Behavior of Keys in Random Databases

Oleg Seleznjev
Moscow State University, Faculty of Mathematics and Mechanics
119 899, Moscow, Russia
seleznev@top.rector.msu.su

Bernhard Thalheim
Cottbus Technical University, Computer Science Institute,
PB 101344, D-03013 Cottbus, Germany
thalheim@informatik.tu-cottbus.de

June 27, 1997

Abstract

We consider random databases formed by a sequence of random vectors (tuples) with common discrete distribution $\mathcal{P} = \{p(\mathbf{k})\}$. A key is a collection of columns (attributes) which identifies a tuple in a database. A minimal key does not contain any key subset. Many database theory and application problems (e.g., data search optimization, database design) are substantially defined by the *complexity* of a database, i.e., the size of key, functional dependency, and minimal key systems. Moreover, the average behavior of these characteristics is crucial for certain problems. Based on the Rényi entropy, $h_{\mathcal{P}} = -\log_2(\sum_{\mathbf{k}} p(\mathbf{k})^2)$, and the Poisson approximation (Stein-Chen) technique, we establish an approximation for the probability that a given set of attributes A is a key (minimal key) or there is a functional dependency between two sets of attributes as a part of the average complexity evaluation. Asymptotics for parameters (e.g., the entropy, the length) of most likely minimal key candidates are also derived. The length of the shortest keys in A have different interpretations in computer science (e.g., the height of digital trees for a certain model). We establish that this value can be approximated by the linear transformation of the Gumbel (double exponential) random variable with a given accuracy. Thus, the Monte-Carlo approach is feasible for an investigation. Several numerical examples and simulation results are presented.

Keywords: database, key, functional dependency, minimal key, relations and tables, Rényi entropy, Poisson approximation, height of digital tree

1 Introduction

Let $R = R(m, n)$ be a database (a sequence of m tuples, also referred to as *table*) and $U = \{1, \ldots, n\}$ be the set of all attributes. In other words, R is a matrix with m rows (tuples) and n columns (attributes) although the order of tuples in R is not essential for considered problems. Suppose all attributes take values in integer sets (*domains*), which may be infinite. Database theory mostly deals with *relations* (*sets* of tuples, i.e., the tuples in a relation take different values). Tables are also widely used in computer science and applications (e.g., in pattern recognition, in diagnosis of failures in digital circuits and systems, in geological forecasting see, e.g., Andreev (1984)). In many applications, the problem of tuple identification arises in a sense of minimization of possible costs.