Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 5th, 2006

A HYBRID EVOLUTIONARY APPROACH TO SOLVING THE ARCHAEOLOGICAL SERIATION PROBLEM

Michael L. Gargano, mgargano@pace.edu
Computer Science, Pace University, NYC, NY 10038

Lorraine Lurie, llurie@pace.edu
Mathematics, Pace University, NYC, NY 10038

ABSTRACT

A general problem confronting archaeologists is the seriation (or sequence-dating) problem. Sir W.M. Flinders Petrie, a famous archeologist, was the first to formulate this interesting and important problem in mathematical terms almost one hundred years ago. The problem of trying to chronologically order grave sites in a cemetery where each grave contains (or does not contain) a number of different stylistic artifacts (e.g., pottery, jewelry, etc.) of a period is called the Petrie Seriation Problem. We have propose an order based hybrid evolutionary modified life cycle model to solve this problem that we believe has many advantages over previously proposed solution strategies.
Keywords: archaeological seriation, modified life cycle model hybrid

Introduction to the Problem

A general problem confronting archaeologists is the seriation (or sequence dating) problem. Sir W.M. Flinders Petrie, a famous archaeologist, was the first to formulate this interesting and important problem in mathematical terms almost one hundred years ago [11]. The problem of trying to chronologically order grave sites in a cemetery where each grave contains (or does not contain) a number of different stylistic artifacts (e.g., pottery, jewelry, etc.) of a period is called the Petrie Seriation (or Sequence-dating) Problem. Of course any archaeological site (other than a cemetery) containing different types of artifacts (or varieties) would do just as well.

The problem can be modeled mathematically by an incidence matrix whose rows are the grave sites and whose columns are a particular artifact type. A chronological ordering of the grave sites is established by finding a permutation of the rows of the incidence matrix which minimizes the total temporal range of an artifact type summed over all types of artifacts (i.e., columns). (The temporal range of an artifact type is the span from the first appearance to the last appearance of a one in the column of the incidence matrix corresponding to that artifact type). This search problem is characterized by a solution space which is generally unstructured (e.g., multimodal) and is intractable

for a large number of grave sites. Previously proposed solutions to this problem thus far have drawbacks. We have proposed an order based hybrid evolutionary algorithm, a biologically inspired heuristic search procedure, to solve this problem. We believe it has many advantages over previously proposed solution strategies. It primarily identifies an optimal, and/or near optimal, row-permutation in a computationally efficient manner for realistic situations typically encountered by archaeologists.

The life cycle paradigm [14] is an adaptive method evolving a population of potential solutions into a new, fitter, population of potential solutions using swarm methods, genetic algorithms, and local search. The process repeats itself until it reaches an optimal (or near optimal) solution.

MATHEMATICAL MODEL

An important mathematical data structure for analyzing the archaeological seriation problem is the incidence matrix. Each row of the incidence matrix represents a grave while each column represents an artifact type (or variety). The incidence matrix A containing g rows (representing g grave sites) and v columns (representing v different stylistic artifact types) is a table consisting of zeroes and ones. The entry in the i-th row and j-th column of A is denoted by A(i, j) and has a value of one if grave i contains any artifacts of type j and has a value of zero otherwise.

Ideally, if the graves (i.e., rows) were in the proper time sequential order, each artifact type (i.e., column) would suggest an ordinal interval (i.e., a time period) during which that artifact was popular and flourished. An important assumption that we will be making is that each artifact type is represented by that interval of time from when it first makes its appearance and begins to flourish to when it becomes unpopular, finally disappears, and thereafter, becomes forever obsolete (i.e., we will assume no archaistic tendencies). Therefore, we would like to permute the rows of the incidence matrix so that in each column the ones are packed tightly together. Ideally, each column would contain a sequence of zeroes (possibly none) followed by a sequence consisting solely of ones followed by another sequence of zeroes (possibly none). A matrix of this type is called a Petrie Matrix. A g by v incidence matrix A that attains such a form via an appropriate permutation of its rows is called a Petriezable matrix. Note that such a permutation may not be unique.

