Distinct counting with a self-learning bitmap

Aiyou Chen, Jin Cao, Larry Shepp, Tuan Nguyen

(Submitted on 8 Jul 2011)

Counting the number of distinct elements (cardinality) in a dataset is a fundamental problem in database management. In recent years, due to many of its modern applications, there has been significant interest to address the distinct counting problem in a data stream setting, where each incoming data can be seen only once and cannot be stored for long periods of time. Many probabilistic approaches based on either sampling or sketching have been proposed in the computer science literature, that only require limited computing and memory resources. However, the performances of these methods are not scale-invariant, in the sense that their relative root mean square estimation errors (RRMSE) depend on the unknown cardinalities. This is not desirable in many applications where cardinalities can be very dynamic or inhomogeneous and many cardinalities need to be estimated. In this paper, we develop a novel approach, called self-learning bitmap (S-bitmap) that is scale-invariant for cardinalities in a specified range. S-bitmap uses a binary vector whose entries are updated from 0 to 1 by an adaptive sampling process for inferring the unknown cardinality, where the sampling rates are reduced sequentially as more and more entries change from 0 to 1. We prove rigorously that the S-bitmap estimate is not only unbiased but scale-invariant. We demonstrate that to achieve a small RRMSE value of ϵ or less, our approach requires significantly less memory and consumes similar or less operations than state-of-the-art methods for many common practice cardinality scales. Both simulation and experimental studies are reported.

Comments:	Journal of the American Statistical Association (accepted)
Subjects:	**Computation (stat.CO)**; Data Structures and Algorithms (cs.DS)
ACM classes:	G.3; H.2
Cite as:	**arXiv:1107.1697 [stat.CO]**
	(or **arXiv:1107.1697v1 [stat.CO]** for this version)

Submission history

From: Aiyou Chen [view email]
[v1] Fri, 8 Jul 2011 18:50:16 GMT (426kb)

Which authors of this paper are endorsers?