Шацкий Ростислав Евгеньевич, студент,
МГТУ им. Н. Э. Баумана, Москва

Быстрицкая Анна Юрьевна, к.т.н.,
МГТУ им. Н. Э. Баумана, Москва

ОБОБЩЕННЫЙ МЕТОД ФИЛЬТРАЦИИ ТОЧЕЧНЫХ ОБЪЕКТОВ ПО КРИТЕРИЮ ПРИНАДЛЕЖНОСТИ К ОБЛАСТИ ПЛОСКОСТИ НА ОСНОВЕ РЕКУРСИВНОГО РАЗБИЕНИЯ

Аннотация: в работе представлена разработка обобщенного метода фильтрации объектов по критерию принадлежности к области плоскости на основе рекурсивного разбиения.

Ключевые слова: критерий принадлежности к области, индексные структуры данных, рекурсивные операции, оптимизация вычислений.

Введение

Фильтрация точечных объектов по критерию принадлежности к области плоскости – задача, с которой можно столкнуться во множестве предметных областей, например, компьютерная графика, анализ геометрических и географических данных. В случае наличия достаточно большого количества объектов целесообразным становится не только использование эффективного по времени и памяти алгоритма для проверки вхождения объекта в одну область, но и использование алгоритма для фильтрации тех областей, в которые объект не входит по каким-либо «простым» свойствам, определенными заранее для каждой области.

Метод Geohash

Метод Geohash используется при работе с географическими координатами [1], но может быть распространен на любую область с заранее известными границами.

Область разбивается на 32 равных по размеру (в градусах угла) области, каждая из которых обозначается определенным символом в заданном алфавите.

Затем каждая подобласть также разбивается на 32 равных областей и аналогичным способом обозначается определенным символом в заданном алфавите [2].

Процесс продолжается до тех пор, пока не будет достигнута заданная точность вычислений.

Для каждой точки с заданными координатами можно определить хеш-значение, рекурсивно определяя область, к которой она принадлежит. Для каждого многоугольника можно определить одно или несколько хеш-значений, которым этот многоугольник принадлежит.

В системе Geohash, область которой – карта Земли, точность кодирования координаты при длине хеш-значения в 17 символов достигает $4{,}55$ нм [3].

Данный метод можно также использовать для кластеризации объектов на области [4]: каждый кластер может быть представлен хеш-значением определенной длины N, а объекты группируются по первым N символам их хеш-значения.

Особенности обобщенного метода

Поскольку разрабатываемый метод является расширением метода на основе рекурсивного разбиения области, для его работы необходимы заданные границы области. В случае Geohash границы области – $[-90°, 90°]$ по широте и $[-180°, 180°]$ по долготе.

Разрабатываемый метод должен предоставлять возможность выбора схем разбиения области на подобласти с целью оптимизации выполнения поисковых запросов в наиболее нагруженных областях плоскости. Такая оптимизация может являться необходимой лишь на определенных уровнях хеша, в то время как все остальные уровни должны быть вычислены идентично. Также необходимо задание максимального уровня вычисляемого хеша.

На рисунке 1 представлена детализированная схема разрабатываемого метода.