Since, it is not uncommon for archaeologists to be confronted with hundreds of graves containing hundreds of varied artifacts the number of possible permutations to consider may be astronomical. The seriation problem becomes intractable for the average archaeologist even with a small number of grave sites to consider. For example, if there are g grave sites then there are $P(g, g - 2)$ or $(g!/2)$ different row permutations to consider (since the outcome produced by a permutation and its reverse permutation produce basically the same ordering). To an archaeologist it is usually a simple task to identify the earlier from the later sites, that is, it is easy to distinguish at which end of the sequence to start (e.g., using carbon 14 dating if there seems to be a long time span). Note also that we are only expecting to retrieve information concerning the sequential order of the graves and are in no way trying to set an actual time scale. We are also, at first, making the assumption that each grave contains representatives of every variety of pottery and/or jewelry extant at the time of interment. In the real world situation this is normally not a good assumption, in that there are missing artifacts in some graves. Hence it is highly likely that there are missing ones in the incidence matrix A, making A non Petriezable.

Consider $G = AA^t$ whose (i, j)-th entry is equal to the number of varieties which are in common to graves i and j. G is called the Similarity Matrix for Graves and is g by g, symmetric, and non-negative. We say G is a Robinson Matrix[12] if its entries never decrease as one progresses along a row towards the main diagonal and if its entries never increase as one continues to progress along that row beyond the main diagonal. Since G is symmetric if G is a Robinson Matrix, this property is also true for each column of G. If A is a g by v incidence matrix then G contains a lot of information about the chronology of the graves that is hidden in A.

PRIOR STRATEGIES

Sir W. M. Flinders Petrie tried to find permutations which minimized the total temporal range summed over all the varieties. That is, he combined all the ranges of each of the columns into a range index and then he tried to minimize this index. (Clearly, the minimum value of the Petrie range index is the sum of the ones in the incidence matrix providing the incidence matrix is Petriezable). Since the number of possible permutations was large, Petrie used a combination of his expert experience, scholarly knowledge, and stylistic insight to find a reasonable, handcrafted solution. D.G. Kendall [9] proved a theorem which states that a permutation P which changes the grave matrix G into a Robinson Matrix (i.e., PGP^t) can also transform the incidence matrix A into a Petrie Matrix (i.e., PA). Therefore, if A is Petriezable then all the information needed to construct a sorting permutation of the type required to change A into Petrie form is contained in G. Of course, an exhaustive search for such a P is impractical and so a simpler heuristic procedure was developed using a multidimensional scaling (MDS) technique.

G can be viewed as a measure of how temporally alike the i-th grave is to the j-th grave. Although there is no temporal metric defined directly on the graves via G, we can construct a distance function D for pairs of graves by using a multidimensional scaling technique developed by R.N. Shepard and J.B. Kruskal[10]. Multidimensional scaling attempts to represent g objects geometrically by g points in k-dimensional real space, so that inter-point distances correspond to experimental dissimilarities between those objects (i.e., dissimilarities and distances are monotonically related). MDS represents each grave i as a point p(i) in k-dimensional real space in such a way so that

$$D[p(i), p(j)] < D[p(x), p(y)] \text{ iff } G(i, j) > G(x, y)$$

to as great an extent as possible. This method uses iterative perturbation techniques to find an answer by first computing an error term and then shifting the points by a small amount so as to reduce the previous error value. The method also requires the choice of an arbitrary initial position to start the iteration.

The process of finding a permutation which transforms the matrix A into a Petrie matrix is equivalent to finding a (0, 1) matrix having the consecutive ones property for columns[5] (i.e., in each column all the ones are consecutive). Booth and Lueker[1] proposed an efficient strategy using PQ-Trees for finding a permutation which transforms a (0, 1) matrix into a matrix that has the consecutive ones property providing the matrix is Petriezable.

A SIMPLE EXAMPLE

Here is an illustration of a very simple instance of this problem [9]. Consider the following incidence matrix A with g = 6 and v = 6. This incidence matrix can be row-permuted into Petrie form in exactly two ways.

