Space Constrained Selection Problems for Data Warehouses and Pervasive Computing

Themistoklis Palpanas[*]
University of California, Riverside
themis@cs.ucr.edu

Nick Koudas
AT&T Labs-Research
koudas@research.att.com

Alberto Mendelzon
University of Toronto
mendel@cs.toronto.edu

Abstract

Space constrained optimization problems arise in a multitude of important applications such as data warehouses and pervasive computing. A typical instance of such problems is to select a set of items of interest, subject to a constraint on the total space occupied by these items. Assuming that each item is associated with a benefit, for a suitably defined notion of benefit, one wishes to optimize the total benefit for the selected items.

In this paper, we show that in many important applications, one faces variants of this basic problem in which the individual items are sets themselves, and each set is associated with a benefit value. We present instances of such problems in the context of data warehouse management and pervasive computing, derive their complexity, and propose several techniques for solving them. Since there are no known approximation algorithms for these problems, we explore the use of greedy and randomized techniques.

We present a detailed performance study of the algorithms, highlighting the efficiency of the proposed solutions and the benefits of each approach. Finally, we present a worst-case analysis of the algorithms, which can be useful in practice for choosing among the alternatives. The solutions proposed in this paper are generic and likely to find applications in many more problems of interest than those mentioned above.

1. Introduction

In recent years there has been a proliferation in the amount of information that is being produced. The data that are being gathered and stored involve several aspects of human activity. Retailers register individual transactions in their stores, businesses keep track of the interactions with customers and suppliers, scientific laboratories record various measurements of interest. Obviously, the volume of data generated in such situations is huge.

It is often times the case that not all data are stored, because of their sheer size. In order to save space people usually aggregate data and only store the aggregates. A typical scenario in a retail store data warehouse, which records user transactions, would be to aggregate the sales at the day level [13]. Similarly, scientific measurements such as temperature, precipitation, air pollution, are aggregated at the hour level. In many cases even the size of the aggregated data is too large, so that portions of it have to be moved to tertiary storage or permanently deleted. Then, we have to choose which parts of the data to keep around.

Limited space can become a restricting factor even when the amount of data is not large, but when the actual available space is limited. Characteristic examples of this scenario are handheld devices, which have small storage capacity, and for which space allocation should be done with care. Mobile users in general face an analogous problem in terms of bandwidth. When the available bandwidth or connection time resources are limited, we have to choose carefully which bytes to transmit over the communication link.

Observe that all the above examples lead to space constrained optimization problems. We wish to fill the given space with the most useful of our data, or in other words with the data that will give us the highest benefit. The benefit is determined by the number and importance of the queries that we can answer based on the stored data. The view selection problem [9] in the context of data warehouses is an instance of such an optimization problem, and has been studied extensively in the literature.

In this paper we study the above optimization problem for the case when the benefit we get for admitting items in the solution is associated to *sets* of items, and not to individual items. In both the Knapsack and the view selection problems the benefit increases with each item that is added to the solution. However, this is not true for the class of applications that we consider. In this setting there are additional constraints inherent in the problem, which dictate that when we select a new item to insert in the solution the added benefit is zero, unless a set of related items are inserted as well. The above constraints make the problem harder, and in Section 6 we discuss how they affect the solution procedure. For the rest of the paper we refer to this optimization problem as the *Constrained Set Selection (COSS)* problem.

Several problems that have been presented in the literature [20, 22, 16] are related to the *COSS* problem. Nevertheless, none of the known results or techniques seem to be applicable in this case [15]. This study is the first thorough examination of practical algorithms that solve the *COSS* problem. Our experimental evaluation can serve both as a practitioner's guide, and also provide intuition about the nature of the problem from a theoretical perspective.

In this paper, we make the following contributions.

- We formulate optimization problems concerning the

[*] Work done while the author was at the University of Toronto.

- selection of *sets* of items that under a space constraint yield the highest benefit, where benefits are associated to sets of items. This kind of problems appear in various domains, and are very interesting in practice.
- We derive the complexity of the above optimization problems, and propose several algorithms for their solution. Since there are no known polynomial time approximation algorithms for these problems, we examine the use of known optimization principles in this context [6], such as greedy and randomization.
- We explore the properties of the above techniques with an experimental evaluation. Our results illustrate the behavior of the algorithms under different settings, and highlight the benefits of each approach.
- Based on our analysis, we present worst case scenarios for the algorithms. This offers insight into the operation of the algorithms, and provides a practical guide for selecting among the techniques proposed.

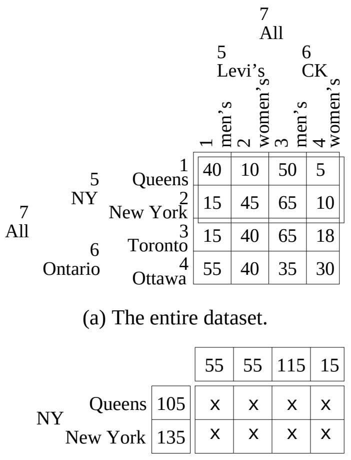
The outline of the paper is as follows. Section 2 illustrates examples in which the *COSS* problem is applicable, and Section 3 reviews the related work. In Section 4 we present the formulation of the optimization problem, and Section 5 proposes algorithms for their solution. In Section 6 we show experimental results evaluating the performance and the utility of the proposed algorithms. Finally, in Section 7 we present an analysis of the algorithms and discuss their relative performance.

2. Applications of the COSS Problem

In the following sections we present with more detail two specific example applications of the problem. The first application comes from the world of data warehousing, and the second from pervasive computing.

2.1. Aggregate Selection for Approximate Querying in Datacubes

The volume of data stored in OLAP tables is typically huge, often times in the order of multiple terabytes to petabytes. In large organizations, where terabytes of data are generated every day, it is common practice to aggregate these data in order to save storage space. During this procedure the detailed information that produced the aggregates is lost. In other cases the detailed data is moved to tertiary storage, which makes the task of accessing them very cumbersome. Nevertheless, users are often interested in inquiring about the data that generated the summarized form. In such cases, generating good estimates for the original data in response to queries is a pressing concern. Efficient solutions to the above problem have been proposed in the literature [4, 23, 3, 21], where the reconstruction techniques are based solely on the aggregated information in order to produce estimates for the detailed values.



	1 men's	2 women's	3 men's	4 women's
1 Queens	40	10	50	5
2 New York	15	45	65	10
3 Toronto	15	40	65	18
4 Ottawa	55	40	35	30

(a) The entire dataset.