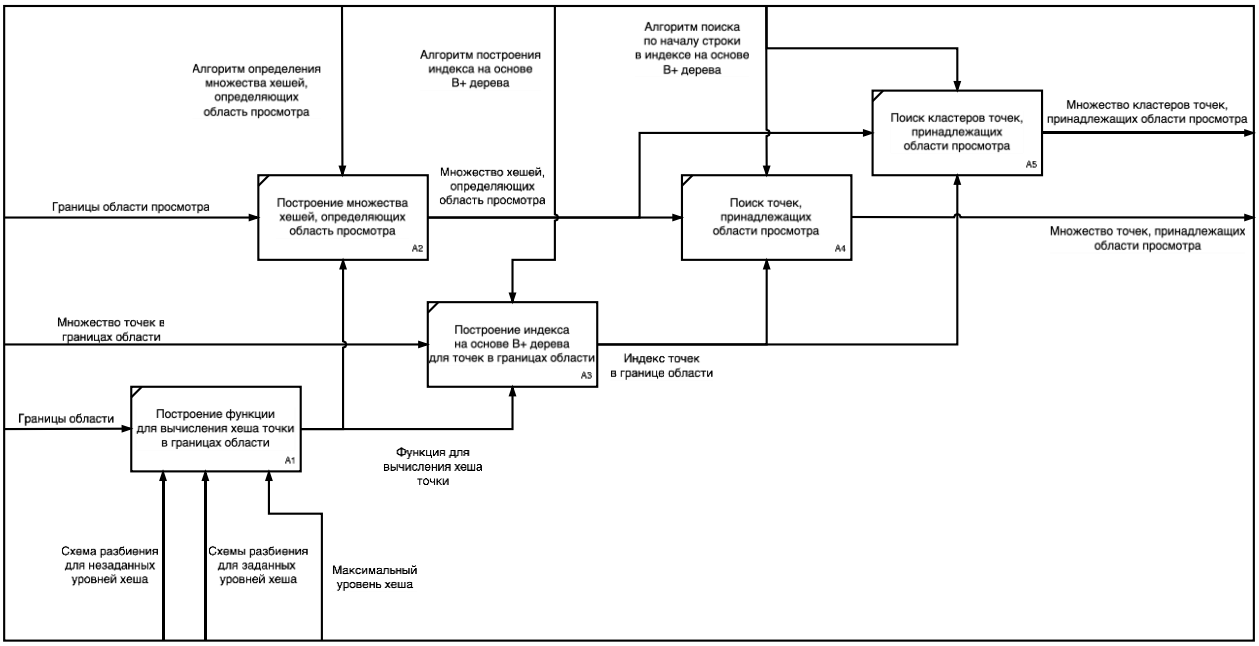

Рис. 1 – Схема обобщенного метода

Типы разбиений на подобласти

Самое простое разбиение области задается количеством разбиений по осям x и y. Например, если область разбивается на 5 подобластей по оси x и на 3 области по оси y, в результате будет 15 подобластей одинакового размера. Количество разбиений по осям x и y будут обозначаться как nx и ny соответственно. Пример такого разбиения изображен на рисунке 2.

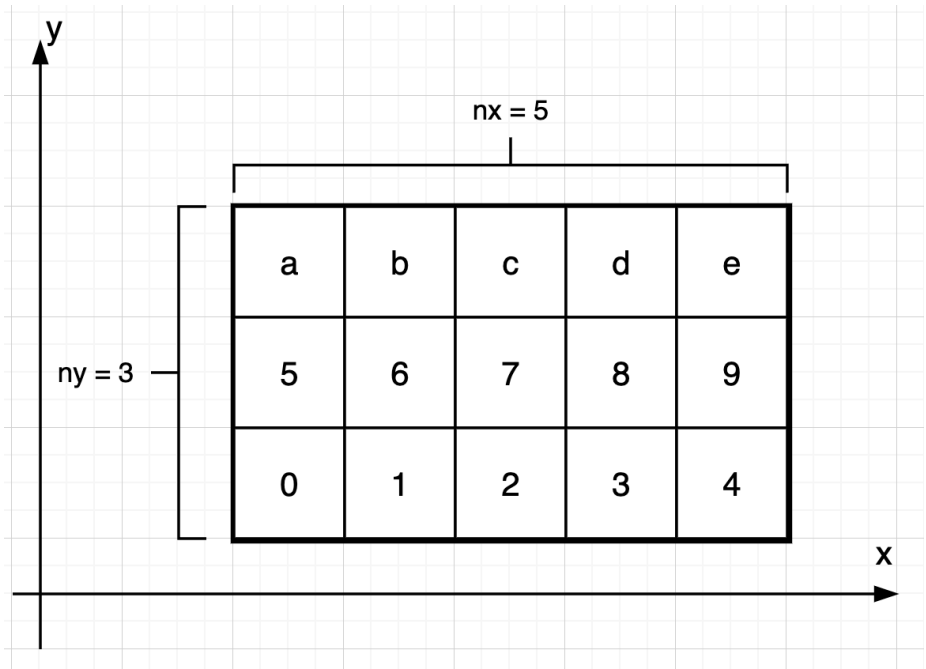

Рис. 2 – разбиение области с $nx = 5$ и $ny = 3$.