Let
$$A = \begin{matrix} 1\,0\,0\,1\,1\,0 \\ 0\,0\,1\,0\,0\,1 \\ 1\,1\,1\,1\,0\,1 \\ 0\,0\,0\,0\,1\,0 \\ 1\,1\,1\,1\,1\,0 \\ 1\,0\,0\,0\,1\,0 \end{matrix}$$

and so
$$G = \begin{matrix} 3\,0\,2\,1\,3\,2 \\ 0\,2\,2\,0\,1\,0 \\ 2\,2\,5\,0\,4\,1 \\ 1\,0\,0\,1\,1\,1 \\ 3\,1\,4\,1\,5\,2 \\ 2\,0\,1\,1\,2\,2 \end{matrix}$$

Multidimensional scaling applied to this problem produces: p(l) = 0.28, p(2) = – 1.62, p(3) = – 0.73, p(4) = 1.50, p(5) = – 0.14, p(6) = 0.70 with the implied permutations: (2 3 5 1 6 4) and its reverse permutation (461532).

Applying this permutation to A we get:

$$PA = \begin{matrix} 0\,0\,1\,0\,0\,1 \\ 1\,1\,1\,1\,0\,1 \\ 1\,1\,1\,1\,1\,0 \\ 1\,0\,0\,1\,1\,0 \\ 1\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,1\,0 \end{matrix}$$

which is in Petrie form having a Petrie range index (discussed in section later) of 18 (i.e., the sum of the ones in A). Applying this permutation to G we get:

$$PGP^t = \begin{matrix} 2\,2\,1\,0\,0\,0 \\ 2\,5\,4\,2\,1\,0 \\ 1\,4\,5\,3\,2\,1 \\ 0\,2\,3\,3\,2\,1 \\ 0\,1\,2\,2\,2\,1 \\ 0\,0\,1\,1\,1\,1 \end{matrix}$$

which is in Robinson form.

DRAWBACKS: PRIOR STRATEGIES

The strategies proposed thus far have many drawbacks. The method originally proposed by Sir W.M. Flinders Petrie considers only a tiny fraction of all the possible permutations and is too seat-of the-pants. Trying to directly find a permutation that would put G into Robinson form is too intractable. The multidimensional scaling approach uses too many iterations, uses an arbitrary initial position to start the iteration, does not guarantee a good or even a reasonable solution, presents only one solution, does not handle constraints, does not have the ability to incorporate any expert heuristic information, has difficulty in handling graves which do not contain all varieties extant at the time of burial (i.e., matrix A is not Petriezable), may encounter scaling difficulties, and the method assumes some arbitrary structural elements. The PQ-Tree approach[1] cannot handle a non-Petriezable matrix (which is the realistic situation encountered by archaeologists), cannot incorporate constraints and heuristics, and incurs large runtime storage and cpu-time overhead. Although in a genetic algorithm was applied to this problem with very good results[7], a more recent paradigm (the hybrid evolutionary life cycle model) that incorporates a genetic algorithm is more desirable.

HYBRID EVOLUTIONARY METHODOLOGY

We propose an order based hybrid evolutionary algorithm for the archaeological seriation problem. The implementation is based on prior research done in other application areas [4,6,7,14]. Our method creates and evolves a population of potential solutions (i.e., sequences of row permutations of the incidence matrix) so as to facilitate the creation of new members by swarming, mating and mutating, or local search.

Fitness (or worth) is naturally scored by the Petrie range index of the permuted incidence matrix. In each column j note the first row r_f and last row r_1 containing a 1. The Petrie index for that column is $r_1 - r_f + 1$. The Petrie range index for the matrix is the sum of these over all columns. This will be the fitness of a permutation (i.e., a possible solution). The smaller the Petrie range index, the best the fitness (i.e., the better the solution).

This method is a hybrid combining swarm methods, genetic algorithms, and local search. An individual population member passes through three phases that are iterated until a satisfactory solution is obtained. As in many processes in nature, each individual member goes through different life cycle paradigms as it evolves. In this adaptive search heuristic a member goes from a swarm search to a genetic search to a local search and back again reiteratively.

