count-min sketch (concept map)

Central node: count-min sketch

Spaces branch (top)

- There are three types of spaces in the graph of memory vs data
  - super-linear space
  - linear space
  - sub-linear space

- **super-linear space** — is a → linear space over a field K with a given decomposition of subspaces of grade 0 and grade 1 → its study is → known as superliner algebra

- **linear space** — It is → Also known as vector space; consists of → A set of elements called points, and a set of elements called lines

- **sub-linear space** — the space is sublinear in data and input size, the data structure used cannot be much accurate → so the operations will be → approximate in nature → Here → sketch word means 2-dimensional matrix → We have → predefined size of the matrix that's why it is called a sublinear

- used for processing the queries → the two basic operations used to answer point queries → are → counting first and computing the minimum next from the matrix of data.

Advantage branch (right)

- Advantage is → size of this data structure will never change based on given input we are getting to calculate the frequency of the given element → It fails → where there is a hash collision → if → when more elements got the same hash values in that case either gets maximum because of hash collision → Means → less number of hash functions we take, there will be a high probability of collision. → It is advisable → To take more hash functions for better result

why branch (right)

- why → As the data grows the usual techniques that use hash tables are not feasible. They take a lot of memory and time to process the data. → So to tackle → the memory inefficiency model, we can use count-min sketch to calculate the frequency in sub-linear space → in which → we will not be storing the complete values of the data stream, instead, we will use a matrix to compute the frequency

How branch (bottom right)

- How → Count min sketch allows calculating frequency of large data streams in sublinear space using the same DES-compact time completely → In Count Mean Sketch → Sketch word means 2-dimensional matrix. Count means counting hash values, and min means to take the minimum of that event's counts over all of the hash functions → Where → number of rows is equal to no of hash functions. The number of columns depends on a minimum number which hash function gives as output → To find → frequency of given element we need to pass that data to all hash functions from the matrix → initially → every cell in the Count-Min Sketch is initialized to 0 → When → an event occurs, the event's id is hashed over every column. Each hash function outputs a new value, and the counter at each resulting row-column combination is incremented → Then → The minimum of that event's counts over all of the hash functions

possess branch (bottom center)

- possess → Probabilistic data structure. → that solves → the problem in an approximate way → it → a frequency table of events in a stream of datas.

Applications branch (left)

- has → Applications
  - machine learning
  - Natural Language Processing
  - computer networking → Majorly faces → Heavy Hitters Problem
  - Database Queries

- **Heavy Hitters Problem**
  - No efficient method to detect → As they are very large in number counts min sketch can be used for approximation
  - Is used → Identify all of the -heavy hitters for an input parameter
  - Can be → Either ip address/port or both

- **Database Queries**
  - are → Used to retrieve information from database tables.
  - can be → Used within fields of database.
  - Running them → can be made considerably cheaper by using the Count-Min Sketch → it → provides a relatively precise frequency approximation of terms in the table R

miro

Advantage is → size of this data structure will never change based on given input we are getting to calculate the frequency of the given element → **It fails** → where there is a hash collision → **if** → one or more elements got the same hash values in that case value grab increases because of hash collision → **Means** → less number of hash functions we take, there will be a high probability of collision. → **It is advisable** → To take more hash functions for better result

why → As the data grows the usual techniques that use hash tables are not feasible. They take a lot of memory and time to process the data. → **So to tackle** → the memory inefficiency model, we can use count min sketch to calculate the frequency in sub-linear space → **in which** → we will not be storing the complete values of the data stream, instead, we will use a matrix to compute the frequency

How → count min sketch allows calculating frequency of large data streams in sublinear space using the same O(1) constant time complexity → **In Count Mean Sketch** → sketch word means 2-dimensional matrix. Count means counting hash values and min means to take the minimum of that event's counts over all of the hash functions → **Where** → number of rows is equal to no of hash functions The number of columns depends on a maximum number which hash function gives as output → **To find** → frequency of given element we need to pass that data to all hash functions from the matrix → **initially** → every cell in the Count-Min-Sketch is initialized to 0 → **When** → an event occurs, the event's id is hashed over every column. Each hash function outputs a row value, and the counter at each resulting row-column combination is incremented → **Take** → The minimum of that event's counts over all of the hash functions

miro

possess

Probabilistic data structure.

that solves

the problem in an approximate way

as

a frequency table of events in a stream of datas.

known as superliner algebra

Its study is

Also known as vector space

It is

linear space over a field K with a given decomposition of subspaces of grade 0 and grade 1

is a

A set of elements called points, and a set of elements called lines

consists of

super-linear space

linear space

sub-linear space

the space is sublinear in data and input size, the data structure used cannot be much accurate

so the operations will be

approximate in nature

Here

sketch word means 2-dimensional matrix

We have

predefined size of the matrix that's why it is called a sublinear

used for processing the queries

the two basic operations used to answer point queries

are

counting first and computing the minimum next from the matrix of data.

There are three types of spaces in the graph of memory vs data

machine learning

Natural Language Processing

As they are very large in number ,count min sketch can be used for approximation

No efficient method to detect

Identify all of the -heavy hitters for an input parameter

Is used

Heavy Hitters Problem

Majorly faces

computer networking

Applicatitons

has

Either ip address/port or both

Can be

Database Queries

are

can be

Running them

Used to retrieve information from database tables.

Used within fields of database.

can be made considerately cheaper by using the Count-Min Sketch

It

provides a relatively precise frequency approximation of items in the table R