REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

J. C. MOORE
W. B. RICHMOND
A. B. WHINSTON

# An optimal search strategy based on user preferences for choice oriented databases

*Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, tome 32, n° 1 (1998), p. 1-32.

<http://www.numdam.org/item?id=RO_1998__32_1_1_0>

Recherche opérationnelle/Operations Research
(vol. 32, n° 1, 1998, pp. 1 à 32)

# AN OPTIMAL SEARCH STRATEGY BASED ON USER PREFERENCES FOR CHOICE ORIENTED DATABASES (*)

by J. C. MOORE (1), W. B. RICHMOND (2) and A. B. WHINSTON (3)

Communicated by Charles TAPIERO

---

Abstract. – *In many instances, when people search a database, they are searching for a best record instead of a record with a particular set of characteristics. Using boolean queries to develop a search strategy to find a best record has proven too difficult, especially for novice users. We construct an optimal decision strategy for finding the best available alternative that takes into account both the expected value of continued search and the cost of search. The optimal strategy is intuitively appealing, and can be used as the basis for heuristics in more complex environments.*


Keywords: Data base, best record, optimal strategy.

Résumé. – *Dans beaucoup de cas, lorsque des personnes effectuent des recherches dans une base de données, elles recherchent en fait un meilleur enregistrement au lieu d'un enregistrement ayant un ensemble particulier de caractéristiques. L'utilisation d'interrogations booléennes pour développer une stratégie de recherche en vue de trouver un meilleur enregistrement s'est montrée trop difficile, spécialement pour les utilisateurs novices. Nous construisons une stratégie de décision optimale pour trouver la meilleure possibilité disponible qui prenne en compte à la fois la valeur moyenne de la continuation de la recherche et le coût de celle-ci. La stratégie optimale est intuitivement attractive, et peut être utilisée comme base d'heuristiques dans des situations plus complexes.* 

Mots clés : Base de données, meilleur enregistrement, stratégie optimale.

## 1. INTRODUCTION

With traditional database systems people define a query in terms of the characteristics of the records they want to retrieve. The database management system searches for records that match the desired characteristics exactly. In many cases, however, people want the record representing the best available alternative and not a record with a specific set of characteristics. These cases include choosing products and services in electronic shopping systems

---

(*) Received June 1995, accepted September 1995.
(1) Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana.
(2) School of Business Administration George Mason University, Fairfax, Virginia.
(3) Management Science & Information Systems Department and $IC^2$ Institute University of Texas at Austin, Austin, Texas.

and selecting members for a project team from a personnel database. This scenario is described more fully in Moore, Richmond and Whinston [4].

Searching choice oriented systems requires understanding the user's preferences and taking into account the cost of search. In many instances, these databases will be very large and the user will be charged based on the length of the search. The system must therefore make a tradeoff between the expected gain from additional search and the expected cost from that search. In reponse to the need for systems that search for a best record and incorporate the cost of search in the search process, Moore, Richmond and Whinston [5] take an economic, decision theoretic approach to information retrieval. They look at selecting the best record from a file. The file is treated as a choice set, and the individual decision makers have preferences over the records in the file. The individuals who want to access the file must all use the same decision process or algorithm. The algorithm balances the cost of computing against the value of further search. Under certain restrictions on the searcher's and the individuals' utility functions, they are able to construct an optimal algorithm for finding the best record in the file. Moore *et al.*'s approach is limited by the restriction of the user's preferences to two attributes.

In this paper, we extend the work of Moore, Richmond and Whinston [5] by developing an optimal search algorithm for a database with $n$ attributes. We integrate the decision problem with the database system by viewing the retrieval process (algorithm) as a formalization of part of the database user's decision strategy. Taking this decision-theoretic approach enables us to incorporate the user's preferences and values when evaluating and selecting a retrieval algorithm. In our approach, the net value of retrieving a particular record depends on the user's information and the cost to the user (in dollars or mental effort) of the retrieval process. Different retrieval algorithms may, therefore, retrieve different records in response to the same query. We use a decision model that incorporates the cost of the decision process in determining an optimal decision, and we model decision making as a two-stage process–gathering information and selecting an alternative.

We draw upon an economic model [3] commonly used in marketing [1], and we model the users' preferences over the records with an additive, linear function of the attributes. The users know the available attributes, but they do not know which alternatives have which attributes. When the users' believe that each alternative is equally likely to have each attribute, and that the existence of any attribute is independent of the existence of any other attribute, the optimal search strategy is to examine the attributes in decreasing order of their importance to the user. This search strategy is optimal because it maximizes the user's expected utility minus his or her expected search costs. Examining an attribute is logically equivalent to using an inverted file to determine which alternatives (records) have a particular attribute. The process of examining attributes ends when the expected cost of obtaining more information outweighs the value of obtaining the information.

The optimal process described above depends on the individual. Each user will have his or her own retrieval algorithm that is based on individual preferences. The structure of the algorithm is common across individuals, however, and can be implemented in a database management system. In addition, the algorithm does not require a common, system-wide similarity measure as is used in some systems, and obtaining a functional representation of the user's preferences can be accomplished using self-explicated weights or various forms of conjoint analysis [2].

This paper is divided into four sections. In Section 2, we present the decision model that is the basis for this work. In Section 3 we construct an optimal decision strategy, and in Section 4, we conclude with a discussion of extensions of this work.

## 2. MODEL

We conceptualize the choice problem by using a decision model that is implemented within the context of a database system. We model the set of alternatives as a database and the retrieval algorithm as a decision process. When faced with a choice problem, such as which car to buy or which set of employees to assign to a project, the database user wants to select the best available record from the database. We assume that there are $m$ records in the database, each represented by $n$ attributes, and the decision maker's payoff is a function of the attributes of the record selected. This view of the choice problem is consistent with the economic model proposed by Lancaster [3], consumer choice models used in marketing [1] and the psychological models proposed by Tversky [8, 9], thus merging the fields of economics, marketing, psychology and database retrieval.

When faced with a choice problem, a database user usually needs to gather information about the possible alternatives. Database systems provide a natural place to store information on the alternatives, and the query process naturally embodies the search strategy. When gathering information, the database management system should make a tradeoff between the value or accuracy of the decision and the cost of acquiring and processing the information, including the cost of making the choice. In general, the user does not want to pay for complete information on the alternatives before evaluating and selecting the final choice. There is a tradeoff between the cost of obtaining more information and the expected improvement in the choice that may result from the additional information. As an example, envision a user searching a database to find a used car. The user's ideal choice is a 1964 red, convertible, E-type Jaguar that costs nothing, and he or she has found a 1964, red, convertible, E-type Jaguar that costs one dollar. It is unlikely that getting more information on the other available cars will result in a better choice.

To explore the process of searching a database for a best record, we use the Moore and Whinston [6, 7] choice model, which explicitly accounts for the tradeoff between the cost and the value of additional information. In the model a choice, **D**, is a function of eight parameters.

$$\mathbf{D} = \langle X, \phi, D, \omega, A, \{M_a | a \in A\}, c, r \rangle$$

where:

$X$ is the state space. It represents the set of possible database instantiations. $x \in X$ is one database instantiation.

$\phi$ is a probability distribution over the set of possible states of the world, $X \cdot \phi(x)$ represents the probability that $x$ is the true state (*i.e.*, that $x$ is the actual database instantiation in use). We use the probability distribution over the set of possible database instantiations to construct a retrieval algorithm that maximizes the net expected payoff for searching an arbitrary database instantation.

$D$ is the set of possible final decisions or choices. Here, it is the set of alternatives represented by the records in the database. The user will choose one record, $d$. The record chosen depends on the database management system's response to the user query (*i.e.*, it depends on the search process the database management system uses).

$\omega$ is a function of the true state of the world (the database instantiation), $x$, and the record(s), $d$, the database management system returns in response to the query. Each individual user has his or her own $\omega(x, d)$, which represents the user's preferences over the alternatives in the database.

$A$ is the set of information gathering actions. It represents the available computations and comparisons. Each $a \in A$ results in a signal or response. In this case, each action, $a \in A$, is similar to an inverted file lookup that identifies the records having a particular attribute, and the signal is the set of records having the particular attribute. Note that we are concerned with the logical search strategy. Implementing each action need not be done with an inverted file lookup.

$\{M_a\}$ is the partition of the state space resulting from action $a$. $\{M_a\} = \{M_{a,1}, \ldots, M_{a,\eta(a)}\}$ where $\eta(a)$ is the number of possible signals resulting from action $a$. In this case, $\eta(a)$ is the number of possible responses from an inverted file lookup. Each $M_{a,i}$ represents the set of possible database instantiations whose records have values for attribute $a$ corresponding to those identified as the $i$-th possible response to action $a$.

$c$ is a function of the information actions, where $c(a)$ is the cost to the user of executing action $a$. In this context, $c$ is the cost of an (logical) inverted file lookup. If the user is charged for using the system, then $c$ is a dollar cost; otherwise, $c$ is the dollar cost of the user's time and the mental cost of using the system.

$r$ is the maximum number of attributes the database management system can examine to solve the query.

The goal is to determine an information and decision strategy (retrieval algorithm), $\sigma$, that maximizes $\Omega(\sigma) - \Gamma(\sigma)$, where:

$\sigma$ defines the sequence of logical inverted file accesses used to answer the query. More formally, $\sigma$ is a sequence of actions that partition the state space. At each element of a partition a (possibly different) action is specified. For each element of the final partition, $\sigma$ determines the records the database management system will display.