		55	55	115	15
Queens	105	x	x	x	x
New York	135	x	x	x	x

(b) Aggregated values for the upper half of the dataset.

Figure 1. Example dimension hierarchies on two dimensional sales data.

Example 1 *Figure 1(a) shows a typical OLAP table, with two dimension attributes (location and jeans), and hierarchies defined in each dimension. Assume that for the state of NY (i.e., the upper half of the table) we only store the* aggregated *sales for each city and for each category as shown in Figure 1(b), and that we have deleted all the detailed values. The users might want to inquire about the number of redtab jeans sold in Queens NY (a point query), or they might request the number of any kind of jeans sold in each city of NY state (a range query). It turns out that we can use the information that is stored in the aggregates, in order to provide approximate answers to the above queries. The reconstruction algorithm will need both of the aggregates shown in Figure 1(b), and will be able to produce estimates for all the values marked as "x".*

Now consider the case where we are allowed only a limited amount of space for storing the aggregates. We can compute the space requirements of the aggregates, either by computing them, or by applying estimation techniques [24]. In this setting, we would like to materialize the subset of the aggregates that satisfies the given space constraint, and at the same time is necessary for the reconstruction of the important queries. The importance, or benefit, of each query can be determined automatically by observing the system workload, or it can be manually set by the user. Note that, unlike the view selection problem in data warehouses [9], in this case we get no partial benefits if we only materialize a subset of the aggregates. This is because the algorithms that reconstruct the detailed values [23, 3, 21] require a specified set of aggregates in its entirety.

2.2. Profile-Driven Data Management

We draw our second example from the area of *pervasive computing* [2]. The premise of pervasive computing is that all the relevant information (at each point in time and space) should be accessible to any mobile user. The users specify what information is relevant by using a profile language, which allows each user to define a number of queries that wants to be answered by the data items in the cache of the computer. Each query is associated with a benefit value, which reflects the query's utility to the user, and needs a set of the data items in order to be answered. The data items have specified storage requirements. Furthermore, each data item may help answer more than one queries.

One of the problems in the aforementioned context is what data to send to a mobile computer. The choice of the data to download is crucial, because it determines which of the queries in the profile can be answered, and consequently the total benefit that can be achieved. This is an interesting problem, in which we have to take into account both the user profile that describes what data are useful, and a space restriction on the mobile computer, or equivalently a connection time restriction (in both cases we can only transmit a limited specified amount of data).

Example 2 *Consider a user visiting a new city [2]. The profile of such a user could contain the following 3 statements (or queries).*

1. *Find location, requires map.*
2. *Find rental car, requires car agency list, and map.*
3. *Find restaurant, requires restaurant list, and map.*

In this example, if only one of the two required data items for queries 2 and 3 is present in the mobile computer, then we cannot answer those queries. There is no use having a list of restaurants if there is no map available, and vice versa. Therefore, the benefit gained is zero. Also observe that the map item is useful in answering all 3 queries.

3. Related Work

In the problem of view selection for data warehouses [9] we want, given a space constraint, to materialize a set of views in order to minimize the response time of queries to the data warehouse. In this case however, by selecting one of the views required by a query we get a fraction of the associated benefit. A subsequent study shows that the heuristics for the view selection problem do not provide any competitiveness guarantee against the optimal solution [12], and proposes the experimental comparison of the available algorithms. Many other studies [8, 25, 26, 1] deal with the view selection problem as well.

A similar optimization problem appears in the context of database design under the name of vertical partitioning [18, 19], where the *Bond Energy Algorithm* [17] has been employeed for its solution. In this domain the problem can be stated as follows. We have a set of applications accessing a particular set of attributes in the relations of a database. Each attribute requires a certain amount of space (to store the values of the attribute in the relation), and may be used by more than one applications. We want to find an allocation of the attributes to the memory hierarchy (e.g., main memory, local disk, network disk) so as to minimize the execution time of the applications. Similar to the view selection problem, partial benefits are credited, which is not true for our case.

The *COSS* problem is related to a family of optimization problems known as the Weighted constrained Maximum Value sub-Hypergraph problem [20], and the studies show that even simple instances of the problem do not accept polynomial solutions. We are not aware of any algorithms proposed for the *COSS* problem in that area.

Much work has been devoted to the *Set Cover* problem [5], which is similar to the optimization problem we are trying to solve, and efficient approximation algorithms [22] have been proposed for its solution. However, the same techniques do not seem to be applicable in our case. Furthermore, the *COSS* problem is also similar to the *k-Catalog Segmentation* problem [16]. But no interesting positive results are known for this problem either. In this study we try instead to examine the *COSS* optimization problem from a practical point of view, and we hope that the results we report will stimulate more work in this area.

The *COSS* problem has also appeared in the domains of aggregate selection for approximate querying in datacubes [21], and profile-driven data management [2], which exemplify its diverse practical applications. However, the solution of the optimization problem was not studied. All the algorithms described in this paper can be applied for the solution of the problem in these domains.

4. Problem Formulation

In the following paragraphs we establish the terminology used in the rest of the paper, we present the formal statement of the problem, and discuss its complexity.

Let *components* be the individual data items that are available for use, and *objects* be the consumers of the components. When all the components required by an object are selected in the solution we say that the object is *satisfied*. Each object is associated with a benefit, which is claimed when the object is satisfied. Components are associated with a space requirement, which is the storage space they will occupy when selected. When a component is selected we say that it is *materialized*.

Example 3 *According to the terminology we introduced above, the problem of view selection for approximate querying in datacubes (see Example 1) can be translated as follows. The queries that we want to be able to reconstruct are the objects, while the sets of aggregates needed by the reconstruction algorithm are the components.*

Similarly, in the profile-driven data management problem (see Example 2) the objects are the queries contained in the user profile, and the components are the data items required to answer the queries.

We can construct a bipartite graph $G(U, V, E)$, where U is the set of objects, V the set of components, and E the edge set of the graph. An edge exists between $u \in U$ and $v \in V$ if component v is required by object u. An example of the general form of the graph is shown in Figure 2.