MODIFIED LIFE CYCLE MODEL

We now state the modified hybrid life cycle algorithm that we used:

Randomly initialize a population of possible solutions in each paradigm.
For all swarm members

```
        Swarm
For all genetic algorithm members
        Mate and Mutate
For all local search members
        Search Locally
While (terminating condition not met)
        For each member
                Switch life cycle phase if no recent improvement
        For all swarm members
                Swarm
        For all genetic algorithm members
                Mate and Mutate
        For all local search members
                Search Locally
```

Since each phase has difference characteristics, this paradigm utilizes the best of each. Some of the differences are follows:

Swarm	**Genetic Algorithm**	**Local search**
Movement	Replacement	Movement
Directed Change	Random Change	Directed Change
Self-organizing	Natural Selection	Greedy
Cultural Transmission	Recombination	None
Neighborhood	Non-local	Neighborhood

SWARM PHASE

In this phase each swarm population member p_i swarm through permutation space. The global best of all the members is noted as member p_g. After being influenced by the global best p_g a particle (member) searches locally by considering all of its neighbors (i.e., permutations that results from a simple transposition of two adjacent positions in the permutation, also known as a simple 2 reversal). Each member then moves to a permutation neighbor of best fitness or stays in place.

Here is the algorithm:

```
Randomly initialize a swarm population of permutations in permutation space
Repeat
```

Find the global best p_g
For each member of the swarm
Mimic part of the global best p_g and move there
Consider all of the neighbors and move to a neighbor of best fitness or stay in place.

For example, if g = 6 and member p_g = 5 2 6 4 1 3 is best then if member p_i = 3 5 2 4 6 1 may first mimic member p_g by moving to 2 5 3 4 6 1 since 2 comes before 3 in p_g, then searching locally by considering all of its neighbors (i.e., permutations that results from a simple transposition of two adjacent positions in the permutation, also known as a simple 2 reversal). Thus, for example a neighbor would be 5 2 3 4 6 1 by reversing the first two positions.

If a member has no place to move or it has the same fitness for some time, then this member will be removed from this population and be put into the genetic algorithm population.

GENETIC ALGORITHM PHASE

The mating convention is such that only high scoring members (lower Petrie range index) will preserve and propagate their "worthy" characteristics from generation to generation and thereby help in continuing the search for an optimal solution. GA implementation requires a suitable encoding in chromosome space to evolve the chromosome members of the population and a mapping of the chromosome into permutation space to score the permutation members of the population. The permutation representation used is related to the Cantor Expansion of an integer [13] and the methodology which we use is similar to that used in the GA solution of the Traveling Salesperson Problem[2]. Each member of the population of potential solutions is encoded as a sequence of g genes in a chromosome where g is the number of gravesites. The value j of the gene in the k-th position of the chromosome (i.e., v_k) can range between 1 and k where chromosome position is measured from right to left starting at position 1. Thus, the chromosome can be represented by $v_g, v_{g-1}, \ldots, v_1$ where v_k=j ($1 < j < k$). This chromosome encoding is mapped into a g-th order permutation as follows: For the chromosome position index k ranging from g down to 1, place k at position i of the permutation (i.e., P_i). That is $P_i = k$ where $i = f(j)$. Here f is a counting function that determines i as the j-th unfilled position in the permutation counting from right-to-left. A simple example illustrates the above. If g = 6, a chromosome encoding obeying the range bounds for each position is: (452221) and its corresponding permutation is:(526341).

Applying this permutation to the rows of an incidence matrix A yields a transformed matrix from which the fitness can be scored by calculating its Petrie range index.

Selection of parents for mating involves choosing one chromosome member from the high scorers by a weighted "roulette wheel" approach and choosing the other chromosome member randomly. The reproductive process is a simple crossover operation where two selected parent members are cut into head and tail sections at some randomly chosen position and then have their tails swapped to create two offspring members. The crossover operation yields offspring chromosomes whose genes always satisfy the range bounds. A grim reaper mechanism replaces low scoring members in the population with newly created higher scoring offspring. Mutation is a GA mechanism where we randomly

choose a chromosome member of the population and change one randomly chosen gene of that chromosome. This process is useful in creating new areas of search to avoid getting caught on local minima of the solution space.

We can now state the genetic algorithm that we used:
Step 1: Randomly initialize a population of chromosomes.
Step 2: Map chromosome members to permutation members.
Step 3: Score any member that has not yet been evaluated.
Step 4: Sort the members of the population by their scores.
Step 5: Select parents for mating from the upper half of the population;
one using a "roulette wheel" approach and one randomly.
Step 6: Generate offspring using simple crossover.
Step 7: Mutate randomly selected members of the population at a randomly
selected gene at each generation.
Step 8: Replace the lower half of the population with offspring.
Step 9: If Petrie range index attains a minimum
Then return permutation
Else If time is up Then return best solution found
Else go to Step 2.