Более сложное разбиение, в котором не все подобласти одинаковы, можно задать объединением нескольких соседних подобластей в простом разбиении. На рисунке 3 изображено такое разбиение, отличающееся от рассмотренного ранее объединениями подобластей $\{0, 1\}$, $\{2, 3\}$, $\{4, 9\}$, $\{5, 6\}$, $\{7, 8\}$, $\{a, b\}$, $\{c, d\}$.

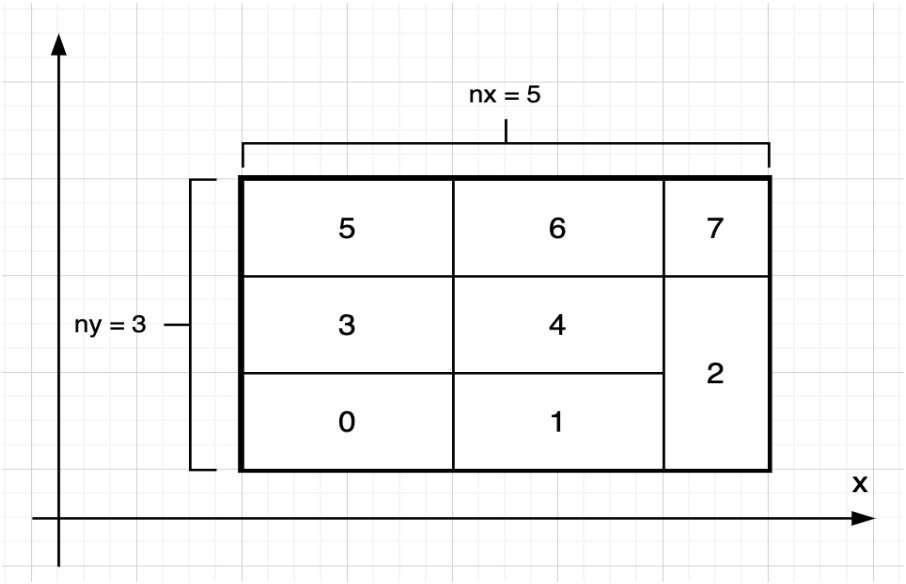

Рис. 3 – Пример разбиения на неравные подобласти

Последовательность разбиений на подобласти

Несколько разбиений области на подобласти объединяются в одну «цепочку» разбиений. Таким образом, если цепочка состоит из двух разбиений:

1) A, разбивающее изначальную область на подобласти $a_1, a_2, \dots, a_n$, где n – количество подобластей в разбиении A;

2) B, разбивающее каждую из подобластей $a_1, a_2, \dots, a_n$ на подобласти $a_i b_1, a_i b_2, \dots, a_i b_k$, где $i = \overline{1,n}$, k – количество подобластей в разбиении B, то максимальная длина хеш-значения при таких разбиениях равна длине цепочки – 2. Разбиение B при этом используется для разделения подобластей с предыдущего разбиения, тем самым очередные подобласти получаются гарантированно меньшего размера. Такие разбиения будут называться *углубляющими*.

Пример такого разбиения изображен на рисунке 4.

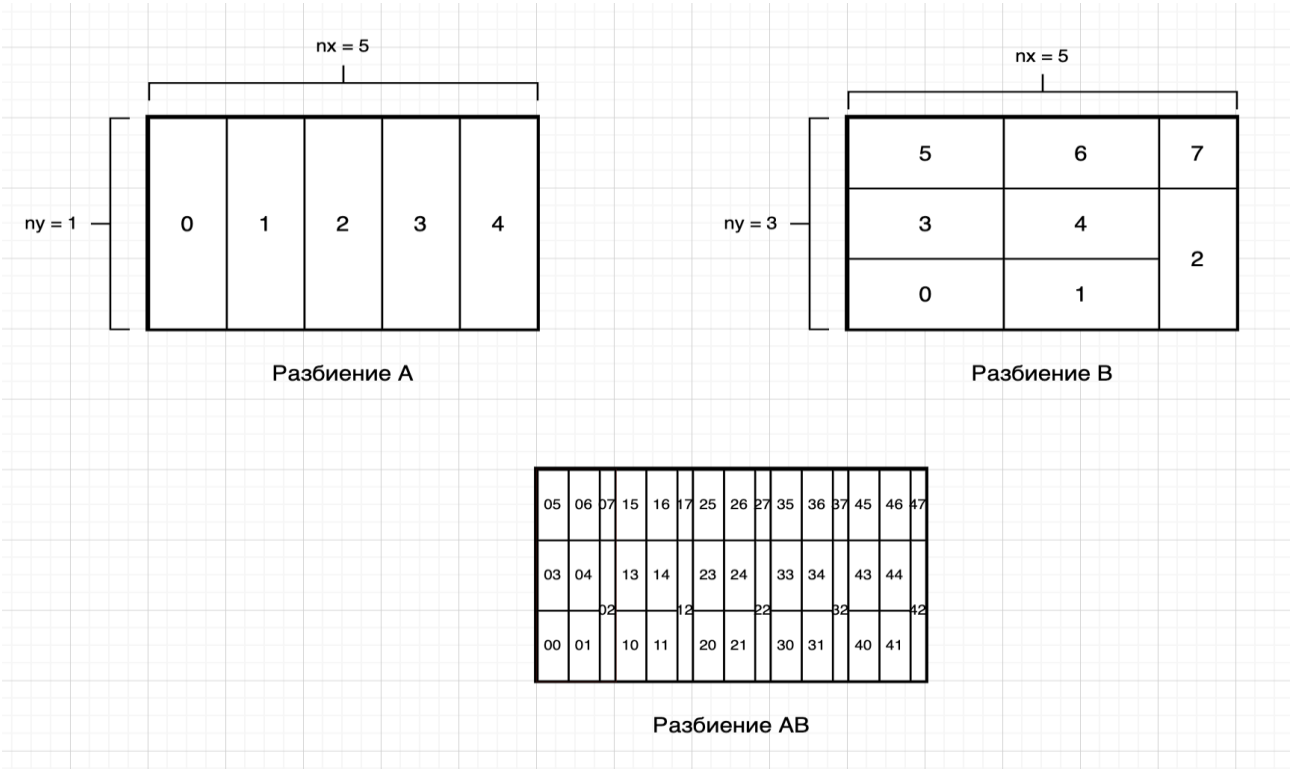

Рис. 4 – Цепочка разбиений AB

Необходимо также предусмотреть ситуацию, когда несколько разбиений применяются к одной и той же области. Такой сценарий предоставляет дополнительную гибкость в составлении цепочки разбиений. Пример такой цепочки разбиений представлен на рисунке 5. Такие разбиения, которые применяются к одной области, будут называться *уточняющими*.