Let $V[i], 1 \leq i \leq |V|$ be a binary vector having 1 in position i if and only if we have selected component v_i

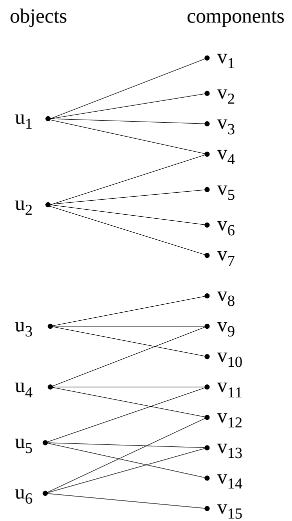


Figure 2. An example of the bipartite graph G.

for materialization, and $s(v_i), 1 \leq i \leq |V|$ be a function determining the space requirements of component v_i. Let $U[j], 1 \leq j \leq |U|$, be a binary vector having 1 in position j if and only if all the components required by object u_j have been materialized. Each object u_j is also associated with a value $b(u_j)$, which specifies its benefit, and is a measure of its importance. Given a constraint W on the total space available for the marginals, we are interested in maximizing the total benefit of queries answered, while satisfying the space constraint. Then, the *Constrained Set Selection (COSS)* optimization problem can be stated as follows.

Problem 1 *Maximize* $\sum_j U[j]b(u_j)$*, subject to* $\sum_i V[i]s(v_i) \leq W$.

It is easy to show that the Knapsack problem reduces to a special case of the *COSS* problem. Hence, according to the following lemma, the optimization problem is NP-Hard.

Lemma 1 *The* COSS *problem is NP-Hard.*

Observe that in the special case where all the queries have the same benefit, the problem is one of trying to satisfy the largest number of objects possible given the space constraint.

5. Algorithms for the COSS Problem

In the following paragraphs we propose several different algorithms for the solution of the *COSS* problem. We start by discussing optimal algorithms, and then present efficient greedy heuristics that can scale up to realistic sizes of the problem. We also explore the applicability of *simulated annealing*, a randomized algorithm, and *tabu search*, a meta-heuristic technique. The above methods have been used extensively in the past for solving a variety of hard problems, including scheduling, routing, and graph optimization [11, 10, 7]. Such algorithms have the ability to explore a larger area in the solution space than greedy algorithms, without getting stuck in local minima.

5.1. On Finding the Optimal Solution

There are two reasons we include an exhaustive algorithm in our discussion. First, it will demonstrate the dramatic difference in execution time between the optimal algorithm and the heuristics. The exhaustive algorithm has to test $O(2^{|U|})$ number of possible solutions, where $|U|$ is the number of objects. Second, and most important, it will serve as a basis for comparison of the quality of the results with the heuristic approaches. Unfortunately, this comparison will only be feasible for very small sizes of the problem.

Evidently, in order to find the optimal solution we do not need to enumerate and examine every single possible answer in the solution space. Solutions with sufficiently low cardinality (such that they are much lower than the space constraint) are bound to be sub-optimal, because we can always add new objects to the solution. Similarly, for answers that have already exceeded the space constraint we can safely prune all the solutions with higher cardinality. Finally, we can also prune the solutions that we know will never exceed the current best solution. These are the partial solutions that even if we fill up all the remaining space with the smallest objects, while assuming that these objects carry the maximum benefit, their total benefit will be smaller than the highest benefit we have seen so far. We will call the straight forward exhaustive algorithm *Naive,* and the one that prunes the search space *NaivePrune*.

Dynamic programming has been applied for the solution of the *Knapsack* problem, some instances of which can be solved in pseudo-polynomial time [5]. However, this technique cannot be applied in our case. The intuition behind this observation is the following. Assume that at some point we want to compute the optimal solution for a subset of the objects (or equivalently components) and for a fraction of the total space allowed. It turns out that there may be more than one optimal solutions at this point. Given that the current solution may affect the later choices of the algorithm we need to keep track of all of them, which leads to exponential time (or space). Thus, the dynamic programming algorithm is not possible to terminate in pseudo-polynomial time.

What makes the *COSS* problem difficult is the fact that in order to satisfy an object we need *all* the corresponding components. Materializing a subset of the components for some object does not credit us part of the object's benefit.

5.2. Solutions Based on Bond Energy

We now present algorithms based on the *Bond Energy Algorithm* [17]. This technique has been used for the vertical partitioning problem in databases [18]. A high level description of the algorithm we propose is shown in Figure 3. The algorithm starts by computing a measure of interrelation between each pair of components $v_i, v_j \subset V$. The interrelation measures for all possible pairs of components is captured in the square matrix A (line 5 of the algorithm):

$$A[i,j] = \sum_{u_k | u_k \in U, e_{ki} \in E, e_{kj} \in E} b(u_k).$$

This measure is a function of the number and the benefit of the objects that require both components, v_i and v_j. The

```
1 procedure SolveCOSS() {
2   let square matrix $A[\cdot] = 0$;
3   let $S = \emptyset$;                    /* selected components $S \subset V$ */
4   for i,j=1 to $|V|$
5     $A[ij] =$interrelation measure between components i and j;
6   $A[\cdot]$ = MakeBlockDiag($A[\cdot]$);
7   $S$ = Split($A[\cdot]$);
8   return($S$);
9 }

10 procedure MakeBlockDiag($A[\cdot]$) {
11   let matrix $A'[\cdot]$ be empty;
12   for i=1 to $|V| - 1$ {
13     select one of the remaining columns in $A[\cdot]$;
14     let $f_{max} = 0$;     /* maximum increase in bond energy so far */
15     for j=1 to i+1 {
16       place new column in $A'[\cdot]$ in position j;
17       $f_{max} = max($bond energy increase after placing new column
                          in position j, $f_{max})$;
18     }
19     keep the $A'[\cdot]$ corresponding to $f_{max}$;
20   }
21   return($A'[\cdot]$);
22 }

23 procedure Split($A[\cdot]$) {
24   let $f_{max} = 0$;      /* total benefit of best solution so far */
25   for i=1 to $|V|$ {
26     make first column of $A[\cdot]$ last, and first row last;
27     let j:=1;
28     let $S = \emptyset$;                    /* selected components $S \subset V$ */
29     while (j< $|V| \wedge$ components in $S$ satisfy the space constraint) {
30       $S$ = $S \cup \{$component $v$ corresponding to j-th column of $A[\cdot]\}$;
31       $f_{max} = max($total benefit of objects satisfied by components
                          in $S$, $f_{max})$;
32     }
33   }
34   return($S$ corresponding to $f_{max}$);
35 }
```

Figure 3. The *BondEn* algorithm.

larger the number of objects and their benefits, the stronger the connection between the pair of components is.

Then, the procedure $MakeBlockDiag()$ permutes the columns of A (or equivalently the rows since A is symmetric), and transforms it into a semiblock diagonal form, where large interrelation values tend to be grouped together. This transformation is expressed by the formula

$$\max\left(\sum_{k=1}^{|V|}\sum_{l=1}^{|V|} A[k,l](A[k,l-1]+A[k,l+1]+A[k-1,l]+A[k+1,l])\right),$$

which is the mathematical representation of the bond energy. We are seeking for the maximum value of this expression over all possible arrangements of the columns of matrix A. The algorithm proceeds greedily by considering a single column at a time. The above procedure is very fast, avoiding to examine the entire exponential search space, and is still able to find a near optimal form for A [17].

In the final step the algorithm splits the set of components V into S and $V - S$. Essentially, S is the set of components that can be used to satisfy the largest fraction of high-benefit objects while restricting the amount of available space. Just a single split point in A determines S and $V - S$. The loop in line 28 makes sure that we won't miss good solutions for S even if S was originally situated in the centre of A. We will refer to this algorithm as *BondEn (Bond Energy)*. Its time complexity is $O(|V|^3)$.

Note that in order to transform matrix A into a semiblock diagonal form, we add one column at a time in the matrix position that results in the largest increase to the overall bond energy. However, the selection of the column to add is arbitrary. Instead, we can enhance the technique by using a greedy column selection approach. That is, choosing in every step among all the available columns (i.e., the ones not already placed) the one that leads to the highest bond energy. We will call this version of the algorithm *BondEnGr (Bond Energy Greedy)*, and its time complexity now becomes $O(|V|^4)$. We also experimented with another two variations of the algorithm, where we make sure that during the split step we do not include in S any components required by objects that are not satisfied. We call the above variations *BondEn-SpAll* and *BondEnGr-SpAll*, which are the extensions of *BondEn* and *BondEnGr* respectively.

5.3. Solutions Based on Greedy Algorithms

The greedy algorithms provide a very fast alternative to solving the *COSS* problem, and are particularly appealing even for very large instances of the problem. The skeleton of the family of greedy algorithms we propose is depicted in Figure 4. They start with an empty solution set, and at each step they add to the solution set those components that satisfy the greedy condition. Before discussing the alternatives, we introduce some new notation that will be necessary for the mathematical formulation. Let S^k be the set of components that the algorithm has selected during the past k iterations, and O^k be the set of objects that are satisfied given the components in S^k. We are interested in deciding how to update those sets during iteration $k+1$. Let $CO^k = U - O^k$ be the set of candidate objects for selection during iteration $k+1$. Let $C_l = \{v_j | e_{lj} \in E \wedge v_j \notin S^k\}$ be the set of components, which are not in S^k, required for object $u_i \in CO^k$. The additional amount of space required by these components is given by the formula $f(u_l) = \sum_{v_j \in C_l} s(v_j)$. As before, we assume that we know the graph $G(U, V, E)$, where U is the set of objects, V is the set of components, and E is the set of edges indicating which components are needed in order to satisfy each object. The greedy step can take any of the following four forms.

```
procedure SolveCOSS() {
  let $O = \emptyset$;          /* selected objects $O \subset U$ */
  let $S = \emptyset$;          /* selected components $S \subset V$ */
  while ($|S| < |V| \wedge$
          components in $S$ satisfy the space constraint){
      among all the objects $\{u | u \in U \wedge u \notin O\}$ select $u_i$
      that satisfies the greedy condition;
      let $O = O \cup \{u_i\}$;
      let $S = S \cup \{v_j | e_{ij} \in E \wedge v_j \notin S\}$;
  }
  return($S$);
}
```

Figure 4. The greedy algorithm.

1. Accept to the solution set those components that will satisfy a new object, and that require the least amount of space. More formally, for each object $u_i \in CO^k$ calculate the additional space required for storing the corresponding components, given by $f(u_i)$. Let $u_l = \arg\min_{u_i} f(u_i)$ be the object that requires the least additional space. Then, $O^{k+1} = O^k \cup u_l$, and $S^{k+1} = S^k \cup C_l$.

This approach is trying to satisfy as many objects as possible. The intuition is that if there are many objects answered then the total benefit of those objects will be high. However, this may not be true if all the satisfied objects happen to have low benefit values. The time complexity of this algorithm is $O(|U|^2)$, and we will refer to it as *GrSp (Greedy Space)*.

2. Accept to the solution set those components that will satisfy the new object with the highest benefit attached to it. More formally, let $u_l = \arg\max_{u_i \in CO^k} b(u_i)$ be the object with the highest benefit among all the candidate objects. Then, $O^{k+1} = O^k \cup u_l$, and $S^{k+1} = S^k \cup C_l$.

The goal of this alternative is to answer as many of the high-benefit objects as possible. This approach is likely to fail if the required components have unexpectedly high space requirements. We will refer to this algorithm as *GrBen (Greedy Benefit)*. Its time complexity is $O(|U|^2)$.

3. Accept to the solution set those components that will satisfy a new object, and the ratio of the object benefit over the total space required by the selected components is minimal. More formally, for each object $u_i \in CO^k$ calculate the additional space required for storing the corresponding components, given by $f(u_i)$. Let $u_l = \arg\max_{u_i} \frac{b(u_i)}{f(u_i)}$ be the object that has the highest benefit per unit of additional space required by its components. Then, $O^{k+1} = O^k \cup u_l$, and $S^{k+1} = S^k \cup C_l$.

This variation of the algorithm is trying to account for the extreme cases we identified as weaknesses to the previous two alternatives. We will refer to it as *GrBenSp (Greedy Benefit per unit Space)*. Its time complexity is $O(|U|^2)$.

4. Accept to the solution at each iteration an individual component. Select the component that fits in the remaining space and yields the maximum B/s ratio, where B is the total benefit of all the objects that are now satisfied because of the selection of the new component, and s is the space requirements of the new component. If the selection of no component causes any new objects to be satisfied then the algorithm picks the component with the smallest space requirements.

This algorithm explores the applicability of choosing components instead of objects. We do not expect it to perform well when most of the objects require more than one component in order to get satisfied. The time complexity of this algorithm is $O(|V|^2)$ and we will refer to it as *GrComp (Greedy Component)*.

5.4. Simulated Annealing

Simulated annealing [14] is a randomized hill climbing algorithm. With a certain probability, that declines over time, this algorithm is allowed to follow directions in the solution space that result in solutions worse than those seen so far. This technique enables the algorithm to avoid local minima and stabilize in a final state that is close to optimal.

