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 формирет общество.

2.2.1 Виденье Stanley Milgram

Stanley Milgram - американский социальный психолог и педагог. в своей статье The Small World Problem [1] Милгрэм рассуждает на тему гипотезы 6 рукопожатий, которая гласит, что любые два человека на планете могут быть связаны через 6-х знакомых. Он обосновывал правдивость этой гипотезы тем, что наше общество состоит из кластеров. То есть, каждый из нас имеет примерно 500 близких знакомых, каждый из которых имеет ещё 500 знакомых(другие кластеры) и так далее. Если посмотреть на полученную геометрическую прогрессию, мы получаем число $500^6 = 1,56 * 10^{16}$. Даже с учётом того факта, что кластеры могут пересекаться, настолько большой результат выглядит очень убедительно

Для подтверждения своей модели, Милгрэм провёл эксперимент Он выдал 300 писем людям из разных городов и попросил доставить их одному человеку из Бостона (США). По результатам исследование, даже не смотря на то, что до человека-цели дошли далеко не все письма (Милгрэм оправдал это тем, что люди имели недостаточно информации о цели, поэтому иногда принимали не оптимальные решения), те, которым это удалось, прошли в среднем через цепь из 5-6 человек.

Однако, при попытке смоделировать данное поведение на компьютере, даже на небольшой выборке значений, это не сработает. Если мы построим граф, который будет состоять из кластеров вершин (вершины, соеденины ребром только в случае, если они находятся рядом), его диаметр будет всё равно очень большим. То есть, чтобы добраться от одной вершины к другой, понадобиться много посредников

2.2.2 Случайные графы Erdős–Rényi

Попробуем подойти к проблеме с другой стороны и предположить, что графы на самом деле полностью случайны. Для простоты будем считать, что вероятность проведения

каждого ребра одинаково и выбирается так, чтобы произведение кол-ва вершин n на неё было фиксировано и равно средней степени вершин. Так мы описали граф, изучением которого занимался Венгерский математик Erdős–Rényi.

Первый вопрос, который может возникнуть: а будет ли такой граф вообще связен? Ведь если нет, то говорить о "кол-ве рукопожатий" вовсе не будет иметь никакого смысла. Оказывается, что да. Erdős доказал, что

$$c \geq 3,\ n \geq 300,\ p = \frac{c \ln n}{n} \Rightarrow \mathbb{P}(\text{G - связен}) \xrightarrow{a.s.} 1 \tag{1}$$

Что буквально говорит о том, что при увеличении выборки вероятность связности графа только растёт

Второй же вопрос тоже назревает сам собой: Чем данный подход может превосходить предыдущий? Идея заключается в том, что так как вероятность рёбер никак не зависит от расстояния между вершинами, будет появляться значительное кол-во связей на средние и далёкие расстояния, что будет способствовать быстрому перемещению по графу.

Однако такой подход тоже не идеален, проблема возникла в другом. Теперь мы легко можем добраться от одной части графа к другой, но столкнёмся с тем, что из-за отсутствия большого кол-ва локальных связей поиск тех или иных вершин будет часто упираться в локальные минимумы из которых не будет выхода

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

Список литературы

[1] [Online; accessed 15. May 2023]. Янв. 2020. URL: http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf.