# THE GEOMETRY OF EFFICIENT HASHING ALGORITHMS

by

Azad Bolour

Memorandum No. ERL-M583

26 April 1976

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

# THE GEOMETRY OF EFFICIENT HASHING ALGORITHMS[†]

by

Azad Bolour

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

## Abstract

We consider the problem of finding balanced hashing algorithms for efficiently answering Boolean queries involving range specifications, from a relation (multi-attribute file), residing in a secondary storage device. A typical query under consideration has the form: retrieve all $\underline{x} = (x_1,\ldots,x_n)$ in the relation, satisfying $\bigvee_{i=1}^{k} \bigwedge_{j \in c_i (\subseteq \{1,\ldots,n\})} (x_j \in r_{ij})$, where for each $j$, the sets $r_{ij}$ are ranges of values of the $j^{th}$ attribute, $1 \leq j \leq n$. A probabilistic model is examined for the occurrence of such queries in which similar queries are assumed to be equiprobable. For this model it is shown that there is often a "box-like" hash function, for which the required average page access to answer a query is near-minimal, in the class of all balanced hash functions from the domain of the relation onto a number of pages in a secondary storage device.

[†]This work was supported in part by the Naval Electronic Systems Command contracts N00039-71-C-0255, N00039-75-C-0034, N00039-76-C-0022 and in part by the National Science Foundation grant GK-10656X.

## 1. Introduction

In this paper we analyze the average performance of hashing algorithms for answering queries from a relation or a multi-attribute file residing in a secondary storage device, and generalize in several directions, some of the pioneering work of R.L. Rivest [15] in this area. The basic result to be presented is that for a large class of queries, namely the class of Boolean queries involving range specifications, and under certain somewhat mild homogeneity restrictions on the probability of occurrence of such queries, hashing algorithms that are efficient, in the sense that they minimize the required average page access to answer a query, often have a simple "box-like" structure.

Hashing, of course, is a method of storing information, in which the location of an information item or a record $\underline{x}$, is determined by a function $h(\underline{x})$. Since its invention in the fifties, a very rich literature has dealt with various aspects of hashing when a record is characterized via a single attribute or key, chosen from a linearly orderable universe of keys. Excellent summaries of this literature appear in Knuth [12], and in a recent article by Knott [11].

Within the past decade, however, attention has been focused on the design of filing schemes in general, and of hash functions in particular, for efficient retrieval of records from multi-attribute files [1-4,8,10,13-18]. In the latter situation one identifies a record $\underline{x}$, with an n-tuple of values $(x_1,\ldots,x_n)$, chosen from an n-dimensional record space $K = K_1 \times \cdots \times K_n$. A retrieval request or a query may then give a partial description of a record through some of its attributes (components). In this paper we shall only be concerned with situations requiring total recall, so that the answer to such a query is the totality of records in a file satisfying the given partial description.

As an example consider a simplified Department of Motor Vehicles file F, in which there is a record for each registered automobile in California, and each record is characterized by the 3-tuple (last name, city of residence, make of car). A typical query about such a file may be,

retrieve all records in F satisfying:
last name = Smith and city of residence = Berkeley . (1)

A hash function designed to organize such a file in storage, should strive to restrict, as much as possible, the number of distinct storage locations (pages, if the file resides in a secondary storage device) that need to be examined in order to answer a query. It should also have a structure which makes it easy, given a query, to identify the storage locations relevant to its answer.

To these ends a number of authors, notably Terry A. Welch [17], Ronald L. Rivest [15], and James B. Rothnie and T. Lozano [16], have proposed using a hash function $h: K_1 \times \cdots \times K_n \to \{1,\ldots,N\}$, whose value is a one-to-one function (e.g. a concatenation) of the values of $n$ simple hash functions, $h_1,\ldots,h_n$, one for each attribute, that is $h(x_1,\ldots,x_n) = \bar{h}(h_1(x_1),\ldots,h_n(x_n))$, where $h_i: K_i \to \{1,\ldots,N_i\}$, $1 \leq i \leq n$, $N = N_1 \cdots N_n$, and $\bar{h}$ is one-to-one (see also Knuth [12] pp. 563-564). Such hash functions are known as multiple-key hash functions after Rothnie [16]. Their simple structure lends itself to easy identifiability of relevant storage locations for a wide class of queries. Furthermore, for answering partial-match queries from a file residing in a secondary storage device, Rivest [15] was able to prove that there is often a multiple-key hash function which minimizes the required average number of pages accessed to answer a query, among all balanced hash functions. Partial-match queries are queries similar to the query (1) above,

which can be represented via a conjunction of attribute-value equalities, and in computing the above average Rivest assumes that all such queries are equiprobable.

In this paper we explore more general conditions under which multiple-key hashing algorithms are optimal in the above sense.

In the first place, partial-match queries constitute but a small, albeit important, subset of queries commonly used to retrieve information from a relation, and the need arises to consider more general query types involving disjunctions and range specifications. In this paper we consider arbitrary "Boolean-range" queries representable as:

$$\text{retrieve all } \underline{x} \in F, \text{ satisfying } \bigvee_{i=1}^{k} \bigwedge_{j \in c_i (\subseteq \{1,\ldots,n\})} (x_j \in r_{ij}) , \quad (2)$$

where for each $j$ the sets $r_{ij}$ are ranges of values of the $j^{th}$ attribute.

Secondly, it would be more realistic to assume not that all queries are equiprobable but only that queries having a similar structure occur with equal probability. In this context we call two queries with defining expressions $\bigvee_{i=1}^{k} \bigwedge_{j \in c_i} (x_j \in r_{ij})$ and $\bigvee_{i=1}^{k'} \bigwedge_{j \in c'_i} (x_j \in r'_{ij})$, similar, if and only if $k = k'$, and there is a permutation $\sigma$: $\{1,\ldots,k\} \to \{1,\ldots,k\}$, such that $c'_i = c_{\sigma(i)}$, and $|r'_{ij}| = |r_{\sigma(i)j}|$, $j \in c'_i$, $1 \leq i \leq k$, where $|r|$ denotes the magnitude or the width of a range $r$. Notice that a partial-match query is a query of the form (2) in which $k = 1$, and $r_{1j}$, $j \in c_1$, consist of single points. Thus two partial-match queries would be similar in the above sense, if and only if they specify particular values for exactly the same set of attributes.

