Applications of Bloom Filter

Data Structures

Reading time: 30 minutes

Bloom Filter is a **probabilistic data structure** which is used to search an element within a large set of elements in constant time that is O(K) where K is the number of hash functions being used in Bloom Filter. This is useful in cases where:

- the data to be searched is large
- the memory available on the system is limited/ low

Hence, Bloom Filter is memory efficient than a Hash Map with the same performance. The only thing to note is that this is a probabilistic data structure so for a small number of cases, it may give wrong results (which can be limited).

The applications of Bloom Filter are:

- Weak password detection
- Internet Cache Protocol
- Safe browsing in Google Chrome
- Wallet synchronization in Bitcoin
- Hash based IP Traceback
- Cyber security like virus scanning

We will now understand each of the applications of Bloom Filter in depth.

Weak password detection

or an existing user updates the password.

Whenever, a new user comes, the new password is checked in the Bloom Filter and if a potential match is detected then the user is warned.

Note that passwords should be stored in hashed form to ensure that even if the Bloom Filter data becomes public, the password of the users are safe.

You can see this application as a string matching application which is memory efficient and works as fast as a Hash Map.

Internet Cache Protocol

A network system use Proxies which are computer systems through which all network requests are sent and received. This provides a central point of contact within a computer network system.

A Proxy hashes all URLs and keep them in its own cache. As there are multiple proxies in the network, all proxies share their cache record with each other. This is known as **Internet Cache Protocol**.

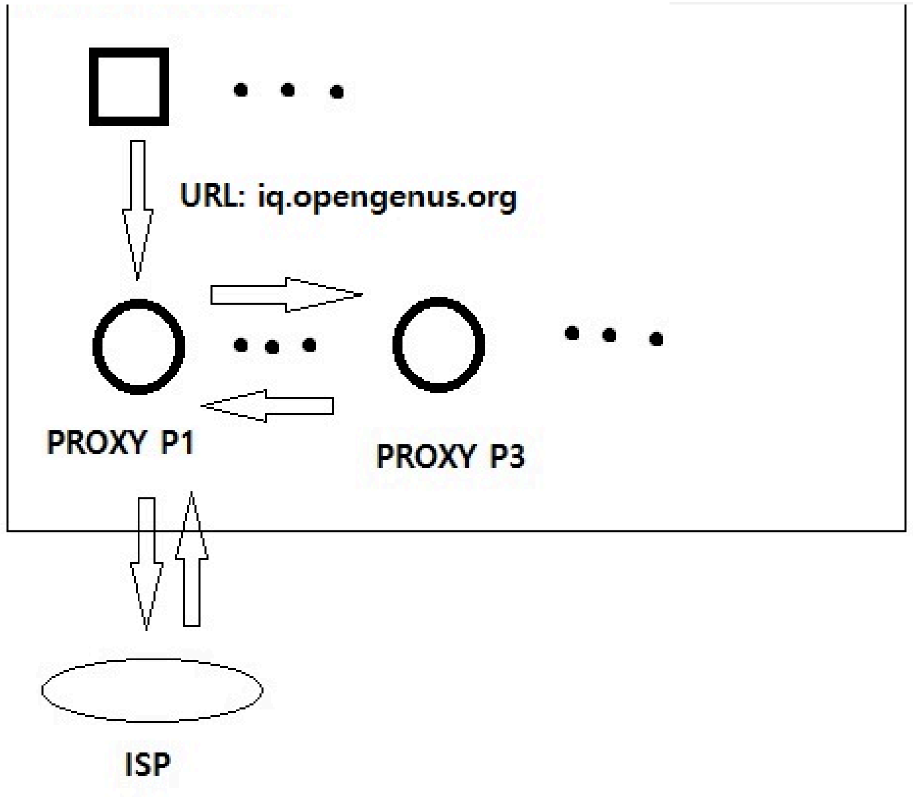
Consider a case when a computer A wants to go to "iq.opengenus.org".

The GET request will go to its nearest proxy P1. Now, proxy P1 will check its cache record and may find out that this page (iq.opengenus.org) has been cached by proxy P3. So, proxy P1 will send a request to proxy P3 and get the page and will return it to the computer A. The searching in the cache is done using Bloom Filter as you can imagine that internet use history of multiple users will be large.

As Bloom Filter is a probabilistic data structure, it may give wrong prediction like:

- Proxy P1 says URL is cached in P3 but it is not cached.

This will send a request to P3 but it will return that the page is not cached. Hence, the next request will be sent to ISP. This results in one extra request (P1 to P3).



One problem is that proxies delete pages after some time but Bloom Filter does not have an feature to delete entries from it. The solution is avoid this extra traffic due clearing of proxy's cache is to use Counting Bloom Filter.

Read this research paper: Summary Cache: A Scalable Wide Area Web Cache Sharing Protocol (PDF) by IEEE to understand better.

Safe browsing in Google Chrome

Google Chrome uses Bloom Filter to check if an URL is a threat or not. If Bloom Filter says that it is a threat, then it goes to another round of testing before alerting the user.

Learn more at this Google code review for this feature.

Wallet synchronization in

Bitcoin

Bitcoin is using Bloom Filter due to its efficient performance and it minimizes the risk of triggering DDoS attacks.

In Bitcoin, all information of blocks are distributed between nodes and as this data is large, this makes the system slow down. The problem is most of the data is thrown away after being received. So, Bloom Filter is used to detect if a particular information will be deleted later or not and following it, the decision to transfer data is made.

Learn more at Bitcoin documentation of BIP

Hash based IP Traceback

A challenge in designing IP protocols is to track the computer from which a packet has originated. This is difficult due to packet forwarding techniques such as NAT and encapsulation even in the case where there is no attempt to hide the source.

The solution is to use a Hash based approach to maintain audit traces which can be used to track the origin machine. As the amount of internet network is large, Bloom Filter is used for this.

Read this paper to understand more: Hash based IP Traceback by BBN Technologies

Cyber security like virus scanning

Common applications of Bloom Filter in the field of Cyber Security are:

- Virus scanning
- Worm detection
- DDoS prevention
- Risky URL detection

that searching can be done in parallel. If Bloom Filter predicts that there may be a match, then the particular sub-string is checked in a Hash Map to be sure.

The use of Bloom Filters in this makes the entire process quite fast and reliable at the same time.

With this, you must have a strong understanding of how Bloom Filters can be used in real life. Enjoy.

OpenGenus Tech Review Team

The official account of OpenGenus's Technical Review Team. This team review all technical articles and incorporates peer feedback. The team consist of experts in the leading domains of Computing.

Read More

Improved & Reviewed by:

SOFTWARE ENGINEERING

System Design —

Data Structures

Anagram Trees

Task Scheduler - Algorithm and Data Structure Project [in JavaScript]

Mean of Array problem solved with Treap

See all 229 posts →

YouTube Data API

We have gone through the Authorization process of YouTube Data API and understand the use of various tokens like client secret key, authorization token, refresh token and much more

SAATWIK BISARIA

DATA STRUCTURES

Bloom Filter: Better than Hash Map

Bloom Filter is a probabilistic Data Structure that is used to determine whether an element is present in a given list of elements. It is quite fast in element searching.

HARSH BARDHAN MISHRA