$\Omega$ is the value of the strategy, $\sigma$, to the user. It is the sum over all possible database instantiations of the probability of a particular instantiation times the value to the user of the record(s) retrieved for the query under strategy $\sigma$. $\Omega$ is the gross expected value of the strategy $\sigma$.

$\Gamma$ is the gross expected cost of the strategy $\sigma$. It is the sum over all possible database instantiations of the probability of a particular instantiation multiplied by the cost of retrieving the record(s) to answer the query under strategy $\sigma$.

The following definition introduces the majority of the terminology used throughout this paper. Figure 1 highlights the key terms.

DEFINITION 2.1: *A decision strategy* $\sigma$ *is represented by a sequence of* $r + 1$ *pairs.* $\sigma = \langle(\mathbf{B}_1, \alpha_1), \ldots, (\mathbf{B}_{r+1}, \delta)\rangle$ *where:*

$\mathbf{B}_t$ *is a partition of the state space and is termed an information structure* ($\mathbf{B}_t = \{B_i\} = \{B_1, B_2, ..., B_k\}$) *and is defined by the logical inverted file look ups performed. Each* $B_i$ *is a set of states.*

$\mathbf{B}_1$ *is the entire state space* (i.e., $\mathbf{B}_1 = X$).

$\alpha_t$ *defines the action taken at the t-th step of the search strategy on each element of the partition* $\mathbf{B}_t$. $\alpha_t : \mathbf{B}_t \rightarrow A$. *For example, let* $\mathbf{B}_t = \{B_1, B_2, B_3\}$ *and let the function* $\alpha_t(\mathbf{B}_t)$ *be given by* $\alpha_t(B_1) = 0$, $\alpha_t(B_2) = 0$, $\alpha_t(B_3) = 3$. *The* $\alpha_t(\mathbf{B}_t)$ *corresponds to taking no information-acquisition action at step t if the true state is an element of* $B_1$ *or* $B_2$, *and taking experiment 3 if the true state is an element of* $B_3$.

$\iota(a, B)$ *is a set of sets of states. It is the partition resulting from taking experiment a on partition element B.* $\iota(a, B) = \{M_a\} \cap B$.

$\delta$ *defines the final decision made after the information gathering process.* $\delta : \mathbf{B}_{r+1} \rightarrow D$. *It defines the record(s) presented to the user in response to an query.*

$\Omega(\sigma) = \sum_{B \in \mathbf{B}_{r+1}} \sum_{x \in B} \phi(x)\, \omega(x, d)$, *where d is the decision taken when the information-gathering strategy ends in partition element B.*

$\Gamma = \sum_{B \in \mathbf{B}_{r+1}} \pi(B)\, C(B)$, *where* $C(B)$ *is the cost of executing the experiments that result in partition element B, and* $\pi(B)$ *is the probability of the true state being in partition element B. This is the cost the user expects to pay to query the database.*

In the database formulation of the choice problem, the elements of the state space, X, are database instantations represented by $m \times n$ matrices, Z, where there are $m$ records or alternatives, each represented by a row of Z; and there are $n$ attributes, represented by the columns of Z, which the alternatives may posses. The user's preferences over the records are represented by $\omega(x, d)$, where $d$ is one of the $m$ rows and $\omega$ is a function of the $n$ attributes. Since there is a cost to obtaining the information, the user may not want the database management system to examine all the attributes, so the user's preference for an alternative must be estimated. We want to find an optimal retrieval algorithm, given an individual's preference function, $\omega$, and assuming that the final choice maximizes expected utility.

## 3. OPTIMAL SEARCH STRATEGY

In this section, we look at the optimal strategy for selecting a record from a database. We assume that the database user's preferences over the

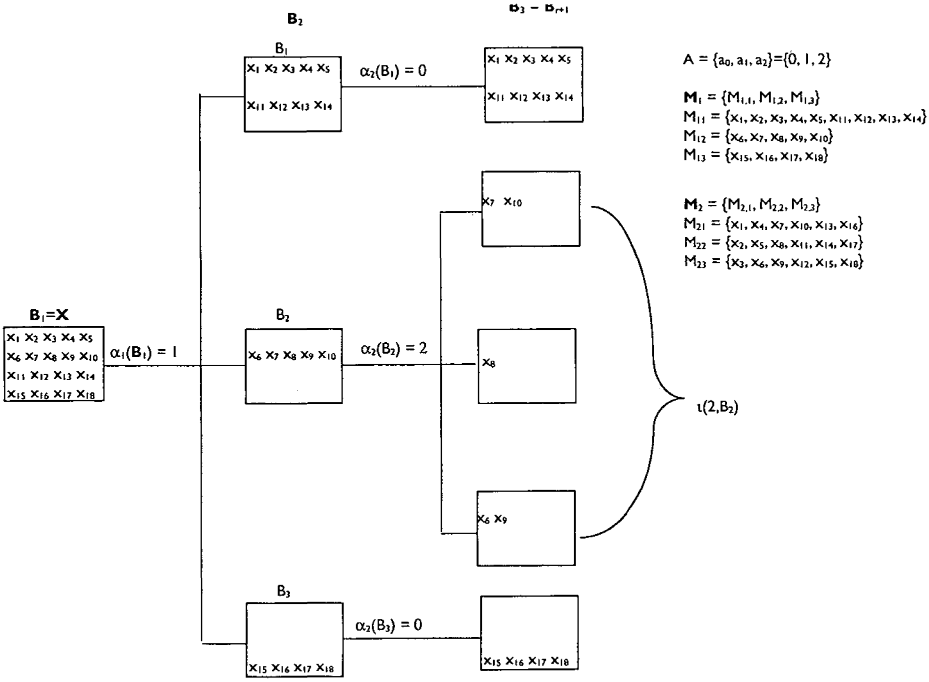



**Figure 1. – Decision Strategy.**

records are represented as an additive, linear (utility) function of the records' attributes. The user enters this function as the basis for his or her query. Each record either has or does not have a particular attribute, and the existence of an attribute is independent of the existence of any other attribute. The database management system obtains the information needed to respond to the query by testing a column (attribute) of the database. The result of a column experiment is a list of the records that have the tested attribute. This type of strategy is similar to using inverted files to search a database.

Figure 2 represents *a* strategy for a database with three records and six attributes. When the user submits a query, the database management system has no information to determine a response. This is represented by the empty box at the left of Figure 2. To solve the query, the database management system examines attribute *A*. The possible results from examing attribute *A* are shown in the column of boxes labeled 1. The actual result depends on the database instantiation.

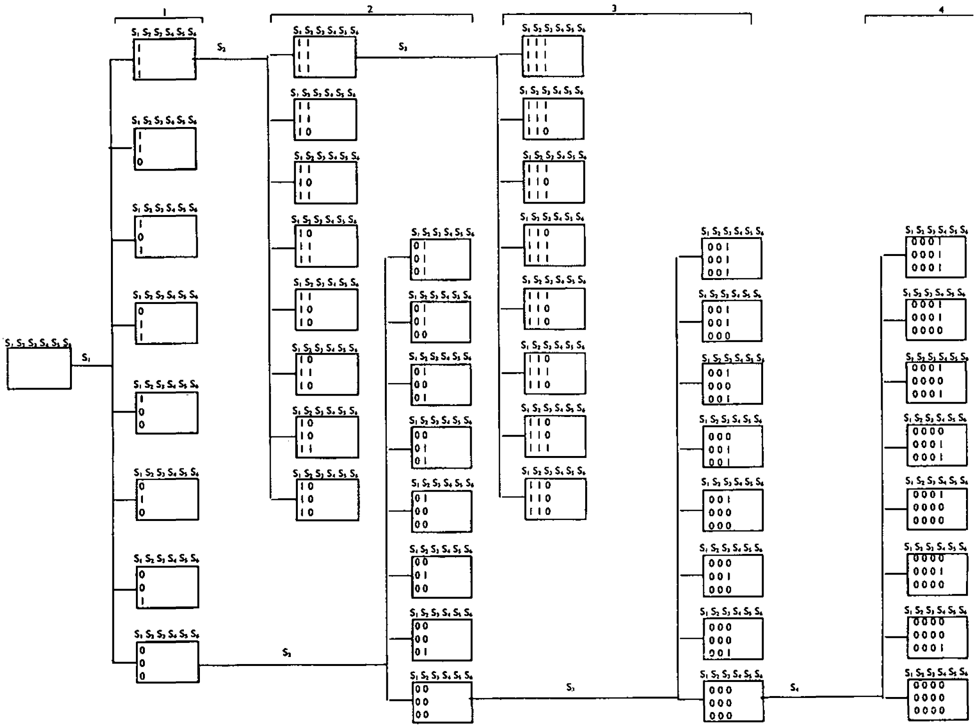


**Figure 2. – Strategy** $\sigma$.

If in the database all records have attribute $S_1$ (represented by

| $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|
| 1 | | | | | |
| 1 | | | | | |
| 1 | | | | | |

), or none of the records have attribute $S_1$ (represented by

| $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|
| 0 | | | | | |
| 0 | | | | | |
| 0 | | | | | |

), the database management system will gather more information before responding to the query; otherwise, it will respond to the query after examining attribute $S_1$.

When all of the records or none of the records have attribute $S_1$, the next attribute examined is attribute $S_2$. The results from examining attribute $S_2$ are shown in the two columns labeled 2. If all records have both attribute $S_1$ and attribute $S_2$, or if none of the records have either attribute $S_1$ or attribute $S_2$, then before answering the query, the database management system gathers more information. In both cases, it examines attribute $S_3$. The results of examining attribute $S_3$ are shown in the two columns labeled 3. Finally, if none of the records have any of the attributes $S_1$, $S_2$ or $S_3$, then the database management system examines a final attribute, attribute $S_4$. Note that even if none of the records have any of these attributes, the database management system will not gather any more information before responding to the query.

The strategy depicted in Figure 2 is not necessarily an optimal strategy. We will refer to the figure throughout this section to explain the lemmas, propositions, theorems and their proofs.

Formally,

$$\mathbf{D} = \langle X,\ \phi,\ D,\ \omega,\ A,\ \{M_a | a \in A\},\ c,\ r\rangle$$

where:

• $X = \{Z\}$, the set of possible database instantiations ($m \times n$ matrices). For each $z \in \{Z\} : x_i$ is the $i$-th record (row of $z$). $x_{.j}$ is the $j$-th attribute (column of $z$), and $x_{ij} \in \{0, 1\}$ for $i = 1, ..., m$ and $j = 1, ..., n$. If $x_{ij} = 1$, the $i$-th record has attribute $j$. If $x_{ij} = 0$, the $i$-th record does not have attribute $j$.

• $\phi(z) = 1/2^{m \cdot n}$. We also assume that $\phi(z|B) = 1/|B|$ where $B$ is any subset of $X$ and $|B|$ is the number of database instantiations in $B$, and all database users believe this probability distribution. Note that we are assuming the probability of $x_{ij} = 1$ is 1/2 for all $i$ and $j$.

• $D = \{x_{1\cdot}, ..., x_{m\cdot}\}$ where $x_{i\cdot}$ is the $i$-th record.

• $\omega(x, d) = \sum_{j=1}^{n} u_j\, x_{dj}$ where $u_j$ is the value or weight placed on attribute $j$. Each user will have an individual payoff function, which will differ in the weights, $u_j$, placed on the attributes.

• $A = \{0, 1, ..., n\}$. Experiment $j \in \{1, ..., n\}$ returns the values for attribute $j$ for all records. Experiment 0 implies do nothing–gather no more information. We also will make use of $A^* = \{1, ..., n\}$, the set of non-null experiments.

• $\{M_a | a \in A\} = \{M_{a1}, ..., M_{a2^m}\}$. Each $M_{aj}$ corresponds to the set of database instantiations having one of the $2^m$ possible vector values for attribute $a$.

• $c$ is constant for all experiments.

• $r \geq n$, the maximum number of inverted file lookups, is greater than or equal to the number of attributes. This ensures that we are able to examine all attributes, if it is economical to do so.

The optimal strategy turns out to be both simple and obvious. However, proving that this strategy is optimal is difficult. Additionally, the strategy is optimal only under the limited conditions specified above. At each step the database management system checks the attribute with the highest utility value for the particular user or stops. The system stops examining columns (attributes) when the expected value of continuing the search exceeds the expected cost. This stopping rule is not myopic. It is possible for it to be suboptimal to check the next attribute, but valuable to check the next 2 (or 3 or 4 or...) attributes. We use induction to prove that the optimal decision strategy examines the attributes in decreasing order by their weight.

• We first show that the last attribute examined on any branch of the optimal strategy must be the attribute with the highest remaining value (*i.e.*, of all the attributes not yet examined in processing the query, the database management system uses the attribute the user values the most to select the record to complete the query).

• We then show that if the second to the last attribute examined, $k$, has a lower weight than the last attribute examined, $k'$, then following a strategy, $\sigma'$, that examines attribute $k$ when $\sigma$ examines $k'$ and examines attribute $k'$ when $\sigma$ examines $k$, does not lower the user's expected payoff.

• We complete the proof by showing that if the $(k+1)$-st through last ($r$-th) attributes used in the solving the query are examined in decreasing order of their weight, $u_j$, and at each level, $k+1, \ldots, r$, either the no attribute is examined and the query is completed, or the database management system examines attribute $a^i$ for $i \in \{k+1, \ldots, r\}$. Then if the value of attribute $k$ is less than the value of attribute $k+1$ (*i.e.*, $u_k < u_{k+1}$), then following a strategy, $\sigma'$, that examines attribute $k$ when $\sigma$ examines $k+1$ and examines attribute $k+1$ when $\sigma$ examines $k$, does not lower the user's expected payoff.

### 3.1. Optimal search strategy proof

To begin, define the set of possible outcomes from examining an attribute as **O**.

DEFINITION 3.1:

$$\mathbf{O} = \{b = (b_1, \ldots, b_m) | b_i \in \{0, 1\}, \ i = 1, \ldots, m\}$$

We can use **O** to refer to any partition element, *B*, arrived at by a strategy, $\sigma$.

$$B \equiv \langle x_{\cdot j} \rangle_{j \in J^*} \quad \text{for} \quad \langle x_{\cdot j} \rangle \in \mathbf{O}^t, \text{ and } t = \# J^*,$$

where:

• $x_{\cdot j}$ represents the $j$-th attribute vector (*i.e.*, it identifies the records that have attribute $j$);

• $J^* = J(B) \subseteq A^* \equiv \{1, \ldots, n\}$ is the set of integers corresponding to the attributes that have been examined in the process of arriving at the information set *B*, so $\langle x_{\cdot j} \rangle_{j \in J^*}$ is the set of known attribute vectors.

• $\# J^*$ is the number of attributes examined in the process of arriving at the information set *B*.

By defining the complement of $J^*$, we can easily express the expected value of a record from any given information set, *B*.

Definition 3.2:

$$K^* = A^* \setminus J^*.$$

We will also need a way to refer to the set of records that the database management system would return in response to the query at any point in the query process. We will call this set $D^*(B)$.

Definition 3.3:

$$D^*(B) = \{x_i \in D | \sum_{j \in J^*} x_{ij} \cdot u_j \geq \sum_{j \in J^*} x_{kj} \cdot u_j \, \forall x_{k \cdot} \in D\}$$

*and denote the known value of record i as* $u_i^*$,

$$u_i^* = \sum_{j \in J^*} u_j \, x_{ij}.$$

To improve the comprehension of the following arguments, renumber the records such that for every database instantiation in partition element *B*,

$$u_1^* \geq u_2^* \geq \ldots \geq u_m^*. \tag{1}$$

### 3.2. The last attribute

The first step is to show that in the optimal strategy, the last attribute examined must have the highest weight of any available attribute. In Figure 2, this applied to the attribute examined on the partition element

| $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | | | |
| 0 | 0 | 0 | | | |
| 0 | 0 | 0 | | | |

and the attribute examined on the partition element

| $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ |
|---|---|---|---|---|---|
| 1 | 1 | | | | |
| 1 | 1 | | | | |
| 1 | 1 | | | | |

. In both cases, the experiment is the last experiment taken on a branch of the strategy. According to the claim, if the strategy in Figure 2 is optimal, then the user values attribute $S_4$ more than attributes $S_5$, and $S_6$, and the user values attribute $S_3$ more than attributes $S_4$, $S_5$ and $S_6$.

DEFINITION 3.4: *A strategy, $\sigma$, weakly dominates a strategy, $\sigma'$, if* $\Omega(\sigma) - \Gamma(\sigma) \geq \Omega(\sigma') - \Gamma(\sigma')$.

LEMMA 3.1: *Let $k$ designate the last attribute examined taken on a branch of a strategy, $\sigma$, so $\alpha_t(B) = k$ is such that $\iota(B, k) = \{B_1, \ldots, B_{2^m}\} \in \mathbf{B}_{r+1}$. If there is an attribute $k'$ such that $k' \notin J(B)$ and $u_{k'} > u_k$ then there is a strategy $\sigma'$ that weakly dominates the strategy $\sigma$.*

*Proof of Lemma 3.1:* Recall that the records in $B$ are ordered in descending order by their expected value (*i.e.*, $u_1^* \geq u_2^* \geq \ldots \geq u_n^*$). Define $\sigma'$ to be the same strategy as $\sigma$ except that $\alpha_t(B) = k'$ in $\sigma'$. Then:

$\Gamma(\sigma') = \Gamma(\sigma)$, and

$$
\begin{aligned}
&\Omega(\sigma') - \Omega(\sigma) \\
&\quad = \sum_{B \in \mathbf{B}_{r+1}^{\sigma'}} \sum_{x \in B} \phi(x)\,\omega(x, d) - \sum_{B \in \mathbf{B}_{r+1}^{\sigma}} \sum_{x \in B} \phi(x)\,\omega(x, d) \\
&\quad = \sum_{B' \in \iota(B, k')} \sum_{x \in B} \phi(x)\,\omega(x, d) - \sum_{B' \in \iota(B, k)} \sum_{x \in B} \phi(x)\,\omega(x, d) \\
&\quad = \frac{1}{2^m} \sum_{i \in \mathbf{O}} \left\{ \max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2}\, u_k \right\} \\
&\qquad - \frac{1}{2^m} \sum_{i \in \mathbf{O}} \left\{ \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2}\, u_{k'} \right\} \\
&\quad = \frac{1}{2^m} \sum_{i \in \mathbf{O}} \{\max_h (u_h^* + u_{k'}\, x_{hi}) - \max_h (u_h^* + u_k\, x_{hi})\} + \frac{1}{2}\, u_k - \frac{1}{2}\, u_{k'}
\end{aligned}
$$

$$= \frac{1}{2^m} \sum_{i \in \mathbf{O} | x_{1i}=0} \{\max_h (u_h^* + u_{k'} x_{hi}) - \max_h (u_h^* + u_k x_{hi})\}$$
$$+ \frac{1}{2} [u_{k'} - u_k + u_k - u_{k'}]$$
$$\geq 0 \quad \square$$

### 3.3. Second to last attribute

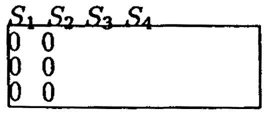
We want to show that if $\sigma$ is a strategy such that:

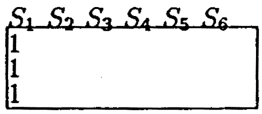
• the next to last attribute examined on some branch of $\sigma$ is $k$,

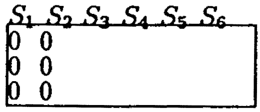
• on some partition elements following from $k$ the database management system examines attribute $k'$,

• no attribute is examined after $k'$, and that

• the value of attribute $k'$ is greater than the value of attribute $k$,

then $\sigma$ is weakly dominated by a strategy $\sigma'$ that examines attribute $k$ instead of $k'$ and attribute $k'$ instead of $k$. In Figure 2, this applies to the pair of attributes examined after partition element 

, and the pair of attributes examined after the partition element 


In Figure 2, when $B^*$ corresponds to 

, then $k$ corresponds to attribute $S_3$ and $k'$ corresponds to attribute $S_4$. The claim states that if the user places more weight on attribute $S_4$ than on attribute $S_3$, then the strategy depicted in Figure 2 is weakly dominated.

LEMMA 3.2: *Let $\sigma$ be a strategy such that for some $B^* \in \mathbf{B}_q$ we have:*

1. $\alpha_q(B^*) = k$
2. $\iota(B^*, k) = \{B_1, ..., B_{2^m}\}$
3. $\alpha_{q+1}(B_i) \in \{0, k'\}$, $i = 1, ..., 2^m$
4. $\iota(B_i, k') \in \mathbf{B}_{r+1}$
5. $u_{k'} > u_k$.

*Then $\sigma$ is weakly dominated.*

The proof shows that the change in the value of switching strategies is non-negative. To prove this, note that on those partition elements where no attribute is examined after $k$, switching to strategy $\sigma'$ can only result in a gain. On the partition elements where attributes $k$ and $k'$ are taken, switching to $\sigma'$ may result in a loss of value. The proof starts by partitioning the set of possible signals from examining attribute $k$ in $\sigma$ into those signals after which no more attributes are examined and those after which attribute $k'$ is examined. Note that a signal is just a list of the records in the database that have the examined attribute. This allows us to eliminate terms and show that for any response to examining attribute $k$, the value gained from switching to attribute $k'$ on the partition elements where another attribute is not examined is at least as great as the loss on the partition elements where both $k$ and $k'$ are examined.

*Proof of Lemma 3.2:* Define $\mathbf{L}$ as the set of vectors resulting from experiment $k$ in $\sigma$ (or experiment $k'$ in $\sigma'$) such that no experiment is taken on the resulting partition element.

$$\mathbf{O} \supseteq \mathbf{L} = \{x_{\cdot k} = (x_1, \ldots, x_m) \in \mathbf{O} \mid \text{ For } \langle x_{\cdot j} \rangle_{j \in J^*} \, x_{\cdot k} : \alpha_{q+1}(B) = 0\}$$

$$\Gamma(\sigma') = \Gamma(\sigma)$$

$$\begin{aligned}
\Omega(\sigma') - \Omega(\sigma) = {} & \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2}\, u_k \right\} \\
& + \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj}) \\
& - \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2}\, u_{k'} \right\} \\
& - \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj}) \\
= {} & \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2}\, u_k \right\} \\
& + \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})
\end{aligned}$$

$$- \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2}\, u_{k'} \right\}$$

$$- \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})$$

Note that

$$\frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{O}/\mathbf{L}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})$$
$$= \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}/\mathbf{L}} \sum_{j \in \mathbf{O}/\mathbf{L}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})$$

so the terms cancel. Also note that

$$\frac{1}{2^{2m}} \sum_{i \in \mathbf{L}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})$$
$$= \frac{1}{2^{2m}} \sum_{i \in \mathbf{L}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})$$

so these terms can be added back in, and we have:

$$\Omega(\sigma') - \Omega(\sigma)$$
$$= \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2} u_k \right\}$$
$$+ \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})$$
$$- \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2} u_{k'} \right\}$$
$$- \frac{1}{2^{2m}} \sum_{i \in \mathbf{O}} \sum_{j \in \mathbf{L}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})$$

Reversing the sums, we get:

$$\begin{aligned}
&\Omega(\sigma') - \Omega(\sigma) \\
&= \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2} u_k \right\} \\
&+ \frac{1}{2^{2m}} \sum_{i \in \mathbf{L}} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj}) \\
&- \frac{1}{2^m} \sum_{i \in \mathbf{L}} \left\{ \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2} u_{k'} \right\} \\
&- \frac{1}{2^{2m}} \sum_{i \in \mathbf{L}} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})
\end{aligned}$$

So we need to prove that for arbitrary $\mathbf{L}$ (specifically, any $i$), we have:

$$\begin{aligned}
&\max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2} u_k + \frac{1}{2^m} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj}) \\
&\geq \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2} u_{k'} + \frac{1}{2^m} \sum_{j \in \mathbf{O}} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj})
\end{aligned}$$

This is true if for all $i$ the following is true:

$$\begin{aligned}
&\max_h (u_h^* + u_{k'}\, x_{hi}) + \frac{1}{2} u_k - \max_h (u_h^* + u_k\, x_{hi}) + \frac{1}{2} u_{k'} \\
&\geq \frac{1}{2^m} \sum_{j \in \mathbf{O}} \{\max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj}) \\
&\qquad - \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})\}
\end{aligned}$$

We can rearrange the second term of the inequality to get:

$$\begin{aligned}
&\frac{1}{2^m} \sum_{j \in \mathbf{O}} \{\max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj}) - \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})\} \\
&= \frac{1}{2} [\max_h (u_h^* + u_{k'}\, x_{hi}) + u_k - \max_h (u_h^* + u_k\, x_{hi}) + u_{k'}] \\
&+ \frac{1}{2^m} [ \sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'}\, x_{hi} + u_k\, x_{hj}) \\
&\qquad - \sum_{j \in \mathbf{O} | h^\dagger j = 0} \max_h (u_h^* + u_k\, x_{hi} + u_{k'}\, x_{hj})]
\end{aligned}$$

where $h^*$ is the alternative that maximizes $\max_h (u_h^* + u_{k'} x_{hi})$ and where $h^\dagger$ is the alternative that maximizes $\max_h (u_h^* + u_k x_{hi})$. So we now need to show that for all $i$ we have:

$$\frac{1}{2} [\max_h (u_h^* + u_{k'} x_{hi}) - \max_h (u_h^* + u_k x_{hi})] \\ \geq \frac{1}{2^m} [ \sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj}) \\ - \sum_{j \in \mathbf{O} | h^\dagger j = 0} \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj})]$$

There are three possible cases.

*Case 1:*

$\max_h (u_h^* + u_{k'} x_{hi})$ and $\max_h (u_h^* + u_k x_{hi})$ are maximized by the same alternative $h$, and $h$ is the first alternative in vector $i$ with a 1. Now the sums are over the same set of vectors (*i.e.*, $h^* j$ is the same as $h^\dagger j$).

So we want to show:

$$\frac{1}{2} (u_{k'} - u_k) \\ \geq \frac{1}{2^m} \sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj}) \\ - \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj})$$

The term:

$$\max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj}) - \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj})$$

is maximized when $x_{hi} = 1$ and $x_{hj} = 0$, so we have:

$$\frac{1}{2} (u_{k'} - u_k) \geq \frac{1}{2^m} \sum_{j=1}^{2^{m-1}} (u_{k'} - u_k) \geq 0$$

*Case 2:*

$\max_h (u_h^* + u_{k'} x_{hi})$ and $\max_h (u_h^* + u_k x_{hi})$ are maximized by the same alternative $h$, and that $h$ is the first alternative. Additionally, $x_{1i} = 0$.

In this case, we have:

$$u_1^* \geq u_h^* + u_{k'} x_{hi} \quad \text{for all } h, \quad \text{and } x_{1i} = 0, \tag{2}$$

so we need to show:

$$u_1^* + \frac{1}{2}u_k + \frac{1}{2}(u_1^* + u_{k'}) - u_1^* - \frac{1}{2}u_{k'} - \frac{1}{2}(u_1^* + u_k)$$
$$\geq \frac{1}{2^m}[\sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj})$$
$$- \sum_{j \in \mathbf{O} | h^\dagger j = 0} \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj})].$$

We have:

$$\sum_{j \in \mathbf{O} | h^\dagger j = 0} \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj}) \quad (3)$$

$$\geq \frac{1}{2^m}[\sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj})], \quad (4)$$

since the $h$ that maximizes the right hand side of Equation 4 must be either $h = 1$ or some $h^*$ such that $x_{h^* i} = 1$ and $x_{h^* j} = 1$. This is by Equation 2.

*Case 3:*

$\max_h (u_h^* + u_{k'} x_{hi})$ and $\max_h (u_h^* + u_k x_{hi})$ are maximized by different alternatives $h^*$ and $h^\dagger$. This implies that $h^*$ is the first alternative with a 1 in vector $i$, and $h^\dagger$ is alternative 1.

In this case, we have:

$$u_1^* \geq u_h^* + u_k x_{hi} \quad \text{for all } h, \quad \text{and } x_{1i} = 0$$
$$u_{h^*}^* + u_{k'} x_{h^* i} \geq u_h^* + u_{k'} x_{hi} \quad \text{for all } h, \quad \text{and } x_{h^* i} = 1$$

So we need to show:

$$\frac{1}{2}[u_{h^*}^* + u_{k'} - u_1]$$
$$\geq \frac{1}{2^m}[\sum_{j \in \mathbf{O} | h^* j = 0} \max_h (u_h^* + u_{k'} x_{hi} + u_k x_{hj})$$
$$- \sum_{j \in \mathbf{O} | h^\dagger j = 0} \max_h (u_h^* + u_k x_{hi} + u_{k'} x_{hj})]$$

We can break up the second part of the equation to get:

$$\frac{1}{2}\left[u^*_{h^*} + u_{k'} - u_1\right]$$

$$\geq \frac{1}{2^m}\left[\sum_{\substack{j\in \mathbf{O}|\\ h^*j=0\\ \wedge 1j=0}} \max_h \left(u^*_h + u_{k'}\, x_{hi} + u_k\, x_{hj}\right) + \sum_{\substack{j\in \mathbf{O}|\\ h^*j=0\\ \wedge 1j=1}} \max_h \left(u^*_h + u_{k'}\, x_{hi} + u_k\, x_{hj}\right)\right.$$

$$\left. - \sum_{\substack{j\in \mathbf{O}|\\ ij=0\\ \wedge h^*j=0}} \max_h \left(u^*_h + u_k\, x_{hi} + u_{k'}\, x_{hj}\right) - \sum_{\substack{j\in \mathbf{O}|\\ ij=0\\ \wedge h^*j=1}} \max_h \left(u^*_h + u_k\, x_{hi} + u_{k'}\, x_{hj}\right)\right]$$

The first and third terms on the right hand side are equal, so they cancel. The second and fourth terms can be rewritten based on the alternative chosen. Before doing so, note that the $j$'s in each of these terms can be paired. The pairs of vectors are identical, except in one, $h^*\, j = 1$ and $1\, j = 0$ and in the other, $h^*\, j = 0$ and $1\, j = 1$. So we have:

$$\frac{1}{2^m}\left[\sum_{\substack{j\in \mathbf{O}|\\ h^*j=0\\ \wedge 1j=0}} \max_h \left(u^*_h + u_{k'}\, x_{hi} + u_k\, x_{hj}\right) - \sum_{\substack{j\in \mathbf{O}|\\ ij=0\\ \wedge h^*j=1}} \max_h \left(u^*_h + u_k\, x_{hi} + u_{k'}\, x_{hj}\right)\right]$$

$$= \frac{1}{2^m} \left[ \sum_{\substack{j \in \mathbf{O} | \\ h^* j = 0 \\ \wedge 1 j = 0 \\ \wedge h = 1}} (u_1^* + u_k) + \sum_{\substack{j \in \mathbf{O} | \\ h^* j = 0 \\ \wedge 1 j = 1 \\ \wedge h = h^*}} (u_{h^*}^* + u_{k'}) + \sum_{\substack{j \in \mathbf{O} | \\ h^* j = 0 \\ \wedge 1 j = 1 \\ \wedge h = h'' \notin \{1, h^*\}}} u_{h''}^* + u_{k'} \, x_{h'' i} + u_k \, x_{h'' j} \right.$$

$$\left. - \sum_{\substack{j \in \mathbf{O} | \\ 1 j = 0 \\ \wedge h^* j = 1 \\ \wedge h = h^*}} (u_{h^*}^* + u_k + u_{k'}) - \sum_{\substack{j \in \mathbf{O} | \\ i j = 0 \\ \wedge h^* j = 1 \\ \wedge h = 1}} (u_1^*) - \sum_{\substack{j \in \mathbf{O} | \\ i j = 0 \\ \wedge h^* j = 1 \\ \wedge h = h'' \notin \{1, h^*\}}} u_{h''}^* + u_k \, x_{h'' i} + u_{k'} \, x_{h'' j} \right]$$

$$\leq \frac{1}{2} [u_{h^*}^* + u_{k'} - u_1]$$

### 3.4. Arbitrary attribute

The last attribute case and the second to last attribute case were proven for arbitrary search strategies. The induction step is true only for optimal strategies, so the proof is more involved. We want to look at examining an arbitrary attribute and show that if $\sigma$ is a strategy such that:

• for some set of database instantiations, $B^*$, $\sigma$ examines attribute $k$;

• for the substrategies, $\sigma(B_i)$, that follow from the partition elements in $\iota(B^*, k)$, the attributes are examined in descending order of their value;

• on the partition elements that follow from experiment $k$ the database management system examines attribute $k'$ or stops and answers the query; and that

• the user values attribute $k'$ more than attribute $k$,

then $\sigma$ is weakly dominated by a strategy $\sigma'$ that examines attribute $k$ when $\sigma$ examines $k'$ and examines attribute $k'$ when $\sigma$ examines $k$. In Figure 2, this applies to both the initial state, where the database management system has no information for answering the query, and to the state

$$\begin{array}{|cccccc|} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ \hline 0 & & & & & \\ 0 & & & & & \\ 0 & & & & & \\ \hline \end{array}.$$

There are two cases for this proof. The first is where examining an attribute is valuable by itself, and the second is where examining an attribute is valuable only in conjunction with the rest of the strategy. For the first case, we:

1. Show where the two strategies, $\sigma$ and $\sigma'$, will result in partition elements where the database management system will return the same record(s), and therefore have equal value. We do this as follows:

(a) Show that stopping after examining attributes $k$ and $k'$ implies one alternative in the query response must have attribute $k'$;

(b) Define the partition elements that have records in the query response with both attributes $k$ and $k'$;

(c) Calculate the number of partition elements, $B$, that have records in the query response with both attributes $k$ and $k'$; and

(d) Show that for these partition elements, no more information is obtained (*i.e.*, after examining attributes $k$ and $k'$, the database management system answers the query, and at least one of the records in the query response has both attributes).

2. Calculate the minimum difference in the value of the two strategies.

Define $\mathbf{L}$ as the set of vector values that occur from examining attribute $k$ in $\sigma$, such that no attributes are examined after than value is obtained.

DEFINITION 3.5:

$$\mathbf{L} = \{x_{\cdot k} \in \mathbf{O} | B \in \mathbf{B}_t, \text{ where } B = \langle x_{\cdot j} \rangle_{j \in J^*},\ x_{\cdot k} \quad \text{and} \quad \alpha_{t+1}(B) = 0\}$$

In Figure 2, for $k$ defined as attribute $S_1$ and $B \in \mathbf{B}_1$ being the state where the database management system has no information for answering the query,

$$\mathbf{L} = \{(1, 1, 0)^T, (1, 0, 1)^T, (0, 1, 1)^T, (1, 0, 0)^T, (0, 1, 0)^T, (0, 0, 1)^T, \}$$

If the true database instantiation is such that any of these are true, then the database management system will answer the query after examining attribute $S_1$.

Now partition **L**, the set of vectors after which experiment $\alpha_{q+1} = 0$ in $\sigma$, based on the first record to have attribute $k$.

DEFINITION 3.6:

$$\mathbf{L} = \mathbf{L}_1 \cup \mathbf{L}_2 \cup \ldots \cup \mathbf{L}_H$$

*where*

$$\mathbf{L}_h = \{x_{\cdot k} \in \mathbf{L} | x_{1,k} = 0, \ldots, x_{i-1,k} = 0, \\ x_{i,k} = 1, x_{i',k} \in \{0, 1\}, i' = i+1, \ldots, m\}$$

So for a database with three records, $\mathbf{L}_1 \subseteq \{(1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0)\}$; $\mathbf{L}_2 \subseteq \{(0, 1, 1), (0, 1, 0)\}$; $\mathbf{L}_3 \subseteq \{(0, 0, 1)\}$. In Figure 2, for the case described above, $\mathbf{L}_1 = \{(1, 1, 0), (1, 0, 1), (1, 0, 0)\}$; $\mathbf{L}_2 = \{(0, 1, 1), (0, 1, 0)\}$; and $\mathbf{L}_3 = \{(0, 0, 1)\}$. This will help define where changing strategies has no effect, a positive effect and a negative effect on the expected value of the query response.

Define the value of examining an attribute conditional on the existing information as $\nu(k, B)$.

DEFINITION 3.7: *Given that the database management system has examined $\# J^*$ attributes, the expected value of examining one more attribute, $k \notin J^*$, is:*

$$\nu(k, B) = \sum_{h=1}^{r^*} \Psi_h [\Delta_h + u_k] - \frac{1}{2} u_k - c,$$

*where equivalence class, $h$, $h \in 1, \ldots, r^*$ has $n_h$ records and the expected utility of any record in class $h$ is greater than the expected utility of any record in class $h+1$. $\Delta_h$ is the difference between the value of equivalence class $h$ and equivalence class 1 (i.e., $\Delta_h = u_1^* - u_j^*$ where $x_j$ is a member of the $h$-th equivalence class). $r^*$ is the largest number such that $\Delta_h + u_k$*

*is non-negative.* $\Psi_h$ *is the probability that the expected value of the choice set will be increased by* $[\Delta_h + u_k]$,

$$\Psi_h = \left(1 - \frac{1}{2^{n_h}}\right) \prod_{j=1}^{h-1} \left(\frac{1}{2^{n_j}}\right).$$

We use the following lemma to bound the expected gain resulting from using strategy $\sigma'$ instead of strategy $\sigma$. It says that if the database management system examines two attributes $k$ and $k'$ and then stops and answers the query, then at least one record in the query response has attribute $k$.

LEMMA 3.3: *Let* $\sigma$ *be a strategy such that for some* $B^* \in \mathbf{B}_q$ *we have:*

1. $\alpha_q(B^*) = k$
2. $\nu(k, B^*) \geq 0$
3. $\iota(B^*, k) = \{B_1, ..., B_{2^n}\}$
4. $\alpha_{q+1}(B_i) \in \{0, k'\}, \ i = 1, ..., 2^n$
5. *For some* $B_i$ *where* $\alpha_{q+1}(B_i) = 0$, $x_{ik} = 0$ *for all* $x_{i\cdot} \in D(x_{\cdot k}; B^*)$

*Then* $\sigma$ *is weakly dominated.*

*Proof of Lemma 3.3:* Let $B' \in \iota(B^*, k)$ be such that $\alpha_{q+1}(B') = 0$ and $x_{ik} = 0$ for all $x_i \in D^*(B')$, then we want to show that $\nu(k, B*) > 0 \Rightarrow \nu(k', B') > 0$. This implies that there is a strategy $\sigma'$ with $\alpha_{q+1}(B') = k'$ that dominates $\sigma$.

$$\begin{aligned}
\nu(k, B) &= \sum_{h=1}^{r^*} \left[\prod_{j=1}^{h-1} \left[\frac{1}{2}\right]^{n_j}\right] \left[1 - \left[\frac{1}{2}\right]^{n_h}\right] [\Delta_h + u_k] - \frac{1}{2} u_k - c \\
&= \sum_{i=1}^{m'} \left[\frac{1}{2}\right]^i [x_i - x_1 + u_k] - \frac{1}{2} u_k - c \\
&= \sum_{i=2}^{m'} \left[\frac{1}{2}\right]^i [x_i - x_1 + u_k] - c \\
&< \sum_{i=2}^{m'} \left[\frac{1}{2}\right]^i [x_i + u_k x_{ik} - x_1 + u_{k'}] - c \\
&\leq \nu(k', B')
\end{aligned}$$

where $m'$ is the number of records that satisfy $\Delta_h + u_k$ being non-negative. □

In Figure 2, let $B^*$ be represented by $\begin{array}{|cccccc|} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ \hline 1 & & & & & \\ 1 & & & & & \\ 1 & & & & & \\ \hline \end{array}$, so $\alpha_q(B^*) = S_2$ and $\alpha_{q+1}(B_i) \in \{0, S_3\}$. According to the lemma, the strategy $\sigma$ is dominated because in the partition element $\begin{array}{|cccccc|} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ \hline 1 & 0 & & & & \\ 1 & 0 & & & & \\ 1 & 0 & & & & \\ \hline \end{array}$, no record has attribute $S_2$, but using this strategy, the database management system will not examine attribute $S_3$ before answering the query.

Now recall that the set of records in an arbitrary database instantiation in $B^*$ is ordered so that

$$u_1^* \geq u_2^* \geq \dots \geq u_m^*.$$

Also assume a lexicographic rule for determining the record the database management system will return. This implies that if the expected utility of $x_{i\cdot}$ equals the expected utility of $x_{j\cdot}$, then the database management system will return $x_{i\cdot}$ as the answer to the query if and only if $i < j$. Using lemma 3.3, we can prove the following proposition, which identifies the records in the query response with an associated $\mathbf{L}_h$.

Proposition 3.1: *Assume*

$$B \equiv \langle x_{\cdot j} \rangle_{j \in J^*} \, x_{\cdot k}, \quad x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L}$$

*and* $x_{ik} = 1$, *then for those* $B' \in \iota(B, k')$, *where* $B' \equiv \langle x_{\cdot j} \rangle_{j \in J^*} \, x_{\cdot k} \, x_{\cdot k'}$ *and* $x_{\cdot k'} \in \mathbf{L}_i$, *we have* $x_{i\cdot} \in D^*(B')$.

*Proof of Proposition 3.1:* The proposition follows if:

$$u_i^* + u_k + u_{k'} \geq u_h^* + u_k \, x_{hk} + u_{k'} \, x_{hk'} \quad \text{for } h = 1, \dots, m.$$

There are two cases.

1. $h < i \Rightarrow u_h^* \geq u_i^*$
2. $h > i \Rightarrow u_i^* \geq u_h^*$

*Case 1:* $h < i \Rightarrow u_h^* \geq u_i^*$

By the definition of $\mathbf{L}_h$ $x_{hk'} = 0$. By Lemma 3.3 and the definition of $\mathbf{L}_h$

$$u_i^* + u_k \geq u_h^* \Rightarrow u_i^* + u_k + u_{k'} \geq u_h^* + u_k \, x_{hk}.$$

*Case 2:* $h > i \Rightarrow u_i^* \geq u_h^*$

$$u_i^* + u_k + u_{k'} - u_h^* - u_k\, x_{hk} - u_{k'}\, x_{hk'} \geq u_i^* - u_i^* - u_h^* \geq 0 \quad \square$$

Next we examine those cases where both strategies $\sigma$ and $\sigma'$ result in the same query response, even though the database management system has examined different attributes and has different knowledge of the database contents. To examine these cases, assume the database management system uses a retrieval algorithm based on strategy $\sigma$, and the set of possible database instantiations is defined as

$$B' \equiv \langle x_{\cdot j} \rangle_{j \in J^*}\, x_{\cdot k}\, x_{\cdot k'}, \quad \textit{where } x_{\cdot k} \in \mathbf{O/L}, \quad \textit{and } x_{\cdot k'} \in \mathbf{L}_h.$$

In these cases, if the database management system uses a retrieval algorithm based on strategy $\sigma'$, the set of possible database instantions will be defined as

$$B'' \equiv \langle x_{\cdot j} \rangle_{j \in J^*}\, x''_{\cdot k'}\, x''_{\cdot k}, \quad \textit{where } x''_{\cdot k'} = x_{\cdot k}, \quad \text{and } x''_{\cdot k} = x_{\cdot k'}$$

and where $x''_{ik'} = 1$ and $x''_{ik} = 1$. This implies that using $\sigma$ and returning record $x_i$ gives the user an expected utility of $u(x_i;\, B,\, x_{\cdot k'},\, x_{\cdot k})$. Using strategy $\sigma'$ and returning $x_i$ gives the user an expected utility of $u(x_i;\, B,\, x_{\cdot k},\, x_{\cdot k'})$ and $u(x_i;\, B,\, x_{\cdot k'},\, x_{\cdot k}) - u(x_i;\, B,\, x_{\cdot k},\, x_{\cdot k'}) = 0$.

For these partition elements, we show that if a record would be in the database management system's response to the query after examining attribute $k'$ and if that record has both attributes $k'$ and $k$ and if the records having attribute $k'$ are given by $x_{\cdot k'} \in \mathbf{L}$ in $\sigma$, then the database management system will not examine any more attributes to determine the query response.

PROPOSITION 3.2: *Let $\sigma$ be a strategy such that for some $B^* \in \mathbf{B}_q$ we have:*

- $\alpha_q(B^*) = k$
- $\nu(k, B^*) > 0$
- $\iota(B^*, k) = \{B_1, \ldots, B_{2^n}\}$
- $\alpha_{q+1}(B_i) \in \{0, k'\},\ i = 1, \ldots, 2^n$
- $\iota(B_i, k') = \{B_1^i, \ldots, B_{2^n}^i\}$
- $\alpha_t\left(B_j^i\right) \in \{0, t\} \quad \textit{for } j = 1, \ldots, 2^n \quad \textit{and for } t > q+1$
- $u_t > u_{t+1} \quad \textit{for } t > q+1$
- $u_{k'} > u_k.$

*If* $\hat{B} \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x_{\cdot k} x_{\cdot k'} \in \iota(B_i, k')$, *where:*

- $x_{\cdot k} \in \mathbf{O}/\mathbf{L}$.
- $x_{\cdot k'} \in \mathbf{L}$.

*and if* $x_i \in D^*(\hat{B})$ *is such that:*

- $x_{ik} = 1$.
- $x_{ik'} = 1$.

*Then* $\hat{B} \in \mathbf{B}_{r+1}$.

*Proof of Proposition 3.2:* By assumption, we know that the expected value of any substrategy that takes an experiment on a partition of the form $B \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x_{\cdot k}$ for $x_{\cdot k} \in \mathbf{L}$ is less than zero. Therefore, the expected value of taking an experiment on a partition of the form $B' \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x_{\cdot k'}$ for $x_{\cdot k} \in \mathbf{L}$ is less than zero.

Rewrite $\hat{B}$ as $\hat{B} \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x_{\cdot k'} x_{\cdot k}$ for $x_{\cdot k'} \in \mathbf{L}$ and $x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L}$ and assume that $x_{ik} = 1$, $x_{ik'} = 1$ and $x_i \in D^*(\hat{B})$. The expected value of taking any experiment on $B$ must be less than or equal to the expected value of taking any experiment on $B'$, since the value of taking experiments stems from their ability to alter the choice set. In $\hat{B}$, the element in the choice set has a value of $u_i^* + u_k + u_{k'}$; whereas in $B'$, the value of an element in the choice set is $u_i^* + u_k$ and the value of $x_{ik'}$ is unknown. All alternatives other than $x_i$ in $\hat{B}$ are no closer, and possibly further (by $u_k$) from the value of $x_i$ than in $B'$, so it is less likely that experimentation will change the choice set. □

Proposition 3.2 characterizes cases where it is optimal to stop examining attributes after examining attribute $k'$. The next lemma specifies the number of possible database instantiations where it is optimal to stop and answer the query after examining attribute $k'$.