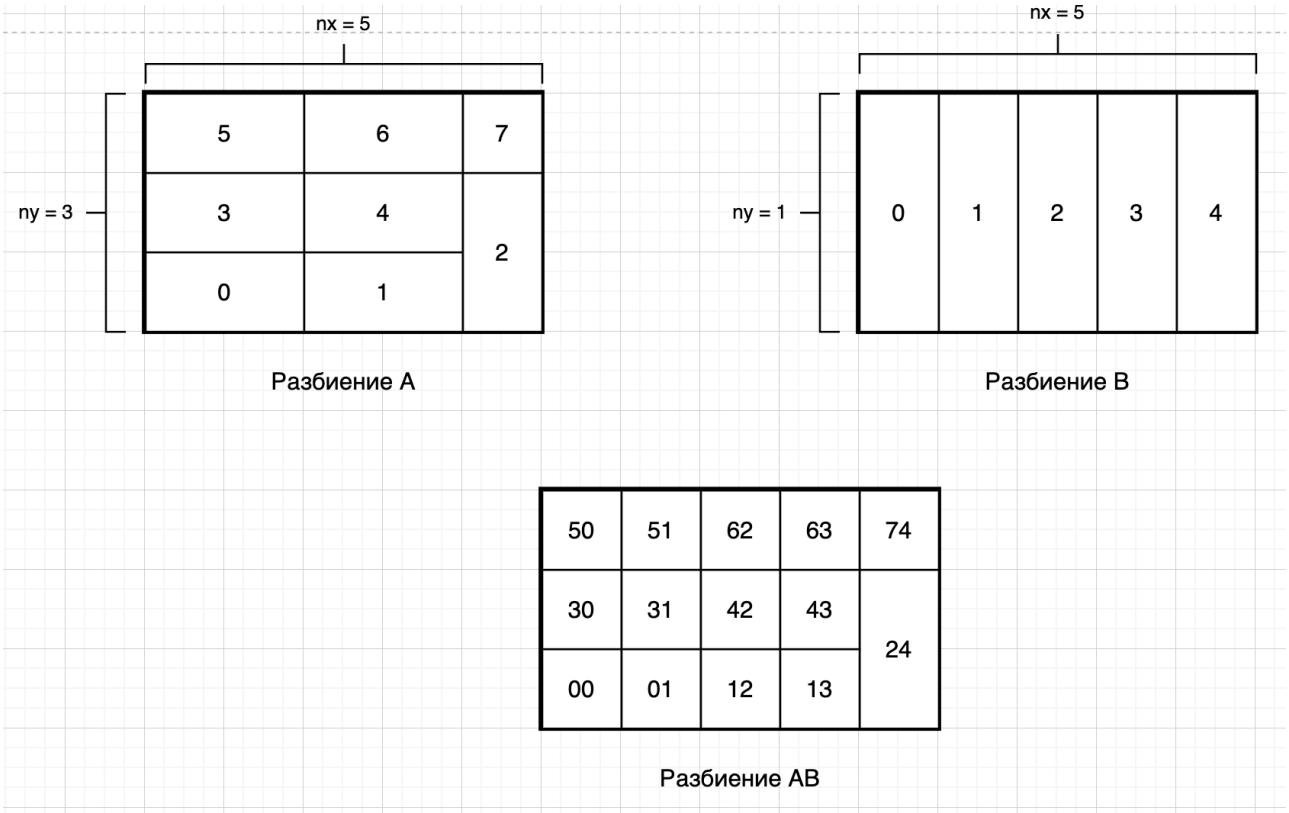

Рис. 5 – Пример применения двух разбиений к одной области

Алгоритм определения хеш-значения точечного объекта

Каждое разбиение характеризуется значениями nx и ny, определяющими минимальный размер одной подобласти по соответствующей оси относительно размера всей разбиваемой области. Если координаты левой верхней вершины области – (x_1, y_1), координаты правой нижней вершины области – (x_2, y_2), то для разбиения A с известными nx и ny и заданной точки (x, y) можно определить, к какой подобласти относится эта точка:

$$l_x = \frac{x_2 - x_1}{nx},$$
$$l_y = \frac{y_1 - y_2}{ny},$$
$$c_x = \left\lfloor \frac{x - x_1}{l_x} \right\rfloor,$$
$$c_y = \left\lfloor \frac{y - y_2}{l_y} \right\rfloor,$$

где l_x, l_y – длины одной подобласти по оси x и y соответственно, c_x, c_y – определяют «координаты» подобласти, в которую попадает точка.

