Dist - Bloom Filter

Kevin Bieri, Till Dreier, Cristian Lluis Araya

20.11.2018

1 Idee

Ein Bloom-Filter ist eine Funktion mit der überprüft werden kann ob ein Element in einer Menge vorkommt. Es kann vorkommen, dass ein Element fälschlicherweise als Element der Menge angegeben wird obwohl es nicht vorkommt (False Positive). Andererseits werden Elemente aus der Menge zu 100% erkannt.

Der Filter verwendet ein Boolean Array der Grösse m und k Hashfunktionen. Mit der Hashfunktion wird dann die Entsprechende Position im Array auf true gesetzt.

2 Vorteile

- Platzsparend, da die Elemente selbst nicht abgespeichert werden.
- Effizient, da zum evaluieren eines Elements nicht durch eine Liste von Elementen iteriert werden muss, sondern direkt die Werte bei den berechneten Array Indizes betrachtet werden können.

3 Nachteile

- Die Wahrscheinlichkeit für False Positives kann zwar minimiert, jedoch nicht ganz eliminiert werden.
- Das Hinzufügen von Elementen in den Bloom-Filter ist durch die mehreren Hashfunktionen aufwändiger als das Hinzufügen in eine "normale" HashMap.
- Die erwartete Menge an Elementen muss zu Beginn bekannt sein, um die Grösse der Datenstruktur und die Anzahl Hashfunktionen bestimmen zu können.

4 Beispiel aus Praxis

Relationale Datenbanken wie z.B. Postgresql benützen einen Bloom filter um Festplattenzugriffe auf nicht extistierenden Einträgen zu reduzieren. Dieses Leistungsoptimierung erhöht die Geschwindigkeit bei Datenbankabfragen.

5 Testing

Um die Fehlerwahrscheinlichkeit unseres Bloom-Filters zu testen haben wir zuerst mit wenigen Wörtern unseren Filter getestet. Danach haben wir eine Testdatei mit rund 6'000 Wörtern erstellt und damit die Fehlerwahrscheinlichkeit überprüft.

6 Resultat

Unser Test mit 6025 Wörtern erkannte 123 False Positives was einer Fehlerwahrscheinlichkeit von 0,0219% entspricht.

Für den Test wurde ein Array in der Länge m = 556'988 und k = 7 verschiedenen Hashfunktionen verwendet. Als gewünschte Fehlerwahrscheinlichkeit wurde 0,01% angegeben.
Zur Berechnung von m und k haben wir die Formeln von Wikipedia verwendet.