DEFINITION 3.8: *Define* $\overline{\overline{\mathbf{O}}}$ *as the number of elements in* $\mathbf{O}$; $\overline{\overline{\mathbf{L}}}$ *as the number of elements in* $\mathbf{L}$ *and* $\overline{\overline{\mathbf{L}}}_h$ *as the number of elements in* $\mathbf{L}_h$

LEMMA 3.4: *There are* $(1/2\,\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}}_h)\,\overline{\overline{\mathbf{L}}}_h$ *partition elements* $B' \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x_{\cdot k} x_{\cdot k'}$ *with* $x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L}$ *and* $x_{\cdot k'} \in \mathbf{L}_h$, *where* $x_{ik} = 1$ *and* $x_{ik'} = 1$ *and where* $x_i \in D^*(B')$. *For each of these* $B'$, *there exist in* $\sigma'$ *a*

$$B'' \equiv \langle x_{\cdot j} \rangle_{j \in J^*} x''_{\cdot k'} x''_{\cdot k}, \quad x''_{\cdot k'} = x_{\cdot k}, \quad x''_{\cdot k} = x_{\cdot k'}$$

*and where* $x''_{ik'} = 1$ *and* $x''_{ik} = 1$. *For* $h = 1, \ldots, H$.

*Proof of Lemma 3.4:*

- $\iota(B^*, k) = \{B_1, \ldots, B_{2^m}\}$
- $\#\iota(B^*, k) = \overline{\overline{\mathbf{O}}}$
- $\Pr(x_{ik} = 1) = 1/2 \Rightarrow \#\{B_i \in \iota(B^*, k) | x_{ik} = 1\} = 1/2\,\overline{\overline{\mathbf{O}}}$.

By the definition of $\mathbf{L}_h$, on exactly $\overline{\overline{\mathbf{L}}}_h$ of these $B_i$ we have $\alpha_{q+1}(B_i) = 0$, so they are not of interest. On the remaining $B_i$, we have $\alpha_{q+1}(B_i) = k'$. For those $B_i$, such that $\alpha_{q+1}(B_i) = k'$, we are only interested in those where $x_{\cdot k'} \in \mathbf{L}$. Of these $B_i$ of interest, exactly $\overline{\overline{\mathbf{L}}}_h$ have $x_{ik'} = 1$, by the definition of $\mathbf{L}_h$. □

Using the previous propositions and lemmas, we can prove the induction step for Case 1.

Induction Step: *Let $\sigma$ be as strategy such that for some $B^* \in \mathbf{B}_q$ we have*:

1. $\alpha_q(B^*) = k$
2. $\nu(k, B^*) > 0$
3. $\iota(B^*, k) = \{B_1, \ldots, B_{2^n}\}$
4. $\alpha_{q+1}(B_i) \in \{0, k'\},\ i = 1, \ldots, 2^n$
5. $\iota(B_i, k') = \{B_1^i, \ldots, B_{2^n}^i\}$
6. $\alpha_t(B_j^i) \in \{0, t\}$ *for* $j = 1, \ldots, 2^n$ *and for* $t > q + 1$
7. $u_t > u_{t+1}$ *for* $t > q + 1$
8. $u_{k'} > u_k$.

*Then $\sigma$ is weakly dominated.*

In the proof, we show that $\Omega(\sigma') - \Omega(\sigma) \geq 0$. We do this by showing that the value gained from switching strategies is bounded from below by

$$\frac{1}{2^m}\overline{\overline{\mathbf{L}}}\frac{1}{2}(u_{k'} - u_k)$$

and the value lost is bounded from above by

$$\frac{1}{2^{2m}}\left[(\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}})\overline{\overline{\mathbf{L}}} - \sum_{i=1}^{H}\left(\frac{1}{2}\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}}_h\right)\overline{\overline{\mathbf{L}}}_h(u_k - u_{k'})\right],$$

which is due to the preceding proposition. Thus, the value gained is greater than or equal to the value lost. We are now done with Case 1.

*Proof of Induction Step*:

### 3.4.1. *Definition of $\sigma'$*

If $\alpha_q(B_i) = k'$ for all $i$, then we are done. $\mathbf{B}_{r+1}^{\sigma} = \mathbf{B}_{r+1}^{\sigma'}$, so $\Omega(\sigma) = \Omega(\sigma')$. Since the cost of the experiments is uniform, $\Gamma(\sigma) = \Gamma(\sigma')$.

So assume that $\alpha_{q+1}(B_i) = 0$ for some $i \in \{1, \ldots, 2^n\}$. Also assume that if $\alpha_{q+1}(B_i) = 0$ then no experiment was taken because the expected value of taking another experiment was less than zero. If this assumption does not hold, then we can replace $\sigma$ with a strategy $\sigma''$ where the assumption does hold, and where the expected payoff from $\sigma''$ is greater than the expected payoff from $\sigma$.

Define $\sigma'$ to be the same strategy as $\sigma$ except in $\sigma'$, and we have:

- $\alpha_q(B^*) = k'$
- $\iota(B^*, k') = \{B'_1, \ldots, B'_{2^n}\}$
- $$\alpha_{q+1}(B'_1) = \begin{cases} k & \text{if in } \sigma \quad \alpha_{q+1}(B_i) = k' \\ & \text{where} \quad B'_i \equiv \langle x_{\cdot j}\rangle_{j \in J^*}\, x'_{\cdot k'} \\ & \qquad\qquad B_i \equiv \langle x_{\cdot j}\rangle_{j \in J^*}\, x_{\cdot k'} \\ & \text{and} \quad x'_{\cdot k'} = x_{\cdot k} \\ 0 & \text{if in } \sigma \quad \alpha_{q+1}(B_i) = 0 \\ & \text{where} \quad B'_i \equiv \langle x_{\cdot j}\rangle_{j \in J^*}\, x'_{\cdot k'} \\ & \qquad\qquad B_i \equiv \langle x_{\cdot j}\rangle_{j \in J^*}\, x_{\cdot k} \\ & \text{and} \quad x'_{\cdot k'} = x_{\cdot k} \end{cases}$$
- $\iota(B'_i, k) = \{B^i_1, \ldots, B^i_{2^n}\}$
- $$\sigma'(B^{i'}_{j'}) = \begin{cases} \sigma(B^i_j) & \text{if } \sigma \quad B^{i'}_{j'} \equiv \langle x_{\cdot h}\rangle_{j \in J^*}\, x'_{\cdot k'}\, x'_{\cdot k} \\ & \qquad x'_{\cdot k'} \in \mathbf{O}\backslash\mathbf{L}\; x'_{\cdot k} \in \mathbf{O}\backslash\mathbf{L} \\ & \text{where} \quad B^i_j \equiv \langle x_{\cdot h}\rangle_{j \in J^*}\, x_{\cdot k}\, x_{\cdot k'} \\ & \qquad x'_{\cdot k} \in \mathbf{O}\backslash\mathbf{L}\; x'_{\cdot k'} \in \mathbf{O}\backslash\mathbf{L} \\ & \text{and} \quad x'_{\cdot k} = x_{\cdot k} \\ & \text{and} \quad x'_{\cdot k'} = x_{\cdot k'} \\ \sigma(B^i_j) & \text{if} \quad B^{i'}_{j'} \equiv \langle x_{\cdot h}\rangle_{j \in J^*}\, x'_{\cdot k'}\, x'_{\cdot k} \\ & \qquad x'_{\cdot k'} \in \mathbf{O}\backslash\mathbf{L}\; x'_{\cdot k} \in \mathbf{L} \\ & \text{where} \quad B^i_j \equiv \langle x_{\cdot h}\rangle_{j \in J^*}\, x_{\cdot k}\, x_{\cdot k'} \\ & \qquad x'_{\cdot k} \in \mathbf{O}\backslash\mathbf{L}\; x'_{\cdot k'} \in \mathbf{L} \\ & \text{and} \quad x'_{\cdot k} = x_{\cdot k'} \\ & \text{and} \quad x'_{\cdot k'} = x_{\cdot k} \end{cases}$$