На рисунке 6 координаты точки A – $(3.6, 2.7)$, $x1 = 2$, $y_1 = 4$, $x2 = 7$, $y2 = 1$, $nx = 5$, $ny = 3$.

Тогда $l_x = \frac{7-2}{5} = 1$ и $l_y = \frac{4-1}{3} = 1$.

$c_x = \left\lfloor \frac{3.6-2}{1} \right\rfloor = \lfloor 1.6 \rfloor = 1$ и $c_y = \left\lfloor \frac{2.7-1}{1} \right\rfloor = \lfloor 1.7 \rfloor = 1$.

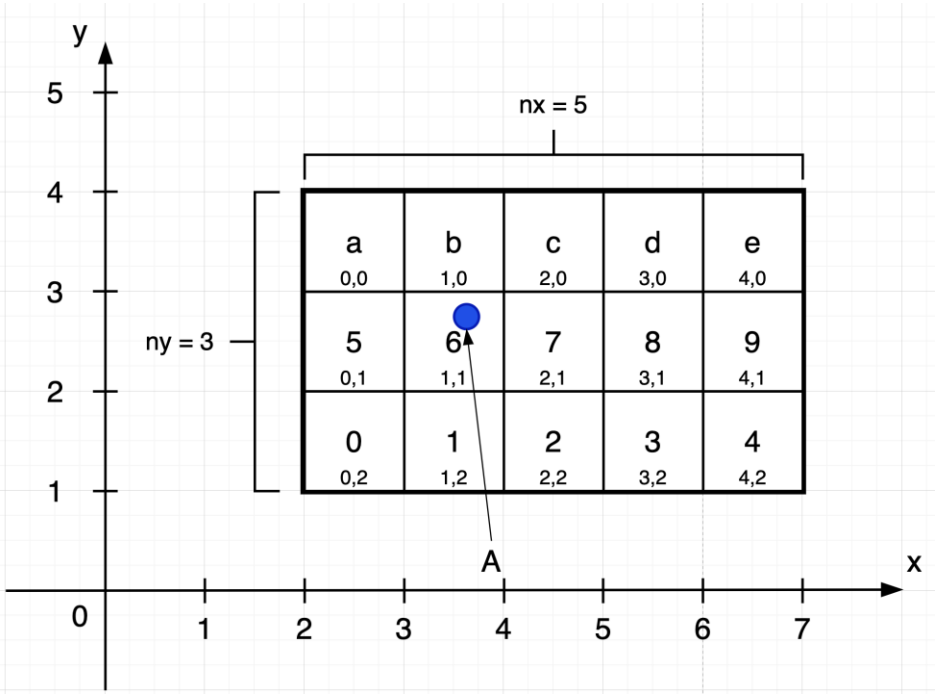

Рис. 6 – Вычисление подобласти для точки A.

По полученным координатам подобласти можно получить хеш-значение подобласти, в которую попадает точка A. Для этого можно использовать хеш-таблицу, ключ в которой – пара координат c_x, c_y, а значение – хеш-символ соответствующей подобласти.

Аналогичный подход можно распространить и на разбиения, в которых подобласти могут иметь разные размеры.

Список литературы:

1. Geohash [Электронный ресурс]. – URL: http://geohash.org/site/tips.html (дата обращения 02.03.2025).

2. Anthony F., Chris E., James H., Skylar L. Spatio-temporal Indexing in Non-realtional Distributed Databases. – International Conference on Big Data. – 2013. – С. 291–299.

3. Microsoft geo_point_to_geohash [Электронный ресурс]. – URL: https://learn.microsoft.com/ru-ru/azure/data-explorer/kusto/query/geo-point-to-geohash-function (дата обращения 03.03.2025).

4. Воробьев А.В., Воробьева Г.Р. Геоинформационная система динамической пространственной кластеризации распределенных источников данных. – Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2023. – № 64. – С. 61–73.