The simulated annealing algorithm that solves the *COSS* problem is depicted in Figure 5. We start with an initial solution (lines 2 and 3) that is provided by one of the greedy algorithms we presented earlier. Lines 7-18 implement the hill climbing procedure. In simulated annealing apart from uphill moves, downhill moves are also allowed under certain circumstances. More specifically, a downhill move is accepted with probability $e^{-(delta)/T}$ (line 13), where $delta$ is the decrease in the cost function from the previous iteration, and T is a time-varying parameter controlling the above probability. When T is high (in the beginning of the process) the probability of accepting downhill moves is high. Then, T is slowly decreased (line 17), and when the system freezes ($T < 1$) no further moves are considered. Note that this process involves two loops. While T remains fixed, the inner loop (lines 8-16) searches for solutions. In our implementation the number of iterations executed is a constant number (dependent on the problem size).

```
1 procedure SolveCOSS() {
2   let $O = O_0$;        /* selected objects $O \subset U$ */
3   let $S = S_0$;        /* selected components $S \subset V$,
                              corresponding to $O$ */
4   let $T = T_0$;
5   let $delta = 0$;
6   let $f_{max} = 0$;        /* total benefit of best solution so far */
7   while ($T > 1$){
8       for $i$=1 to total number of iterations{
9           $O_{new}$ =GetNewSimAnSolution($O$);
10          let $delta$ =(benefit of $O_{new}$)$-$ (benefit of $O$);
11          if($delta \geq 0$)
12            $O = O_{new}$;        /* accept the new solution */
13          if ($delta < 0$) then with probability $e^{-\frac{delta}{T}}$
14            $O = O_{new}$;        /* accept the new solution */
15          $f_{max} = max$(benefit of $O$, $f_{max}$);
16      }
17      $T = \alpha T$;        /* $\alpha < 1$ */
18  }
19  $S =$ marginals needed to answer queries in the $O$
          corresponding to $f_{max}$;
20  return($S$);
21 }

22 procedure GetNewSimAnSolution($O$) {
23   let $S =$ marginals needed to answer the queries in $O$;
24   pick object $O_{admit} \in \{U - O\}$ at random,
     and insert it to $O$;
25   if(total space needed by $S > W$)
26     pick an object $O_{evict} \in \{O - O_{admit}\}$ at random,
       and remove it from $O$;
27   if (total space needed by $S > W$)
28       penalize the total benefit of $O$ proportionally to
             the benefit per space ratio of $O_{admit}$
             and the amount by which $W$ is exceeded;
29   return($O$);
30 }
```

Figure 5. The *SimAn* algorithm.

The function *GetNewSimAnSolution()* (lines 22-28) determines what the proposed solution for the next iteration of the algorithm is going to be. The next solution is chosen at random among all the neighbours of the current solution. We term two solutions as neighbours if we can derive one from the other by adding a single object to one of the solutions, optionally followed by a deletion of another object. This last step ensures that the current solution satisfies the space constraint (lines 25-26). Nevertheless, the requirement that the current solution should satisfy the space constraint at every step is not strict. In some cases, when the new solution seems promising, we allow it to violate the space constraint at the cost of a small penalty to the total benefit.

The complexity of the simulated annealing approach is determined by the number of iterations. The work done in each iteration is really minimal, and we can safely consider it as constant. The number of iterations is controlled by the user, who sets the parameter T, and defines the way it is reduced. In our experiments the parameter T was initially set to four times the total benefit of the initial solution, the α parameter controlling the decrease of T was set to 0.95, and for each fixed value of T the inner loop of the algorithm executed a number of iterations equal to $1/3$ of the total number of objects. Varying the above parameters did not have significant effects on the quality of solutions found.

5.5. Tabu Search

Tabu search [7] is a metastrategy for guiding known heuristics to overcome local optimality. A structure called *tabu list*, used as auxiliary memory, describes a set of moves that are not permitted. This way the algorithm can avoid visiting solutions that has already visited in the past, and thus it is less probable to get stuck in local minima or cycles.

The outline of the tabu search algorithm for the *COSS* problem is shown in Figure 6. The algorithm starts with an initial solution (lines 2 and 3), which in our implementation is derived using the greedy heuristics we presented earlier. Then, the main loop (lines 6-11) iterates over the proposed solutions in order to pick the best one. The function *GetNewTabuSolution()* (lines 15-21) selects the solution (defined in terms of the selected objects) that will be examined during the next iteration of the algorithm. The last step of the function is to make sure that the proposed solution satisfies the space constraint (lines 18 and 19). However, as with simulated annealing, this requirement is not strict.

```
1 procedure SolveCOSS() {
2   let $O = O_0$;        /* selected objects $O \subset U$ */
3   let $S = S_0$;        /* selected components $S \subset V$,
                             corresponding to $O$ */
4   let $f_{max} = 0$;    /* total benefit of best solution so far */
5   let $O_{tabu} = \emptyset$;   /* set of tabu objects */
6   while (search not finished){
7       $O_{new}$ =GetNewTabuSolution($O, O_{tabu}$);
8       $f_{max} = max$(benefit of $O_{new}, f_{max}$);
9       $O = O_{new}$;
10      update $O_{tabu}$ based on the changes to $O$;
11  }
12  $S =$ components needed to satisfy objects in the $O$
          corresponding to $f_{max}$;
13  return($S$);
14 }

15 procedure GetNewTabuSolution($O, O_{tabu}$) {
16   let $S =$ components needed to satisfy objects in $O$;
17   pick an object $O_{admit} \in \{U - O - O_{tabu}\}$,
     and insert it to $O$;
18   if(total space needed by $S > W$)
19     pick an object $O_{evict} \in \{O - O_{admit} - O_{tabu}\}$,
       and remove it from $O$;
20   if (total space needed by $S > W$)
21       penalize the total benefit of $O$ proportionally to
             the benefit per space ratio of $O_{admit}$
             and the amount by which $W$ is exceeded;
22   return($O$);
23 }
```

Figure 6. The *Tabu* algorithm.

Note that the changes that lead to the new solution cannot involve any of the objects in the tabu list Q_{tabu}. In our implementation the tabu list keeps track of the recently added or deleted objects. Nevertheless, this restriction can be overridden when the new solution is the best obtained so far. When the new solution is available, we update the tabu list (line 10), and proceed to the next iteration.

