Unique Visitors Query System – UNIQUE

Minor Project

IBM Career Education

Disclaimer

This Software Requirements Specification document is a guideline. The document details all the high level requirements. The document also describes the broad scope of the project. While developing the solution if the developer has a valid point to add more details being within the scope specified then it can be accommodated after consultation with IBM designated Mentor.

INTRODUCTION

The purpose of this document is to define scope and requirements of a web analytics module to count unique website visitors between any two dates – UNIQUE for a leading media house. Currently the bank's web analytics module to query unique visitors was running very slow. It used a simple algorithm to count the unique visitors that was failing with growth in web site traffic touching around million visitors/month. In other words, the algorithm failed to perform in context of "Big Data".

The proposed system will provide a fast & responsive way to query the number of unique visitors to its website.

This document is the primary input to the development team to architect a solution for this project.

System Users

The management & editorial staff will primarily use the Unique Visitors Query System, UNIQUE.

Assumptions

1. The input to the system will be the standard web log generated by common web servers such as Apache.
2. The data is updated each day in batch mode on the next day.

REQUIREMENTS

UNIQUE will determine the count of unique visitors from the web log and insert a new row for that date along with the unique visitor's count in a database table. The old algorithm will be replaced by a state-of-the-art algorithm that leverages a probabilistic data structure to determine the count at fast speed so that the analytics server do not take exceptionally long time and high CPU %age for the task.

Basic System Operation

The system will leverage "Linear Counting" to speed up the unique visitor counting task. It is based on a probabilistic data structure.

Linear Counting is based on a memory efficient probabilistic data structure for counting unique elements in a "big data set".

The administrator uploads the web log in the system. Upon successful upload, UNIQUE builds the "bit mask" for each source IP (or domain name) in the web log. The unique visitor count is obtained from this bit mask by counting the number of 1's in it.

It is recommended to keep the load factor as 10 but it should be a configurable parameter along with the standard error of estimate, which will default to 0.01.

About Linear Counting

Linear Counting is based on a probabilistic data structure. It uses a bit mask of size "m". The value of "m" is computed using a formula described below. It uses a hash function to determine the "bit location" that is set to "1" in the bit mask.

$m > \max(5, 1/(\varepsilon t)^2) \cdot (e^t - t - 1)$	A practical formula that allow one to choose m by the standard error of the estimation. m - mask size ε - standard error of the estimation t - load factor, n/m

The ratio of number of distinct items in the data set to m is called as the "load factor" (= n/m). It is easy to visualize that with t < 1, number of collisions will be less and number of 1's (also called weight – "w") will be a good estimate of unique count compared to when t = 1 or t > 1; in case of very high load factor, the estimates will not be acceptable! A formula exists to estimate unique count from the number of 1's in the bit mask for t-values are slightly higher than 1:

Estimated count = –m x ln((m – w)/m)

The memory efficiency results from the bit mask usage. The memory consumption only grows linearly as a function of the expected cardinality (n).

DEVELOPMENT ENVIRONMENT

UNIQUE will be developed as a web application using Java/JSP and DB2 database. Eclipse will be used as the IDE for the same. You may consider using a JavaScript framework like Prototype. You may also refer to http://www.serve.net/buz/hash.adt/java.001.html URL for concepts on hashing and to http://dblab.kaist.ac.kr/Publication/pdf/ACM90_TODS_v15n2.pdf URL for deeper concepts on Linear Counting.