It turns out that even in the much more general setting described above, one can usually find a multiple-key hashing algorithm, for which the required average page access to answer a query is near-minimal in the class

of all balanced hash functions. This constitutes the principal result of the present paper. In fact we will prove that a near-optimal hash function $h^*$ in the above context often has a "box-like" structure, in that its clusters $h^{*-1}(i)$, $1 \leq i \leq N$, are identical rectangular boxes in $K$.

It may be noted here that this paper deals with continuous files, whose records are elements of a subset of the n-dimensional Euclidean space $R^n$, this being an approximation to a discrete situation in which the number of possible values of an attribute is reasonably large.

## 2. Basic Definitions and Assumptions

This section introduces a formal framework for the analysis of the average performance of hashing algorithms in answering "Boolean-range" queries. Much of the problem formulation is this section and the section to follow is based on the earlier works of Welch [17] and Rivest [15].

A record is an n-tuple of the form $(x_1,\ldots,x_n)$, $x_i \in K_i = [a_i,b_i] \subset R$, $1 \leq i \leq n$. The components of a record are also known as attributes, and the record space $K = K_1 \times \cdots \times K_n$ will also be referred to as the attribute-value space. In what follows we shall assume without loss of generality that $K = [0,1]^n$. A file or a relation is a finite set of records. We assume that the file under consideration has approximately constant size, and that the records in it are uniformly distributed within $K$.[†]

We consider a paged environment in which the pages in a secondary storage device provided for storing a file $F$ have equal size, and we let $b$ be

[†]This assumption can be replaced by the somewhat weaker assumption that the distribution of the records in $K$ is independent in the $n$ attributes. See Bolour [2], and section 5 of the present paper for more details.

the number of pages necessary to store $F$. A hash function into $b$ pages is a function $h: K \to \{1,\ldots,b\}$, with the interpretation that $h(\underline{x})$ is the page wherein a record $\underline{x}$ is to be stored, in case it belongs to the file under consideration. We use the shorthand $\underline{n}$ for the set $\{1,2,\ldots,n\}$ of positive integers less than or equal to an integer $n$. Given a hash function $h: K \to \underline{b}$, let $X_i = h^{-1}(i)$, $i \in \underline{b}$; $X_i$ is known as the $i^{th}$ cluster or the extent of the $i^{th}$ page, associated with $h$. In what follows we shall find it more convenient to view a hash function $h: K \to \underline{b}$, as a partition of $K$ into the $b$ clusters $X_1,\ldots,X_b$. Throughout the remainder of this paper we shall be concerned with measurable subsets of $K$ only, and in particular we assume that $X_1,\ldots,X_b$ have measurable characteristic functions. An explicit statement of this assumption in every instance in which it is needed will henceforth be omitted. We use the notation $|\ |$ to denote the size of a set, so that for a discrete set $S$, $|S|$ will denote its cardinality, whereas if $S$ is a subset of $R^n$, then $|S|$ will denote its volume (Lebesgue measure) in $R^n$. A hash function is said to be balanced if its associated clusters have equal size, $|X_i| = 1/b$, $i \in \underline{b}$. Since the records are assumed to be uniformly distributed, and the pages have equal size, we will restrict our attention to balanced hash functions only. This insures against consistent overflow/underflow, and we shall henceforth assume that page overflow is negligible.

A multiple-key hash function is a hash function that can be represented as $h(\underline{x}) = h(x_1,\ldots,x_n) = \bar{h}(h_1(x_1),\ldots,h_n(x_n))$, where $h_i: K_i \to \underline{N_i}$, and $\bar{h}$ is a one-to-one function. It is easy to see that the clusters associated with a multiple-key hash function have the form,

$$Y^1_{i_1} \times \cdots \times Y^n_{i_n}, \quad i_j \in \underline{N_j}, \quad j \in \underline{n}, \tag{3}$$

where

$$\bigcup_{i_j \in N_j} Y^j_{i_j} = K_j \ , \quad j \in \underline{n} \ .$$

A subset $B$ of $R^n$ is said to be a rectangular box or simply a box, if $B = r_1 \times \cdots \times r_n$, and if for $j \in \underline{n}$, $r_j$ is either a finite closed interval $[v_j, v_j+a_j]$, or the entire real line. We call $r_j$ the $j^{th}$ component of the box $B$. If the component clusters $Y^j_{i_j}$, $i_j \in N_j$, $j \in \underline{n}$, associated with a multiple-key hash function are intervals, then the clusters (3) are boxes in $K$, and we call the corresponding hash function box-like. Notice that the clusters associated with a balanced box-like hash function, must be isomorphic rectangular boxes in $K$.

Let $Q = Q(\underline{x})$ denote a Boolean expression involving $\wedge$, $\vee$, and atomic range statements of the form "$x_i \in [v_i, v_i']$", for example, $(0 \leq x_1 \leq \frac{1}{4} \wedge x_2 = \frac{1}{2}) \vee (\frac{1}{2} \leq x_3 \leq \frac{3}{4})$. A Boolean-range query is a query of the form, retrieve all $\underline{x} \in F$, satisfying $Q(\underline{x})$. (4)

In what follows since there will be no chance of confusion we will not distinguish between a Boolean-range query and the Boolean expression defining it. In particular the notation $Q$ (or $Q(\underline{x})$) will be used to denote the query (4), as well as its defining Boolean expression. Given a query $Q$, let $E(Q) = \{\underline{x} \in R^n \mid \underline{x}$ satisfies $Q(\underline{x})\}$; $E(Q)$ will be known as the extent of the query $Q$. Notice that the answer to $Q$ is simply $F \cap E(Q)$.

If the expression $Q(\underline{x})$ is in disjunctive normal form, then the corresponding query will be referred to as a canonical query. We note that an arbitrary Boolean-range query can easily be converted into an equivalent canonical query. This paper is concerned only with canonical queries, which can without loss of generality be represented as,

$$Q(\underline{x}) = \bigvee_{i=1}^{k} \bigwedge_{j \in c_i (\subseteq \underline{n})} x_j \in r_{ij} = [v_{ij}, v_{ij}+a_{ij}] \quad (-\infty < v_{ij} < \infty, \ 0 \leq a_{ij} < \infty, \ j \in c_i, \ 1 \leq i \leq k).$$

In case $Q(\underline{x})$ is composed of a single disjunct, its extent $E(Q) = E(\underset{j \in c}{\wedge} x_j \in r_j = [v_j, v_j + a_j])$ is a rectangular box in $R^n$. Q will then be referred to as a box query; for example $"0 \leq x_1 \leq \frac{1}{2} \wedge x_3 = \frac{1}{4}"$ (5) defines a box query whose extent is $[0, \frac{1}{2}] \times R \times [\frac{1}{4}, \frac{1}{4}] \times R^{n-3}$. If in addition each $r_j$, $j \in c$, consists of a single point, then the corresponding query will be called a partial-match query; for example $"x_1 = \frac{1}{2} \wedge x_3 = \frac{1}{4}."$ (6)

Given a canonical query $Q(x) = \underset{i=1}{\overset{k}{\vee}} \underset{j \in c_i}{\wedge} (x_j \in r_{ij})$, let $Q_i(\underline{x}) = \underset{j \in c_i}{\wedge} (x_j \in r_{ij})$, $1 \leq i \leq k$. The queries defined by $Q_i$, $1 \leq i \leq k$, may then be regarded as the components of the query Q. Since the order of these components is unimportant, it follows that a canonical query Q corresponds uniquely to a subset $\{Q_1, \ldots, Q_k\}$ of box queries, or equivalently to a collection $B(Q) = \{B_1, \ldots, B_k\} = \{E(Q_1), \ldots, E(Q_k)\}$ of boxes in $R^n$. Notice that $E(Q) = \underset{1}{\overset{k}{\cup}} E(Q_i)$, so that the answer to Q is just $\underset{1}{\overset{k}{\cup}} F \cap E(Q_i)$. In what follows we shall assume without loss of generality that $K \cap E(Q_i) \neq \emptyset$, $1 \leq i \leq k$.

Let $X_1$, $X_2$ be subsets of $R^n$. We say $X_1$ is equivalent to $X_2$ under translation (denoted by $X_1 \equiv X_2$) if there is a translation $T: R^n \to R^n$, $T(x_1, \ldots, x_n) = (x_1 + \alpha_1, \ldots, x_n + \alpha_n)$, $\alpha_i \in R$, $1 \leq i \leq n$, under which $X_1$ is mapped onto $X_2$. Two box queries having extents B and B' are said to be similar, if and only if $B \equiv B'$. Thus, the query (5) is similar to the query $"\frac{1}{4} \leq x_1 \leq \frac{3}{4} \wedge x_3 = \frac{1}{3}"$; but is not similar to the query (6).

It will be helpful in our later development to characterize similarity classes of box queries as follows. Let c denote a generic subset $\{i_1, \ldots, i_j\}$ of the set of attributes $\underline{n}$. If $B = r_1 \times \cdots \times r_n$ is a box whose finite components are $[v_{i_1}, v_{i_1} + a_{i_1}], \ldots, [v_{i_j}, v_{i_j} + a_{i_j}]$, then B may be characterized by $c = c(B) = \{i_1, \ldots, i_j\}$, an initial vector $\underline{v}_c = \underline{v}_c(B) = (v_{i_1}, \ldots, v_{i_j}) \in R^j$, and a dimension vector $\underline{a}_c = \underline{a}_c(B) = (a_{i_1}, \ldots, a_{i_j}) \in R^{+(j)}$

$= [0,\infty)^j$. Two boxes $B$ and $B'$ are then equivalent under translation if and only if $c(B) = c(B') = c$, and $\underline{a}_c(B) = \underline{a}_c(B')$. Thus a similarity class of box queries corresponds to a family of boxes in $R^n$ characterized by the set $c$ of coordinates along which they have finite components, and by a common dimension vector $\underline{a}_c \in R^{+(j)}$; it is parameterized by a set of vectors $\underline{v}_c \in R^j$. We will denote such a similarity class by $S_c(\underline{a}_c)$, and we let $S$ be the set of similarity classes of box queries.

Two canonical queries $Q$ and $Q'$, characterized by collections $\mathcal{B}(Q) = \{B_1,\ldots,B_k\}$, and $\mathcal{B}(Q') = \{B'_1,\ldots,B'_{k'}\}$ of boxes in $R^n$, are said to be similar if and only if $k = k'$, and there is a permutation $\sigma: \underline{k} \to \underline{k}$, such that $\mathcal{B}_i \equiv \mathcal{B}_{\sigma(i)}$, $1 \leq i \leq k$. Thus $Q$ and $Q'$ are similar if and only if they are composed of similar box queries. For example the query "$(0 \leq x_1 \leq \frac{1}{2} \wedge x_2 = \frac{1}{2}) \vee (\frac{1}{3} \leq x_3 \leq \frac{1}{2})$" is similar to the query "$(\frac{1}{4} \leq x_3 \leq \frac{5}{12}) \vee (\frac{1}{4} \leq x_1 \leq \frac{3}{4} \wedge x_2 = \frac{3}{4})$". A similarity class of canonical queries can then be characterized by an element of $S^*$, the set of finite subsets of $S$. Let $S^*$ be such a similarity class corresponding to the subset $\{S_{c_1}(\underline{a}_{c_1}),\ldots,S_{c_k}(\underline{a}_{c_k})\}$ of $S^*$. The extents $E(Q)$, $Q \in S^*$, then define a family of subsets of $R^n$ parameterized by a set of vectors $(\underline{v}_{c_1},\ldots,\underline{v}_{c_k})$ $\in R^{(\sum_1^k |c_i|)}$, where $\underline{v}_{c_i}$, $1 \leq i \leq k$, are the initial vectors of boxes associated with a given query in $S^*$. To avoid cumbersome notation, in what follows we shall again not distinguish between a similarity class $S^*$ of canonical queries, and the family of sets $E(Q)$, $Q \in S^*$.

Our assumption about the frequency of occurrence of queries is simply that <u>all similar queries are equiprobable</u>. The occurrence of queries is then governed by an arbitrary probability measure $P_{S^*}$ on the set $S^*$ of equivalence classes of queries, which induces a probability measure $P_Q$ on

the set $Q$ of canonical queries, since all similar queries are assumed to be equiprobable.

## 3. Problem Statement

Let a file $F$ be stored according to a particular balanced hash function $h$, characterized by clusters $X_1, X_2, \ldots, X_b$. In answering queries we assume that no auxiliary information is available about the present content of $F$. Since the answer to a query $Q$ is $E(Q) \cap F$, the retrieval algorithm must then access the $i^{th}$ page in answering $Q$, if and only if the $i^{th}$ page <u>may</u> contain a relevant record to $Q$, that is, if and only if $E(Q) \cap X_i \neq \emptyset$. So let $z_i(Q)$, $1 \leq i \leq b$, be random variables defined on the set $Q$ of queries as follows, $z_i(Q) = \begin{cases} 1 , & X_i \cap E(Q) \neq \emptyset \\ 0 , & X_i \cap E(Q) = \emptyset \end{cases}$. Then the number of pages that need to be examined in answering a query $Q$ is just, $z(Q) = \sum_{i=1}^{b} z_i(Q)$. The problem is then to choose a balanced hash function $h$, or equivalently a balanced partition of $K$ into clusters $X_1, \ldots, X_b$, minimizing the average number, $\bar{z}$, of pages that need to be examined in order to answer a query; where $\bar{z} = \sum_{i=1}^{b} \bar{z}_i(Q) = \sum_{i=1}^{b} P[E(Q) \cap X_i \neq \emptyset]$, and $P[E(Q) \cap X_i \neq \emptyset] = P_Q(\{Q \mid E(Q) \cap X_i \neq \emptyset\})$.

For answering partial-match queries this problem has been considered by Terry A. Welch [17] and later by Ronald L. Rivest [15], under the assumption that all partial-match queries are equiprobable.

Rivest's strategy in tackling the latter problem is quite simple, but proves to be extremely useful in simplifying it, and provides a basis for our approach to the more general problem at hand. His idea is simply this: the average page access is minimized if the average access to each individual page, i.e. $P[X_i \cap E(Q) \neq \emptyset]$ is minimized.

Formally the problem is to find a partition $X_1,\ldots,X_b$ of $K$ minimizing,

$$\bar{z} = \sum_{i=1}^{b} P[E(Q) \cap X_i \neq \emptyset] \ , \tag{7}$$

subject to the constraints, $|X_i| = \frac{1}{b}$, $1 \leq i \leq b$.

Now let

$$\bar{z}^*_{(b)} = \inf_{\substack{|X| = 1/b \\ X \subset K = [0,1]^n}} P[X \cap E(Q) \neq \emptyset] \ , \tag{8}$$

and let us call clusters achieving this infimum, optimal clusters. It then follows from (7) that

$$\bar{z} \geq b\bar{z}^*_{(b)} \ . \tag{9}$$

This lower bound on the average page access is achievable if and only if $K$ can be partitioned into $b$ optimal clusters.

The principal result of this paper (proved in the next section) is that if similar queries are equiprobable, then optimal clusters can without loss of generality be taken to be equivalent rectangular boxes, whose dimensions are determined by the problem parameter $P_{S^*}$. Notice that a collection of identical rectangular boxes is about the simplest collection of geometrical objects to fit together. Thus if it is possible to partition $K$ into $b$ optimal boxes, then the induced hash function is optimal: it achieves minimal average page access. Otherwise a partition of $K$ into approximately optimal boxes is usually possible and induces a box-like hash function from $K$ into (approximately) $b$ pages, achieving near-minimal average page access.

## 4. Characterization of Optimal Clusters

The main result of this paper is embodied in the following theorem, whose proof is the object of the present section.

Theorem 1. If similar canonical queries are equiprobable then

$$\bar{z}^*_{(b)} = \inf_{\substack{X \subset [0,1]^n \\ |X| = 1/b}} P[X \cap E(Q) \neq \emptyset] \qquad (10)$$

is achieved by a rectangular box in $[0,1]^n$.

To prove Theorem 1 we first need an expression for $P[X \cap E(Q) \neq \emptyset]$ in terms of the given probability measure $P_{S^*}(S^*)$ on similarity classes of queries.

Let $F = \{E_{\underline{v}} \mid \underline{v} \in V \subset R^m\}$ be a family of subsets of $R^n$, parameterized by a subset of points in $R^m$, and let $X \subset R^n$. The intersection measure

of $X$ with respect to $F$ (denoted by $|X|_F$) is then defined as $|X|_F = |\{\underline{v}|\ E_{\underline{v}} \cap X \neq \emptyset\}|$, and measures the size of the subset of $F$ whose elements intersect $X$. Identifying $F$ with a similarity class $S^*$ of queries, we see that since all queries in $S^*$ are assumed equiprobable,

$$P[X \cap E(Q) \neq \emptyset |\ Q \in S^*] = |X|_{S^*}/|K|_{S^*}$$

and hence that

$$P[X \cap E(Q) \neq \emptyset] = \int_{S^*} |X|_{S^*} \frac{dP_{S^*}(S^*)}{|K|_{S^*}} . \qquad (11)$$

We are now in a position to prove a stronger assertion than that of Theorem 1.

Theorem 2. Given a set $X \subset [0,1]^n$ of volume $1/b$, there exists a box $B(X) \subset [0,1]^n$ of the same volume $1/b$, whose intersection measure with respect to every similarity class $S^*$ of canonical queries is no larger than the corresponding measure of $X$; that is,

$$|B(X)|_{S^*} \leq |X|_{S^*} , \quad S^* \in S^* .$$

To see that Theorem 1 follows from Theorem 2, suppose the latter is valid. Then from (11) it follows that the infimum (10) is achieved if and only if it is achieved by a rectangular box. What then needs proof is that among all boxes of a given size $1/b$ within $[0,1]^n$, there is one for which $\overline{z}_b^*$ is achieved. However, a box $B$ having dimension vector $\underline{a}(B) = (a_1,\ldots,a_n)$, satisfies the constraints $|B| = 1/b$, $B \subset [0,1]^n$, if and only if,

$$\prod_{i=1}^{n} a_i = \frac{1}{b} \quad \text{and} \quad \frac{1}{b} \leq a_i \leq 1 , \ 1 \leq i \leq n . \qquad (12)$$

It is then easy to see from (11) that $P[B \cap E(Q) \neq \emptyset]$ is a continuous function of the dimension vector $\underline{a}(B)$ which from (12) spans a compact region of $R^n$, and therefore, that $\inf_{\substack{B \subset [0,1]^n \\ |B| = 1/b}} P[B \cap E(Q) \neq \emptyset]$ is achievable. Q.E.D.

We first prove Theorem 2 for similarity classes of partial-match queries. The result is then used to prove the more general statement about canonical queries. To begin with, however, we need to develop a few elementary concepts.

Since the dimension vector associated with a partial-match query is $(0,0,\ldots,0)$, similarity classes of partial-match queries can be characterized by a subset $c$ of the coordinates. If $S_c$ is such a similarity class, then the expression defining a typical query in $S_c$ has the form, $\underline{x}_c = (x_{i_1},\ldots,x_{i_j}) = (v_{i_1},\ldots,v_{i_j}) = \underline{v}_c$. Letting $\pi_c(X)$ denote the projection of a set $X$ $(\subseteq R^n)$ on the subset $c$ of the coordinates,

$$\pi_c(X) = \{\underline{v}_c = (v_{i_1},\ldots,v_{i_j}) \mid \exists \underline{x} = (x_1,\ldots,x_n) \in X,\ x_{i_1} = v_{i_1} \wedge \cdots \wedge x_{i_j} = v_{i_j}\},$$

we see that the extent of a query $Q_{\underline{v}_c} : \underline{x}_c = \underline{v}_c$ intersects $X$, if and only if $\underline{v}_c \in \pi_c(X)$. Hence $|X|_{E_c} = |\{\underline{v}_c \mid E(Q_{\underline{v}_c}) \cap X \neq \emptyset\}| = |\pi_c(X)|$. The appropriate definition of $|\pi_\emptyset(X)|$ here is $|\pi_\emptyset(X)| = \begin{cases} 1, & X \neq \emptyset \\ 0, & X = \emptyset \end{cases}$.

In our later proofs, as in most inductive proofs in n-dimensional geometry, we will need the notion of a section of a set in $R^n$. Let $t \in R$. The n-1 dimensional hyperplane in $R^n$ defined by the equation $X_i = t$ will be denoted by $H_i(t)$. Given a set $X \subseteq R^n$, we let $X_i(t) = \pi_{\underline{n}-\{i\}}(X \cap H_i(t))$; $X_i(t)$ will be known as the section of $X$, perpendicular to the $i^{th}$ coordinate axis, defined by $t$.

The following assertions relate the projected volume of a set $X \subset R^n$ to the projected volumes of its sections perpendicular to a given coordinate axis.

Assertion 1. Let $c \subseteq \underline{n}$, and suppose $i \in c$. Let $c' = c - \{i\}$. Then

$$|\pi_c(X)| = \int_R |\pi_{c'}(X_i(t))| dt \quad .$$

Assertion 2. Let $c \subseteq n$, and suppose now that $i \notin c$. Then

$$|\pi_c(X)| \geq |\pi_c(X_i(t))| \quad , \quad t \in R \ .$$

The proofs are trivial and will be omitted.

Theorem 3 (A Minimal Projection Property of Rectangular Sets in $R^n$) Given a set $X \subset R^n$ of size $|X|$, there exists a box $B(X) \subset R^n$ of the same size (as $X$), whose projected volume on every hyperplane defined by equating a subset of the coordinates to zero, is no larger than the corresponding projected volume of $X$; that is,

$$|B(X)| = |X| \quad ,$$

but

$$|\pi_c(B(X))| \leq |\pi_c(X)| \quad , \text{ for all } \ c \subset \underline{n} \ . \tag{13}$$

(In particular the dimensions of $B(X)$ are no larger than the corresponding dimensions of the smallest box enclosing $X$, so that if $X \subset [0,1]^n$, then we can assume without loss of generality that $B(X) \subset [0,1]^n$.)

Proof. Given a set $c$ of coordinates, let $\underline{c}$ be an $n$-vector of zeroes and ones defined as, $\underline{c}(i) = 1$ if and only if $i \in c$. For example if $n = 4$ and $c = \{1,4\}$, then $\underline{c}$ would be $(1,0,0,1)$. We let $\underline{1}_n = (1,\ldots,1)$ be the vector of $n$ ones. The following lemma turns out to be equivalent to Theorem 3.

Lemma 1. Let $c_1,\ldots,c_k$ be not necessarily distinct subsets of $\underline{n}$, and let $\lambda_1,\ldots,\lambda_k$ be positive reals such that $\sum_{i=1}^{k} \lambda_i \underline{c}_i = \underline{1}_n$. Then

$$\prod_{1}^{k} |\pi_{c_i}(X)|^{\lambda_i} \geq |X| \quad . \tag{14}$$

Since the inequality (14) is tight when $X$ is a box, Lemma 1 is seen to be an easy consequence of Theorem 3.

We will now prove the sufficiency of Lemma 1 for Theorem 3. The case $|X| = 0$ is trivial. Assume then that $|X| > 0$. It is easily justified that for any subset $c \subset \underline{n}$, and its complement $\bar{c} = \underline{n} - c$, $|\pi_c(X)||\pi_{\bar{c}}(X)| \geq |X|$. It follows that $|\pi_c(X)| > 0$ for all $c \subset \underline{n}$.

Let us now consider the problem of maximizing the volume of a rectangular box $B$ of dimensions $y_1,\ldots,y_n$, satisfying the constraints (13):

$$\begin{aligned} &\text{maximize } \prod_{1}^{n} y_i \text{ such that} \\ &|\pi_c(B)| = \prod_{i \in c} y_i \leq |\pi_c(X)| \ , \quad c \subset \underline{n} \quad . \end{aligned} \tag{15}$$

Letting $\epsilon = \min_{c \subset \underline{n}} |\pi_c(X)|^{1/|c|}$ we see that $\epsilon > 0$ and $(\epsilon,\epsilon,\ldots,\epsilon)$ is feasible for the above problem. Hence no box having a zero dimension can be optimal for (15). We can therefore use the standard transformations $x_i = \log y_i$, $1 \leq i \leq n$, reducing the problem (15) to the linear programming problem,

$$\begin{aligned} &\text{maximize } x_0 = \sum_{1}^{n} x_i = (1,1,\ldots,1)\cdot\underline{x} = \underline{1}_n \cdot \underline{x} \ , \\ &\text{such that for all } c \subset \underline{n}, \ \sum_{i \in c} x_i = \underline{c}\cdot\underline{x} \leq \log |\pi_c(X)| \ , \\ &\underline{x} \text{ unconstrained} \ . \end{aligned} \tag{16}$$

The dual linear program to (16) is,

$$\text{minimize} \quad \sum_{c \subset \underline{n}} \lambda_c \log |\pi_c(X)| \ ,$$

$$\text{such that} \quad \sum_{c \subset \underline{n}} \lambda_c \underline{c} = \underline{1}_n \tag{17}$$

$$\text{and} \quad \lambda_c \geq 0 \ , \quad c \subset \underline{n} \ .$$

Now both the primal and the dual systems above are feasible: $\underline{x} = (\log e, \ldots, \log e)$ is feasible for (16), while $\lambda_{\{1\}} = \cdots = \lambda_{\{n\}} = 1$, $\lambda_c = 0$ otherwise, is feasible for (17). Hence by the Duality Theorem of Linear Programming [6, section 6.3] the optimal solution to (16) may be written as,

$$x_0^* = \sum_{c \subset \underline{n}} \lambda_c \log |\pi_c(X)| \ ,$$

where

$$\sum_{c \subset \underline{n}} \lambda_c \underline{c} = \underline{1}_n \ , \quad \lambda_c \geq 0 \ , \tag{18}$$

Thus the maximum volume of a box satisfying the constraints (15) is $\prod_{c \subset \underline{n}} |\pi_c(X)|^{\lambda_c}$, with the $\lambda_c$'s satisfying (18), and is therefore no less than $|X|$ if Lemma 1 is valid. Theorem 3 would then follow by shrinking the maximizing box to the correct size $|X|$.

Proof of Lemma 1. We proceed by induction on $n$. For $n = 1$, we need to prove that $|\pi_{\{1\}}(X)|^1 |\pi_{\emptyset}(X)|^{\lambda_{\emptyset}} \geq |X|$, which is trivially true since $\pi_{\{1\}}(X) = X$ and $|\pi_{\emptyset}(X)| = 1$. Suppose now that the lemma is true for $n = m-1$. Let $X \subset R^m$, and let $c_1, \ldots, c_k \subset \underline{m}$, $\lambda_1, \ldots, \lambda_k > 0$, and $\sum_1^k \lambda_i \underline{c}_i = \underline{1}_m$. Without loss of generality let

$$m \in c_i \quad \text{if and only if} \quad 1 \leq i \leq j \leq k \ , \tag{19}$$

and let $c_i' = c_i - \{m\}$ for $1 \leq i \leq k$. Then $c_i'$, $1 \leq i \leq k$, satisfy

$$\sum_1^k \lambda_i \underline{c}_i' = \underline{1}_{m-1} \ , \tag{20}$$

and we are in a position to use the induction hypothesis on sections of $X$ perpendicular to the $m^{th}$ axis, obtaining

$$\prod_{1}^{k} |\pi_{c_i'}(X_m(t))|^{\lambda_i} \geq |X_m(t)| , \quad t \in R .$$

Hence using Assertion 1,

$$\begin{aligned}
|X| &= \int_R |X_m(t)|dt \leq \int_R \prod_{1}^{k} |\pi_{c_i'}(X_m(t))|^{\lambda_i} dt \\
&= \int_R \prod_{i=1}^{j} |\pi_{c_i'}(X_m(t))|^{\lambda_i} \prod_{j<i\leq k} |\pi_{c_i'}(X_m(t))|^{\lambda_i} dt \\
&\leq \int_R \prod_{i=1}^{j} |\pi_{c_i'}(X_{m+1}(t))|^{\lambda_i} \prod_{j<i\leq k} |\pi_{c_i}(X)|^{\lambda_i} dt \\
&\quad \text{(by Assertion 2 since for } j < i \leq k, \ c_i = c_i' \text{ and } m \notin c_i) \\
&= \prod_{j<i\leq k} |\pi_{c_i}(X)|^{\lambda_i} \int_R \prod_{i=1}^{j} |\pi_{c_i'}(X_m(t))|^{\lambda_i} dt \\
&\leq \prod_{j<i\leq k} |\pi_{c_i}(X)|^{\lambda_i} \prod_{i=1}^{j} \left[\int_R |\pi_{c_i'}(X_m(t))|dt\right]^{\lambda_i} \\
&\quad \text{(by Hölder's inequality [9, par. 188], since} \\
&\quad \text{from (18) and (19), } \sum_{i=1}^{j} \lambda_i = 1) \\
&= \prod_{i=1}^{k} |\pi_{c_i}(X)|^{\lambda_i} \\
&\quad \text{(by Assertion 1, since for } 1 \leq i \leq j, \ c_i = c_i' \cup \{m\}).
\end{aligned}$$

Q.E.D.

This proves Theorem 2 for similarity classes of partial-match queries.

We are now going to show that the box $B(X)$ associated with a set $X$ in Theorem 3 has no larger intersection measure than $X$ with respect to any equivalence class of boxes in $R^n$; that is,

$$|X|_{S_c(\underline{a}_c)} \geq |B(X)|_{S_c(\underline{a}_c)} , \quad S_c(\underline{a}_c) \in S .$$

The proof of this claim is an immediate consequence of the following lemma and Theorem 3.

Lemma 2. Let $S_c(\underline{a}_c) = S_c(a_{i_1},\ldots,a_{i_j})$ be an equivalence class of boxes in $R^n$. Then

$$|X|_{S_c(\underline{a}_c)} \geq \sum_{c' \subseteq c} \left( \prod_{i \in c'} a_i \right) |\pi_{c-c'}(X)| \quad . \tag{21}$$

Theorem 2 for box queries would then follow from Lemma 2 and Theorem 3, by noting that the inequality (21) is tight when $X$ is a finite box so that,

$$\begin{aligned} |X|_{S_c(\underline{a}_c)} &\geq \sum_{c' \subseteq c} \left( \prod_{i \in c'} a_i \right) |\pi_{c-c'}(X)| \\ &\geq \sum_{c' \subseteq c'} \left( \prod_{i \in c'} a_i \right) |\pi_{c-c'}(B(X))| \quad \text{(using Theorem 3)} \\ &= |B(X)|_{S_c(\underline{a}_c)} \cdot \end{aligned}$$

Proof of Lemma 2. Here we will only set up the machinery for an inductive proof of Lemma 2. The actual proof will be omitted as it is similar to one used by Davenport [7] in connection with a different problem.