If a member has the same fitness for some time, then this member will be removed from this population and be put into the local search population.

LOCAL SEARCH PHASE

In this phase each member considers all of its neighbors (i.e., a permutation that results from a simple transposition of two adjacent positions of the permutation, also known as a simple 2 reversal). Each member then moves to a permutation neighbor of best fitness (if there are two or more choose one randomly). If a member has no place to move or it has the same fitness for some time, then this member will be removed from this population and be put back into the swarm population.

HYBRID EVOLUTIONARY METHOD ADVANTAGES

The advantages of this approach are:

1). it will usually produce an optimal and/or a near optimal solution(s).
2). convergence to the optimal permutation can be established when just one member of the population scores the minimum Petrie range index which is known apriori (i.e., the minimum Petrie range index possible = number of 1s in A).

3). its computational complexity is polynomial.
4). it can easily handle constraints and incorporate heuristics.
5). it can handle graves which do not contain all varities extant at the time of burial.
6). it uses the A matrix directly rather than the square symmetric matrix G.
7). it is faster and more effective than any of swarm, genetic algorithm, or local search individually

We believe that this approach most closely captures the flavor and the spirit of what Sir W.M. Flinders Petrie originally had in mind.

CONCLUSIONS

We considered using a hybrid evolutionary model based on a modified life cycle paradigm in order to solve the Petrie archaeological seriation problem that is NP hard. We then demonstrated that using this self-adapting algorithm that employs various different properties of the well known search strategies improves the algorithm's efficiency. In solving this NP hard problem, the algorithm employed different life cycle phases when appropriate adapted to its current search needs making more efficient.

ACKNOWLEDGEMENTS

We wish to thank Pace University's Seidenberg School of Computer Science and Information Systems for partially supporting this research.

REFERENCES

[1] Booth, K.S. and Lueker, G.S. 1976. Testing For The Consecutive Ones Property, Interval Graphs, and Graph Planarity Using PQ-Tree Algorithms. Journal of Computer and System Sciences 13: 335-379.

[2] Dewdney, A.K. 1988. The Armchair Universe - An Exploration of Computer Worlds, W. H. Freeman & Co. pg 252.

[3] Davis, L., 1991. Handbook of Genetic Algorithms, Van Nostrand Reinhold.

[4] Edelson, W. and Gargano, M. L. 1995. A Genetic Algorithm Approach To Optimizing Portfolio Merging Problems. in Proceedings of The Third International Conference on Artificial Intelligence Applications on Wall Street. pgs 168-173.

[5] Fulkerson, D.R. and Gross, O.A. 1965. Incidence matrices and interval graphs. Pacific Journal of Mathematics. 15:835-55.

[6] Gargano, M.L. and Rajpal, N. 1994. Using Genetic Algorithm Optimization To Evolve Popular Modern Abstract Art. Proceedings of the 1994 Long Island Conference on Artificial Intelligence and Computer Graphics. 38-52.

[7] Gargano, M. L. and Edelson, W. A Genetic Algorithm Approach To Solving The Archaeological Seriation Problem, Congressus Numerantium, 119 (1996), pp. 193-203.

[8] Goldberg, D., 1989. Genetic Algorithms, Addison Wesley Publishing Co.

[9] Kendall, D.G. 1969. Incidence matrices, interval graphs, and seriation in archaeololgy. Pacific Journal of Mathematics.

[10] Kruskal, J.B. 1964. Multidimensional scaling. Psychometrika. 29:1-42.

[11] Petrie, W.M. Flinders, 1899. Sequences in prehistoric remains. Journal of Anthrop. Inst. 29:295-301.

[12] Robinson, W.S. 1951. A method for chronologically ordering archaeological deposits. American Antiquity. 16:293-301.

[13] Rosen, K. H. 1991. Discrete Mathematics and Its Applications. Second Edition. McGraw-Hill, pgs 129, 286.

[14] Uran, B. and Gargano, M. L., Discovering Effective, Comprehensible Data Classification Rules Using Hybrid Evolutionary/Swarm Model for Data Mining, Congressus Numerantium, 176 (2005), pp. 49-63.