Similar to the simulated annealing method, the complexity of tabu search is controlled by the user. An important difference in the case of tabu search is the selection of the next solution. It can be as simple as a random change in the current solution (like simulated annealing), or as involved as exhaustive enumeration. Therefore, there is a tradeoff between the number of iterations the algorithm will execute, and the complexity of each iteration. In our experiments we use a greedy algorithm to select the new solution.

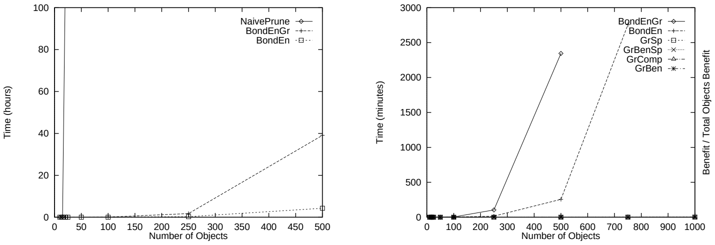
6. Experimental Evaluation

For the evaluation of the efficiency and behavior of the algorithms we use synthetic datasets, where the number of objects range from 10 to 1000. The number of components is in each case twice the number of objects, and more than half of the components help satisfy multiple objects. In order to assign benefit values to objects and space requirements to components we generated random numbers following uniform, Gaussian, and Zipfian distributions. The results for the datasets constructed using the Gaussian distribution are similar to those obtained from the uniform datasets, and we omit them for brevity.

In the experiments we measure the sum of the benefits of the objects that are satisfied given the components that are selected by the algorithms. We are also interested in the computation time of the proposed algorithms. In all cases we report the benefit of the solutions normalized by the total benefit of all the objects in the problem. Similarly, the space constraint is normalized by the total space requirements of all the components in the problem.

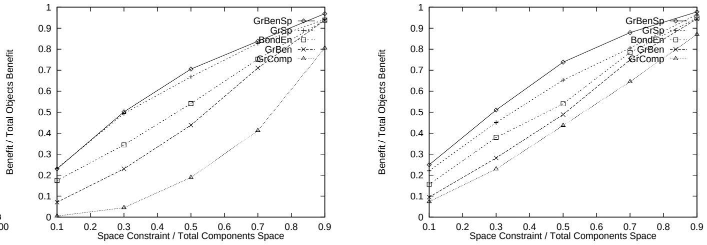
6.1. Scalability of the Algorithms

The first set of experiments examines the efficiency of the algorithms in terms of the time required to produce the solution. Figure 7 shows how the run-time of the algorithms changes when the number of objects increases. In Figure 7(a) we depict the tremendous difference in the computation time needed by the naive approach and the rest of the algorithms. *NaivePrune*, which is represented by the vertical line at the very left of the graph, is able to produce answers in a reasonable time-frame only for problems involving fewer than 20 objects. As shown in Figure 7(b) the bond energy algorithms



(a) *NaivePrune* and *BondEn* (b) *BondEn* and greedy

Figure 7. Scalability of the algorithms.



(a) uniform distribution (b) Zipfian distribution

Figure 9. Varying the space constraint.

scale more gracefully. Nevertheless, when the problem size becomes large, i.e., more than 600 objects for *BondEn* and more than 400 objects for *BondEnGr*, these algorithms require more than 24 hours to produce a solution. Evidently, it is only the greedy algorithms that are able to scale to thousands of objects. The time they required was under 2 minutes in all cases we tested.

6.2. Evaluating the Quality of the Solutions

In this set of experiments we evaluate the quality of the solutions produced by the proposed algorithms. First we compare the variations of the bond energy family of algorithms. Figure 8 illustrates the benefits for the solutions produced by *BondEn*, *BondEnGr*, *BondEn-SpAll*, and *BondEnGr-SpAll* for various values of the space constraint. The differences among the algorithms are minimal in all the cases we considered. Observe also that the greedy method of constructing the bond energy matrix (*BondEnGr*) in some cases improves the quality of the solutions. However, the more involved split procedure (represented by *BondEnGr-SpAll*) is not able to achieve a better solution than the simpler approach (*BondEnGr*). Since the above algorithms perform without significant differences, for the rest of the experiments we only demonstrate *BondEn*, which is the fastest among them.

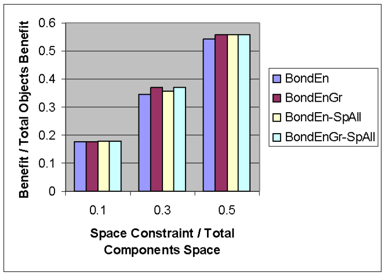


Figure 8. Performance of the *BondEn* **algorithms.**

In the next set of experiments we compare the quality of the solutions of the bond energy and the greedy algorithms. The graphs in Figure 9 illustrate the normalized total benefit of the solutions when we vary the space constraint, and the graph G remains the same. The two graphs correspond to the cases where the assigned object benefits and component space requirements follow uniform, and Zipfian distributions, respectively. Figure 10 depicts the relative ordering of the algorithms in terms of the quality of the solution, when we vary the graph G. As we move in the graph from left to right there is an increase in the average number of objects that are connected to each component (i.e., we increase the degrees of the component nodes in G). The best performance across all experiments is achieved by *GrBenSp*, closely followed by *GrSp*. The *BondEn* and *GrBen* algorithms perform in the middle range, while *GrComp* performs the worst.

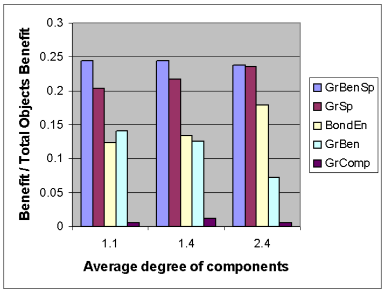
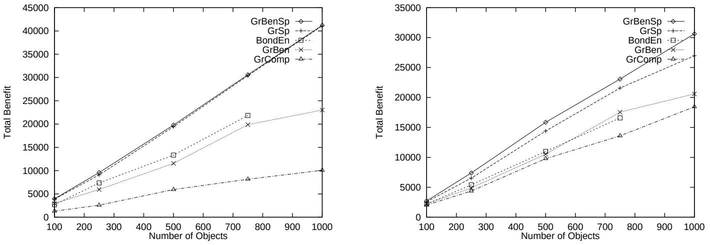


Figure 10. Varying the graph G**.**

The poor performance of *GrComp* is explained by the nature of the algorithm, which builds the solution one component at a time, instead of sets of components like the rest of the algorithms. This choice restricts *GrComp*, and does not allow it to start accumulating benefit until all the components related to a specific object are brought in the solution. This explains the slow rate at which the algorithm improves the solution at the lower left part of the graph in Figure 9(a).