Let $S_{\underline{n}}(\underline{a}) = S_{\underline{n}}(a_1,\ldots,a_n)$ be an equivalence class of finite boxes in $R^n$. (The case $S_c(\underline{a}_c)$, $c \subset \underline{n}$, can be treated in a similar fashion.) Let us denote by $B(\underline{v};\underline{a})$ a box in $S_{\underline{n}}(\underline{a})$, with initial vector $\underline{v}$, and let $\hat{\mu}_i$ be the unit vector in the $i^{th}$ direction, $1 \leq i \leq n$. Given a set $X \subset R^n$ define the set $M_i(X) \subset R^n$ as, $M_i(X) = \bigcup_{0 \leq b_i \leq a_i} X - b_i\hat{\mu}_i$: $M_i(X)$ is the subset of $R^n$ swept out by a translation of $X$ through a distance of $a_i$ units parallel to the $i^{th}$ coordinate axis, in the negative direction.

Claim.

$$\{\underline{v} \mid B(\underline{v};\underline{a}) \cap X \neq \emptyset\} = M_n M_{n-1} \cdots M_1(X) = \bigcup_{0 \leq b_1 \leq a_1, \ldots, 0 \leq b_n \leq a_n} (X - \sum_{i=1}^{n} b_i \hat{\mu}_i)$$

(and hence $|X|_{S_{\underline{n}}(\underline{a})} = |M_n \cdots M_1(X)|$).

Proof.
$$\begin{aligned} B(\underline{v};\underline{a}) \cap X \neq \emptyset &\Leftrightarrow \exists \underline{x} \in X, \ \underline{x} \in B(\underline{v};\underline{a}) \\ &\Leftrightarrow \exists \underline{x} \in X, \ \underline{v} \in B(\underline{x}-\underline{a};\underline{a}) \\ &\Leftrightarrow \exists \underline{x} \in X, \ \underline{v} = \underline{x} - \sum_{i=1}^{n} b_i \hat{\mu}_i, \ 0 \leq b_i \leq a_i, \ i \in \underline{n} \\ &\Leftrightarrow \underline{v} \in M_n \cdots M_1(X) \ . \end{aligned}$$ Q.E.D.

The assertion $|M_n \cdots M_1(X)| \geq \sum_{c \subseteq \underline{n}} (\prod_{i \in c} a_i) |\pi_{\underline{c}}(X)|$ can now be proved by induction on $n$, using the induction hypothesis on sections of $M_{n-1} \cdots M_1(X)$ perpendicular to the $n^{th}$ axis. The details are similar to the proof in Davenport [7], and the interested reader should have little difficulty carrying out the latter proof in the context of the present problem.

Proof of Theorem 3. Let $S^* = \{S_{c_1}(\underline{a}_{c_1}), \ldots, S_{c_k}(\underline{a}_{c_k})\}$ be a similarity class of Boolean-range queries. It is then easily justified that

$$|X|_{S^*} = \prod_{i=1}^{k} |K|_{S_{c_i}(\underline{a}_{c_i})} - \prod_{i=1}^{k} (|K|_{S_{c_i}(\underline{a}_{c_i})} - |X|_{S_{c_i}(\underline{a}_{c_i})}) \ . \qquad (22)$$

This can be proved by a straightforward manipulation of the definitions of intersection measure and similarity. However it is perhaps most easily seen by dividing both sides of (22) by $\prod_{i=1}^{k} |K|_{S_{c_i}(\underline{a}_{c_i})}$, so as to convert (22) into a statement about probabilities. It then asserts that the probability that the extent of a query in $S^*$ does not intersect $X$, is the product of the probabilities that the extent of a query in $S_{c_i}(\underline{a}_{c_i})$ does

not intersect $X$, $1 \leq i \leq k$. Hence $|X|_{S^*}$ is an increasing function of $(|X|_{S_{c_1}(\underline{a}_{c_1})},\ldots,|X|_{S_{c_k}(\underline{a}_{c_k})})$, and since by Lemma 2,

$$(|X|_{S_{c_1}(\underline{a}_{c_1})},\ldots,|X|_{S_{c_k}(\underline{a}_{c_k})}) \geq (|B(X)|_{S_{c_1}(\underline{a}_{c_1})},\ldots,|B(X)|_{S_{c_k}(\underline{a}_{c_k})})$$

it follows that

$$|X|_{S^*} \geq |B(X)|_{S^*} . \qquad \text{Q.E.D.}$$

## 5. Conclusions

We have thus far been able to reduce the problem of finding optimal clusters among all subsets of $K$ with a given volume, to that of finding optimal rectangular clusters. Given $P_{S^*}$, the probability of accessing a rectangular cluster with dimension vector $\underline{a} = (a_1,\ldots,a_n)$, is a polynomial function of $\underline{a}$, say $q(\underline{a})$. Hence the dimensions of an optimal cluster can be found by solving the following constrained minimization problem in $R^n$,

$$\begin{aligned} &\text{minimize } q(\underline{a}) = q(a_1,\ldots,a_n) \\ &\text{such that } 0 \leq a_i \leq 1, \quad 1 \leq i \leq n, \\ &\text{and } \prod_{1}^{n} a_i = \frac{1}{b} . \end{aligned} \tag{23}$$

Let $(a_1^*,\ldots,a_n^*)$ be an optimal solution to (23). It follows from (9) that if $[0,1]^n$ can be partitioned into boxes of the above dimensions then the induced hashing algorithm is optimal. However, this is possible only if each dimension $a_i^*$ of an optimal box is an integral divisor of 1, which is of course extremely unlikely. One may therefore attempt to find boxes of approximately the same shape and size as the optimal box, whose dimensions are in fact integral divisors of 1. Of course to avoid page overflow the size of the

approximating boxes should be no larger than $1/b$. Our preliminary computational results indicate that roundoff errors incurred in the latter process are generally quite small. There are, of course, two sources of errors here. In the first place, the average page access after roundoff may be larger than the minimum possible using $b$ pages. On the other hand since the boxes may have a smaller volume than $1/b$, the number of clusters could be greater than $b$ and more pages may be required in memory for storing the file.

