Hashing of Data Bases Based on Indirect Observations of Hamming Distances *

V.B.Balakirsky
The author is with the Department of Information Theory,
University of Lund, P.O. Box 118, S-221 00 Lund, Sweden
on leave from The Data Security Association "Confident",
St.-Petersburg, Russia
E-mails:
volodya@dit.lth.se,
vbal@newton.spb.su,
vbal@mathematik.uni-bielefeld.de

Abstract. We describe hashing of data bases as a problem of information and coding theory. It is shown that the triangle inequality for the Hamming distances between binary vectors may essentially decrease the computational efforts of a search for a pattern in the data base. Introduction of the Lee distance in the space, which consists of the Hamming distances, leads to a new metric space where the triangle inequality can be effectively used.

Index. Hashing, Searching for Patterns, Coding, Decoding

*The work was supported by a Scholarship from the Swedish Institute, Stockholm, Sweden.

1 Introduction

One of important problems in computer science can be represented as follows : we have given a collection of items, and we wish to store these items and upon demand retrieve the items whose key-values match given key-values. A particular approach to the storage and retrieval problem is known as hashing when we use the key-value of an item to compute an address for the storage of the item. Since the mapping between keys and addresses is not one-to-one, the events when different keys have the same address may take place; these events are known as collisions, and their resolution is of the main interest for a hashing scheme [1-4].

Similar ideas form the basis for the procedures which are referred to as external hashing [5]. It is known that the cost per storage increases as the access time decreases. Main memories are usually of the random access time. Since their size is limited with the cost requirements, the data bases are stored in the secondary memory with a rather slow access [6]. The hashing technique can essentially reduce the total number of accesses required if the values of a hash function applied to the keys of each item of the data base are stored in the main memory. To find the item with a certain key we calculate the value of the hash function for that key and access only the items whose key-values have the same value of the hash function. In the present paper, we address the external hashing and formulate the following problem (to simplify formalization, the item and its key value are associated) : we have given a collection of items which are binary vectors of the same length stored on the external memory, and a part of random access memory (RAM) that can be filled with the values of a hash function corresponding to each item. For a given pattern, which is a binary vector of the same length as the items, we must find all the items located at the Hamming distance less than a fixed threshold value.

We develop the external hashing in two directions. First, the value of the threshold can be positive, and it puts some restrictions on the hash functions allowed since we want to get information about the Hamming distances between binary vectors based on the comparison of the values of the hash function corresponding to these vectors. To make it possible, we use the metric properties of the Hamming space and extend some approaches by V.Koshelev [7] to the problem of finding the neighbouring points in random digital fields (these results are briefly stated in Section 3). Second, we repre-