We also conducted a series of experiments where we varied the number of objects. Figure 11 depicts the total benefit achieved by each algorithm. We report the results of running the algorithms on the same graph G, where the query benefits and the component space requirements were produced from uniform (Figure 11(a)), and Zipfian (Figure 11(b)) distributions. These experiments show that the relative ordering in performance for all the algorithms we consider remains the same across various problem sizes.

In Table 1 we report the results of the experiments with *SimAn* and *Tabu*. We use the solution provided by *GrBenSp* as the base solution, and report the improvement on this solution achieved by each one of the two algorithms. The initial solution for both *SimAn* and *Tabu* is provided by *GrBenSp*, and this is also the method used by *Tabu* to select solutions at each iteration. We experimented with varying the number of objects and the space constraint, and we allowed the algo-



(a) uniform distribution (b) Zipfian distribution

Figure 11. Varying the number of objects.

rithms to run an equal amount of time to the time required by *GrBenSp* to produce the solution.

		GrBenSp	SimAn	Tabu
objects	*constraint*			
100	*0.5*	1	**1.012**	**1.012**
250	*0.5*	1	1	**1.003**
500	*0.5*	1	**1.003**	**1.003**

(a) Varying the number of objects

		GrBenSp	SimAn	Tabu
objects	*constraint*			
250	*0.1*	1	1	**1.05**
250	*0.3*	1	1	**1.002**
250	*0.5*	1	1	**1.003**

(b) Varying the space constraint

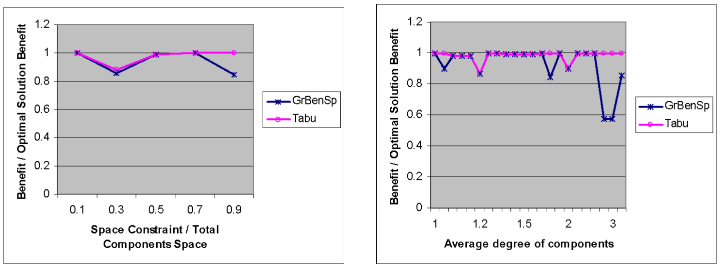
Table 1. Improvement in the *GrBenSp* solution by *SimAn* and *Tabu*.

Unlike *SimAn*, *Tabu* was able to improve on the initial solution in all the cases tested. We believe that the reason *Tabu* outperforms *SimAn* is because the solution space is too large, and this makes it extremely difficult for *SimAn* to head to the correct direction. Remember that *SimAn* moves in the solution space by selecting at random one of the numerous solutions that are neighbors of the current solution. On the other hand, the *Tabu* algorithm directs its search in the solution space more effectively, because it employs a more structured way of taking steps at each iteration.

6.2.1 Comparison to Optimal Solution

An interesting observation is the fact that both *SimAn* and *Tabu* improve the solution only by a small amount (less than or equal to 5% for the cases we tested). A natural question then is how close to optimal is the solution provided by *GrBenSp* in the first place. Unfortunately, it is not easy to find the optimal solution for the experiments we presented, because the size of the problem is prohibitively large. Therefore, we conducted a series of experiments where the number of objects was set to 10, for which we could get the optimal solution to compare against the other algorithms. The results of these experiments are depicted in Figure 12. The graphs show the benefit of the solutions produced by *GrBenSp* and *Tabu* normalized by the benefit of the optimal solution (which is always 1). In the graph shown in Figure 12(a) we vary the space constraint. In Figure 12(b) each point in the graph represents an experiment with a different graph *G* connecting objects to components. As we move in the graph from left to right there is an increase in the average number of objects that are connected to each component. The left-most point in the graph corresponds to the case where each component is connected to a single object, and the problem degenerates to the Knapsack problem.

Note that in many settings in both graphs *GrBenSp* finds a solution very close to optimal. Though, there are also cases where it achieves slightly more than half the benefit of the optimal. These experiments indicate that the *GrBenSp* algorithm is in many circumstances effective at finding near-optimal solutions. This explains the fact that *SimAn* and *Tabu* could not find much better solutions than the greedy algorithm in our previous experiments (see Table 1). Nevertheless, the graphs show that the *Tabu* algorithm is able to improve upon *GrBenSp*, and find a very good solution in almost all the cases where *GrBenSp* performs poorly.



(a) Varying the space constraint (b) Varying the graph *G*

Figure 12. *GrBenSp* and *Tabu* compared to optimal.

7. Discussion

The experiments demonstrate that the bond energy algorithms perform worse than some of the greedy approaches. At a first glance this is a rather surprising result, given that the *BondEn* algorithms have higher complexity and seem to make choices with greater care. However, we observe that all the decisions that *BondEn* makes are based on just *pairs* of components, despite the fact that in many cases it is a larger number of components that are interrelated. The experiments show that this prevents *BondEn* from capturing the true, more complex, associations inherent in the problem, narrows the information upon which the algorithm operates, and leads to poor decisions. Thus, *BondEn* should not be the algorithm of choice, especially since it does not scale as well as the greedy algorithms.

The greedy approaches are better than the bond energy algorithms for another reason as well. They are able to incrementally compute a new solution when the space constraint is relaxed. This is an important factor in the selection of the algorithm, since it removes the requirement of a hard space constraint, and gives the user the flexibility to choose among a range of slightly different space requirements which may lead to solutions with significant variations in the benefit values. On the contrary, the bond energy algorithms cannot readily provide the new solution, because they have to rerun the split procedure (see Figure 3), which is expensive.

The distinction among the greedy algorithms is quite clear between the best (*GrBenSp* and *GrSp*) and the worst (*GrBen* and *GrComp*) performers (see Figures 8(b) and 9(a)). However, it is not clear from the experiments whether *GrBenSp* or *GrSp* is a better choice, since they both have the same time complexity and provide solutions of similar quality. The following theorems[1] present theoretical results on the behaviour of the greedy algorithms when compared to optimal, that help answer the above question.

Theorem 1 *In the worst case, the* GrComp *algorithm provides a solution that is at least* $(\frac{W}{2} - 1)b$ *times worse than the optimal solution, for any* $b > 0$ *and* $W > 3$.