We computed the optimal dimensions of a cluster, i.e. the optimal solution of (23), in answering box queries for more than 100 cases in 3 and 4 dimensions with $b = 1000$, and $b = 10{,}000$, in which the coefficients of $q(\underline{a})$ were chosen randomly from the integers 1 to 100. In carrying out the roundoff process we found that no more than 10% extra storage space was needed in any of the trials. In fact in 70% of the cases considered the extra storage space required was less than 5%. Considering the fact that it is desirable to leave some slack in each page to avoid excessive overflow, the extra storage cost incurred seems quite tolerable.

More importantly, we found that the average page access after roundoff was quite close to the derived lower bound (9). In more than 90% of the cases considered the extra retrieval time due to roundoff was less than 5%, and in the remainder it was less than 10%.

Although these preliminary results deal with box queries only, we see no reason to believe that the above errors should be any larger for general canonical queries.

Thus one can safely assert that in case similar queries are equiprobable, there is often a box-like hash function for which the required average page access to answer a query is near minimal in the class of all balanced hash functions from the attribute-value space $K$ onto approximately $b$ pages

in a secondary storage device.

From a practical point of view perhaps the most serious shortcoming of the above model is the assumption of uniform distribution of records in K. Under a suitable extension of the notion of similarity, however, the results presented easily generalize to a situation in which the distribution of the records in K is independent in the n attributes, i.e. where the occurrence of records in K is governed by a probability density $p(\underline{x}) = p_1(x_1)p_2(x_2)\cdots p_n(x_n)$. A balanced hash function here is one whose associated clusters have equal probability with respect to $p(\underline{x})$, and the problem of finding an optimal balanced hash function can be reduced to one in which $p(\underline{x})$ is uniform (and therefore the clusters have equal size), by a suitable transformation of the coordinates (see Bolour [2] for more details).

However, in most applications one finds examples of both dependent and independent attributes in a given file. Since multiple-key hash functions treat the attributes independently of one another, in general no such hash function can be balanced for a file with dependent attributes in the sense that it distributes the records in a typical file evenly among the given pages. It seems, therefore, that for files in which there are strong dependencies, hash functions that are efficient both in terms of storage utilization and retrieval time must necessarily have a more complex structure. The simplest way perhaps to extend the results of this paper to such files, is to partition the attribute-value space <u>a priori</u> into a small number of boxes in each of which the assumptions of this paper are approximately justified, and to hash each box according to an appropriate optimal box-like hash function.

Acknowledgments

It is a privilege to be able to thank the many people whose help and encouragement have made this work an enjoyable and an instructive experience. In particular, I would like to express my thanks to Lotfi A. Zadeh for his patient guidance, support, and encouragement throughout my stay at Berkeley, to Richard M. Karp for his careful attention and criticism during many hours of discussion leading up to the present work, to S.S. Chern for his heartening conviction that I should be able to prove Theorem 3, to Istvan Fáry for giving me access to his collected papers on related problems in n-dimensional geometry, to Ronald L. Rivest and Terry A. Welch whose earlier work in this area provides a basis for the present paper, and to Colin McMaster for reading parts of the manuscript.

## References

1. Abraham, C.T., S.P. Ghosh and D.K. Ray-Chaudhuri, File Organization Schemes Based on Finite Geometries, Information and Control 12 (February 1968), 143-163.
2. Bolour, Azad, On the Optimality of Rectangular Hashing Schemes for Answering Basic Queries from a Relation with Independent Attributes, Memorandum No. ERL-M535, August 1975, Electronics Research Laboratory, University of California, Berkeley.
3. Bose, R.C. and Gary G. Koch, The Design of Combinatorial Information Retrieval Systems for Files with Multiple Valued Attributes, SIAM Journal of Applied Mathematics 17 (November 1969), 1203-1214.
4. Chow, D.K., New Balanced File Organization Schemes, Information and Control 15 (1969), 377-396.
5. Codd, E.F., A Relational Model of Data for Large Shared Data Banks, CACM 13.6 (June 1970), 377-387.

6. Dantzig, G.B., Linear Programming and Extensions, Princeton University Press (1963).

7. Davenport, H., On a Principle of Lipschitz, The Journal of the London Mathematical Society 26 (1951).

8. Gustafson, R.A., Elements of the Randomized Combinatorial File Structure, Proceedings of the Symposium on Information Storage and Retrieval, ACM SIGIR, University of Maryland (April 1971), 163-174.

9. Hardy, G.H., J.E. Littlewood and G. Polya, Inequalities, Cambridge University Press (1934).

10. Hsiao, David and Frank Harary, A Formal System of Information Retrieval from Files, CACM 13 (February 1970), 67-73.

11. Knott, G.D., Hashing Functions, The Computer Journal 18,3 (1975).

12. Knuth, Donald E., The Art of Computer Programming, Vol. 3, Sorting and Searching, Addison Wesley (1972).

13. Lefkovitz, David, File Structures for On-Line Systems, Spartan Books (1969).

14. Lum, Vincent Y., Multi Attribute Retrieval with Combined Indices, CACM 13.11 (November 1970).

15. Rivest, Ronald L., Analysis of Associative Retrieval Algorithms, Rapport de Recherche No. 54 (February 1974), Laboratoire de Recherche en Informatique et Automatique, Rocquencourt, France.

16. Rothnie, J.B. Jr. and T. Lozano, Attribute Based File Organization in a Paged Memory Environment, CACM 17.2 (February 1974), 63-69.

17. Welch, Terry, Bounds on Information Retrieval Efficiency in Static File Structures, Technical Report MAC TR-88, Project MAC MIT, Cambridge, Mass. 02139.

18. Wong, E. and T.C. Chiang, Canonical Structure in Attribute Based File Organization, CACM 14 (September 1971), 593-597.