i.d, то можно использовать УЗБЧ:

$$cc(G) = \overline{cc(v)} \xrightarrow{a.s.} \mathbb{E}[cc(v)] = p = \frac{\lambda}{n} \to 0,\ n \to \infty$$

Теперь сформулируем гипотезу о независимости:

$$H_0 : cc(u_i) - iid,\ \forall u \in V : k \ll n$$

$$H_1 : alternative$$

2.2.3 Тесные графы

Часто истина кроится по середине, поэтому в современное обоснование проблемы 6 рукопожатий говорит о том, что знакомства в обществе можно представить в виде графа со следующими свойствами:

- У каждой вершины есть большое кол-во близких связей (решают проблему локальных минимумов)

- У каждой вершины есть ограниченый набор длинных вершин (решают проблемы с навигацией по графу)

Формализуем данные понятия, чтобы в дальнейшем было удобно сравнивать эти характеристики численно:

2.3 Подход Клайнберга

Клайнберг предложил немного иной подход, к созданию

2.4 Подход Пономаренко

2.5 Предложени по модификации

3 Практическая часть

3.1 Описание абстракций

3.2 Эксперименты и наблюдения(параметры)

3.3 Проверка ранее вывдвинутых гипотез

4 Заключение