Theorem 2 *In the worst case, the* GrBen *algorithm provides a solution that is at least* $(W - 1)(1 - \frac{1}{b})$ *times worse than the optimal solution, for any* $b > 0$ *and* $W > 2$.

Theorem 3 *In the worst case, the* GrSp *algorithm provides a solution that is at least* $1.33b$ *times worse than the optimal solution, for any* $b > 0$ *and in the limit as* $W \to \infty$.

Theorem 4 *In the worst case, the* GrBenSp *algorithm provides a solution that is at least* 1.77 *times worse than the optimal solution, in the limit as* $b \to \infty$.

These theorems show that three of the greedy approaches we examined, namely, *GrComp*, *GrBen*, and *GrSp*, may perform arbitrarily poorly under certain circumstances. The same does not seem to be true for the *GrBenSp* algorithm. We believe that *GrBenSp* is a better choice in this sense, since it avoids making poor decisions that lead to solutions very far from optimal.

Finally, we should note that *Tabu* can in many cases improve on the solution of the greedy algorithms. An interesting observation is that when we turned off the tabu list functionality the performance of the algorithm deteriorated. This indicates that *Tabu* is indeed making well-calculated moves, and that the tabu list enables the algorithm to avoid local minima and explore new areas in the solution space. Therefore, it is beneficial to run this algorithm when the time allows it.

8. Conclusions

Space constrained optimization problems are still significant despite the advances in storage technologies. The tremendous amount of information produced every day, as well as the limited capacities of certain mobile devices, confirm the need for efficient solutions to such optimization problems.

In this paper we focus on the specific problem of *COSS*, where benefit values are associated to *sets* of items, instead of individual items. This optimization problem appears in important applications in various domains, such as data warehouses and pervasive computing. We derive the complexity of this problem, and propose several algorithms for its solution. We present an extensive experimental evaluation of the algorithms that illustrates the relative performance of the different approaches, and demonstrates the scalability of the greedy solutions. We also identify cases where some of the greedy algorithms perform poorly, and experimentally demonstrate that *GrBenSp* is a practical solution for the *COSS* problem.

Acknowledgements

We would like to thank Daniel Kobler for useful discussions regarding the tabu search method, and Ken Sevcik and Panayiotis Tsaparas for many insightful commnents.

References

[1] E. Baralis, S. Paraboschi, and E. Teniente. Materialized Views Selection in a Multidimensional Database. In *VLDB*, pages 156–165, Athens, Greece, Sept. 1997.

[2] M. Cherniack, M. J. Franklin, and S. B. Zdonik. Data Management for Pervasive Computing. In *VLDB*, page Tutorial, Rome, Italy, Sept. 2001.

[3] A. Deshpande, M. N. Garofalakis, and R. Rastogi. Independence is Good: Dependency-Based Histogram Synopses for High-Dimensional Data. In *ACM SIGMOD*, pages 199–210, Santa Barbara, CA, USA, May 2001.

[4] C. Faloutsos, H. V. Jagadish, and N. Sidiropoulos. Recovering Information from Summary Data. *VLDB, Athens, Greece*, pages 36–45, Aug. 1997.

[5] M. Garey and D. Johnson. *Computers and Intractability.* W. H. Freeman, 1979.

[6] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* New York: Academic Press, 1981.

[7] F. Glover and M. Laguna. *Tabu Search.* Kluwer Academic Publishers, 1997.

[8] H. Gupta. Selection of Views to Materialize in a Data Warehouse. In *ICDT International Conference*, pages 98–112, Delphi, Greece, Jan. 1997.

[9] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing Data Cubes Efficiently. In *ACM SIGMOD*, pages 205–216, Montreal, Canada, June 1996.

[10] Y. E. Ioannidis and E. Wong. Query Optimization by Simulated Annealing. In *ACM SIGMOD*, pages 9–22, San Francisco, CA, USA, May 1987.

[11] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon. Optimization by Simulated Annealing: An Experimental Evaluation. *Operations Research*, 37(6):865–892, 1989.

[12] H. J. Karloff and M. Mihail. On the Complexity of the View-Selection Problem. In *ACM PODS International Conference*, pages 167–173, Philadelphia, PA, USA, May 1999.

[13] R. Kimball. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley & Sons, 1996.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[15] J. Kleinberg. Personal Communication. 2002.

[16] J. M. Kleinberg, C. H. Papadimitriou, and P. Raghavan. Segmentation Problems. In *ACM Symposium on the Theory of Computing*, pages 473–482, Dallas, TX, USA, May 1998.

[17] W. T. McCormick, P. J. Schweitzer, and T. W. White. Problem Decomposition and Data Reorganization by a Clustering Technique. *Operations Research*, 20(5):993–1009, 1972.

[18] S. Navathe, S. Ceri, G. Wiederhold, and J. Dou. Vertical Partitioning Algorithms for Database Design. *ACM Transactions on Database Systems*, 9(4):680–710, 1984.

[19] S. B. Navathe and M. Ra. Vertical Partitioning for Database Design: A Graphical Algorithm. In *ACM SIGMOD*, pages 440–450, Portland, OR, USA, May 1989.

[20] D. Nehme and G. Yu. The Cardinality and Precedence Constrained Maximum Value Sub-Hypergraph Problem and its Applications. *Discrete Applied Mathematics*, 74:57–68, 1997.

[21] T. Palpanas and N. Koudas. Entropy Based Approximate Querying and Exploration of Datacubes. In *International Conference on Scientific and Statistical Database Management*, pages 81–90, Fairfax, VA, USA, July 2001.

[22] S. Rajagopalan and V. V. Vazirani. Primal-Dual RNC Approximation Algorithms for Set Cover and Covering Integer Programs. *SIAM Journal on Computing*, 28(2):525–540, 1998.

[23] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes. In *EDBT*, pages 168–182, Valencia, Spain, Mar. 1998.

[24] A. Shukla, P. Deshpande, J. F. Naughton, and K. Ramasamy. Storage Estimation for Multidimensional Aggregates in the Presence of Hierarchies. In *VLDB*, pages 522–531, Mumbai (Bombay), India, Sept. 1996.

[25] D. Theodoratos and T. K. Sellis. Data Warehouse Configuration. In *VLDB*, pages 126–135, Athens, Greece, Sept. 1997.

[26] J. Yang, K. Karlapalem, and Q. Li. Algorithms for Materialized View Design in Data Warehousing Environment. In *VLDB*, pages 136–145, Athens, Greece, Sept. 1997.

[1] The proofs of the theorems are available in the full version of this paper.