Given this definition of $\sigma'$, we have:

$$\Omega(\sigma') - \Omega(\sigma) \tag{5}$$

$$= \sum_{B \equiv \langle x_{\cdot j}\rangle_{j \in J(B^*)}\, x_{\cdot k'};\; x_{\cdot k'} \in \mathbf{L}} \; \sum_{x \in B} \phi(x)\, \omega(x, \delta(B)) \tag{6}$$

$$+ \sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k'} \, x_{\cdot k};\, x_{\cdot k'} \in \mathbf{O} \backslash \mathbf{L} \, x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L}} \sum_{x \in B'} \phi(x) \, \omega(x, \delta(B')) \tag{7}$$

$$+ \sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k'} x_{\cdot k};\, x_{\cdot k'} \in \mathbf{O} \backslash \mathbf{L} \, x_{\cdot k} \in \mathbf{L}} \sum_{x \in B'} \phi(x) \, \omega(x, \delta(B')) \tag{8}$$

$$- \sum_{B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k};\, x_{\cdot k} \in \mathbf{L}} \sum_{x \in B} \phi(x) \, \omega(x, \delta(B)) \tag{9}$$

$$- \sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k} \, x_{\cdot k'};\, x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L} \, x_{\cdot k'} \in \mathbf{O} \backslash \mathbf{L}} \sum_{x \in B'} \phi(x) \, \omega(x, \delta(B')) \tag{10}$$

$$- \sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k} \, x_{\cdot k'};\, x_{\cdot k} \in \mathbf{O} \backslash \mathbf{L} \, x_{\cdot k'} \in \mathbf{L}} \sum_{x \in B'} \phi(x) \, \omega(x, \delta(B')) \tag{11}$$

$$= \frac{\overline{\overline{\mathbf{L}}}}{2^m} \frac{1}{2} (u_{k'} - u_k) \tag{12}$$

$$+ \sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k'} \, x_{\cdot k};\, x_{\cdot k'} \in \mathbf{O} \backslash \mathbf{L} \, x_{\cdot k} \in \mathbf{L}} \sum_{x \in B'} \phi(x) \, \omega(x, \delta(B')) \tag{13}$$

$$\sum_{B' \in \mathbf{B}_{r+1}(B) | B \equiv \langle x_{\cdot j} \rangle_{j \in J(B^*)} x_{\cdot k} x_{\cdot k'} ; x_{\cdot k} \in \mathbf{O} \setminus \mathbf{L}\, x_{\cdot k'} \in \mathbf{L}} \sum_{x \in B'} \phi(x)\, \omega(x, \delta(B')) \tag{14}$$

$$\leq \frac{\overline{\overline{\mathbf{L}}}}{2^m} \frac{1}{2} (u_{k'} - u_k) - \frac{1}{2^{2m}} \left[ (\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}}) \overline{\overline{\mathbf{L}}} - \sum_{i=1}^{H} \left( \frac{1}{2} \overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}}_h \right) \overline{\overline{\mathbf{L}}}_h \right] u_k - u_{k'}$$

$$= \frac{1}{2} \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} - \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} + \frac{\overline{\overline{\mathbf{L}\mathbf{L}}}}{\overline{\overline{\mathbf{O}\mathbf{O}}}} + \sum_{i=1}^{H} \left( \frac{1}{2} \frac{\overline{\overline{\mathbf{L}}}_h}{\overline{\overline{\mathbf{O}}}} - \right) - \sum_{i=1}^{H} \left( \frac{\overline{\overline{\mathbf{L}}}_h \overline{\overline{\mathbf{L}}}_h}{\overline{\overline{\mathbf{O}\mathbf{O}}}} \right) (u_{k'} - u_k)$$

$$= \frac{\overline{\overline{\mathbf{L}\mathbf{L}}}}{\overline{\overline{\mathbf{O}\mathbf{O}}}} - \sum_{i=1}^{H} \left( \frac{\overline{\overline{\mathbf{L}}}_h \overline{\overline{\mathbf{L}}}_h}{\overline{\overline{\mathbf{O}\mathbf{O}}}} \right) (u_{k'} - u_k) \geq 0 \tag{15}$$

where equation 13 is due to Lemma 3.3.

LEMMA 3.5: *Let $B \in \mathbf{B}_q$. If experiment $k$ is taken on $B$ and $x_{ik} = 1$ for some $x_i \in D^*(B)$, then $\nu(j', B') \leq \nu(j', B)$ for all $B' \in \iota(B, k)$ such that $x_{ik} = 1$ and $j' \notin J^* \cup k$.*

*Proof of Lemma 3.5:*

First order the alternatives in *B* by their expected value. Then $\nu(j, B)$ is given by:

$$\nu(j, B) = \sum_{h=1}^{m'} \left[ \frac{1}{2} \right]^h u_h^* - u_1^* + \frac{1}{2} u_j - c$$

where $m'$ is the highest numbered alternative such that $u_{m'}^* + u_j > u_1^*$.

Therefore

$$\nu(j, B) - \nu(j, B')$$
$$= \sum_{h=1}^{m'} \frac{1}{2} \left[ \frac{1}{2} \right]^{h-1} u_h^* - u_1^* + \frac{1}{2} u_j - c$$
$$- \sum_{h=1}^{m''} \frac{1}{2} \left[ \frac{1}{2} \right]^{h-1} u_h'^* - (u_1^* + u_j) + \frac{1}{2} u_j - c$$
$$= \sum_{h=1}^{m'} \frac{1}{2} \left[ \frac{1}{2} \right]^{h-1} u_h^* - u_1^* + - \sum_{h=1}^{m''} \frac{1}{2} \left[ \frac{1}{2} \right]^{h-1} u_h'^* - (u_1^* + u_j) \geq 0$$

since the known value of the $h$-th element in $B'$ is less than or equal to the known value of the $h$-th element of $B$ plus $u_j$, the difference between the $h$-th element and the first element is less in $B'$ than it is in $B$ for $h = 1, ...m$. □

For Case 2 there are two possibilities. The first possibility is that $\nu(k, B) < 0$ and $\nu(k', B) < 0$, but there is a joint strategy using experiments $k$ and $k'$ with a positive expected value. The second possibility is that $\nu(k, B) < 0$ and $\nu(k', B) > 0$. This second possibility reduces to Case 1, so we consider only the first scenario and show that $\overline{\overline{\mathbf{L}}}/\overline{\overline{\mathbf{O}}} > 1/2$, which implies that it is optimal to switch experiments.

From Lemma 3.5 it is easy to see that experiment $k'$ will only be taken on those elements of $\iota(B, k)$ where $x_{ik} = 0$ for all $x_i \in D^*(B)$. In addition, no experiment will be taken on the partition element $B' \equiv \langle x_{\cdot j} \rangle_{j \in J^*} \, x_{\cdot k}$ where $x_{\cdot k}$ is the vector of all 0s. This implies that $\overline{\overline{\mathbf{L}}}/\overline{\overline{\mathbf{O}}} > 1/2$, since the number of vectors where $x_{ik} = 0$ is 1/2. In addition, there are the $(\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}})^2$ partition elements that are the same under both strategies. So we have

$$\begin{aligned}\Omega(\sigma') - \Omega(\sigma) &\geq \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} \frac{1}{2} (u_{k'} - u_k) \\ &\quad - \frac{(\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}})\overline{\overline{\mathbf{O}}} - (\overline{\overline{\mathbf{O}}} - \overline{\overline{\mathbf{L}}})^2}{\overline{\overline{\mathbf{O}}}^2} (u_{k'} - u_k) \\ &= \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} \frac{1}{2} + \frac{\overline{\overline{\mathbf{L}}}^2}{\overline{\overline{\mathbf{O}}}^2} - \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} = \left( \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} - \frac{1}{2} \right) \frac{\overline{\overline{\mathbf{L}}}}{\overline{\overline{\mathbf{O}}}} > 0\end{aligned}$$

by Lemma 3.5.

## 4. CONCLUSION

By treating queries as choice problems, the database system becomes responsible for evaluating the alternatives and presenting a manageable set of information that is relevant to the choice problem. The retrieval strategy presented in this paper is orthogonal to the search strategy typically examined in the economic literature, but it conforms to decision strategies described in the behavioral psychology literature. Unlike traditional retrieval algorithms, it balances the accuracy of the information with the cost of retrieving appropriate data, which is particularly important for very large databases.

The optimal strategy developed in the paper depends heavily on the assumptions made at the beginning of Section 4. Changing the assumption that all attributes exist with probability 1.2 to all attributes exist with probability $p$ has little effect on the strategy. At each step, the attribute that maximized the net expected value would be examined. However, allowing each attribute to have a probability different from the others destroys the structure of the problem, and finding the best alternative will require the use of heuristics. One potential heuristic for this case is to select the attribute at each step that maximizes

$$\nu(k, B) = \sum_{h=1}^{r^*} \left[ \prod_{j=1}^{h-1} (1 - p_k)^{n_j} \right] [1 - p_k^{n_k}] [\Delta_h + u_k] - p_k u_k - c,$$

where $p_k$ is the probability that an alternative will have attribute $k$. This is a simple greedy heuristic that contains the optimal search strategy as a special case.

## REFERENCES


1. J. E. Bettman, *An Information Processing Theory of Consumer Choice.* Addison-Wesley, 1979.
2. M. R. Hagerty, The cost of simplifying preference models, *Marketing Science*, 1986, *5*, (4), pp. 298-323.
3. K. J. Lancaster, A new approach to consumer theory, *Journal of Political Economy,* 1966, *74*, (2), pp. 132-157.
4. J. C. Moore, W. B. Richmond and A. B. Whinston, *Searching choice oriented databases*, Working Paper, George Mason University Fairfax, VA 22030.
5. J. C. Moore, W. B. Richmond and A. B. Whinston, A decision theoretic approach to information retrieval, *ACM Transactions on Database Systems*, 1991, *15*, (3), pp. 311-340.
6. J. C. Moore and A. B. Whinston, A model for decision making with sequential information-acquisition, part 1, *Decision Support Systems,* 1986, *2*, pp. 285-307.
7. J. C. Moore and A. B. Whinston, A model for decision making with sequential information-acquisition, part 2, *Decision Support Systems,* 1987, *3*, pp. 47-72.
8. A. Tversky, Elimination by aspects: A theory of choice, *Psychological Review,* 1972, *79*, pp. 281-299.
9. A. Tversky and S. Sattath, Preference trees, *Psychological Review,* 1979, *86*, (6), pp. 542-573.