Fachhochschule Nordwestschweiz
Hochschule für Technik

Bloom-Filter

Diskrete Stochastik

Experte

Andreas Vogt

Team

Lukas Keller

Matthias Keller

Stefan Mettler

Idee des Bloom-filters

Der Bloom-filter dient zur Überprüfung, ob ein Wort in einem Datenstrom vorhanden ist.

Vorteile

Um zu überprüfen, ob ein Wort in einem Datenstrom vorhanden ist, muss nicht der gesamte Strom überprüft werden. Der Bloom-filter ist sehr Zeit/Speichereffizient.

Nachteile

Der Bloom-filter kann nicht mit vollständiger Wahrscheinlichkeit aussagen, ob ein Wort in einem Stream vorhanden ist. Es ist nur sicher, dass ein abgewiesenes Wort nicht enthalten ist.

Transformation des Hashwerts in einen Arrayindex

Der Hashwert wird von murmur3 in einen Long ausgelesen. Dieser Wert wird Modulo die Anzahl der Arrayelemente gerechnet. Der Absolutwert davon wird in einen Integer umgewandelt und als Arrayindex verwendet.

Verwendung des Bloom-Filters

Der Webbrowser Google Chrome verwendet einen Bloom-filter um «böse» URLs zu infizieren. Da die gesamte Liste aller «bösen» URLs sehr gross ist, wird sie nicht mit dem Browser zusammen verschickt, sondern liegt auf einem Server. Der Browser erhält nur einen Bloom-Filter aller URLs. Sobald der Filter bei einer URL, welche besucht werden will, «möglicherweise enthalten» zurückgibt, wird eine Anfrage an den Server geschickt, ob diese URL wirklich «böse» ist.

Test der Fehlerwahrscheinlichkeit

Wir validieren die Fehlerwahrscheinlichkeit folgendermassen. Die Hälfte eines Sets von Wörtern wird dem Filter beigebracht, die andere Hälfte nicht. Mit der zweiten Hälfte der Wörter wird überprüft, ob diese jeweils nicht im Filter vorhanden sind. Wir haben ein $p = 10\%$ gewählt und folgende Wahrscheinlichkeit für false positives erhalten:

3.569707401032702 %