Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

Visualization Tools for Clustering, Trees and Ordered Trees

Stephen C. Hirtle
Thea Ghiselli-Crippa
Department of Information Science
752 LIS Building
University of Pittsburgh
Pittsburgh, PA 15260, USA
sch@lis.pitt.edu

Consider the situation of a set of individuals visiting a library and browsing through a collection. Each individual will, of course, select different books to browse depending on the original purpose of the visit. In addition, the order in which they search through a catalog will differ. The order is often ignored, yet might provide interesting cues as to the importance of certain books or topics. This scenario is not limited to browsing physical collections, but rather is repeated almost continually in other information-seeking behaviors including searching of on-line databases and hypertext/hypermedia systems.

The focus of this discussion is not on importance, *per se*, but rather on a second cue given by order. Specifically, the collection of ordered visits provides interesting data as to the structure of the collection, itself. To highlight this point consider a specific example in which three individuals search a collection. The first individual might examine documents on cognitive psychology and neuroscience. The second individual might examine documents on cognitive psychology and education. The third individual might examine documents on education, computer literacy, and grant writing. This might be reflected in the following simplified data, where the number indicates the document examined and the sequence indicates the order in which the documents were examined:

Individual A: 1,2,3,4,5,6,7,8
Individual B: 1,5,3,2,4,9,10,11,12,13
Individual C: 9,11,12,13,10,14,15,16

Even if we did not know the underlying structure, by comparing these three sets, we see empirical evidence to suggest that documents 1-5 form one set and 9-13 for another set. Furthermore, 6-8 and 14-16 might form two additional sets. We might also argue that items 1 and 9 are important items, as they were encountered first by all individuals examining that set. In this paper we examine the structure of sets of documents, or items, as revealed by repeated, ordered traversals through a space. In general, the actual data that we will discuss comes from the work in cognitive science on spatial memory and spatial problem solving, but the techniques have wider applicability. The primary point of this paper will be to suggest new graphical methods for the presentation of ordered data. As a result, clusters and other complex structures in the document space will be uncovered through a combination of scaling techniques and data visualization.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

ORDERED TREE ALGORITHM

We begin by considering the ordered tree clustering algorithm, which was discussed in detail by the first author in the 1st ASIS SIG/CR Classification Research Workshop (Hirtle, 1991a). The algorithm was developed by Reitman and Rueter (1980) to account for free-recall data. The algorithm is based on examining the order of items in repeated free-recalls. The data for the algorithm must be a set of linear orders, typically recall-orders, over a fixed set of items, such as repeatedly recalling all 50 U.S. states. In previous studies the number of items to be recalled has varied from 10 to 50, and the number of recalls has varied from 4 to 30. The algorithm proceeds by parsing the orders, in a top-down manner, to extract chunks of items recalled contiguously. The set of chunks can then be written as tree. The ordered tree algorithm is implemented in a program called TIGER, written in RATFOR by Henry Rueter, and is available in an executable version for an IBM PC from the authors.

The resulting representation, which reflects the underlying mental representation of the imbedded concepts, is an ordered tree, where the nodes at any level of the tree can be either ordered, as a unidirectional or bidirectional node, or unordered, as a nondirectional node. The alphabet would be an example of a strictly ordered set of items (indicating unidirectionality), whereas the numbers 1 through 10 (easily recalled in either a forward or backward order) might be taken as an example of a bi-ordered set of items (indicating bidirectionality). Despite many common examples of ordered sets of information, most classification schemes can not account for such regularity, but instead impose only hierarchical constraints on the data. Previous applications of ordered tree analysis include representing the expert knowledge of computer programmers and representing spatial knowledge (McKeithen et al., 1981; McNamara et al., 1989).

VIZUALIZATION OF CLUSTERS

Most recently, we have been concerned with the problem of visualizing clusters, particularly when there is an underlying spatial representation to the points being clustered. To be concrete, we will discuss two collections of data. In the first experiment, subjects were asked to recall one of three fixed sets of cities in the United States, repeatedly. In the second experiment, subjects were asked to solve a variant of the traveling salesman problem, in which the start and end locations were fixed. Even though these experiments are quite different in terms of memory load, subject demands, and so on, the techniques discussed proved useful in both cases. We will discuss each dataset in turn.

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

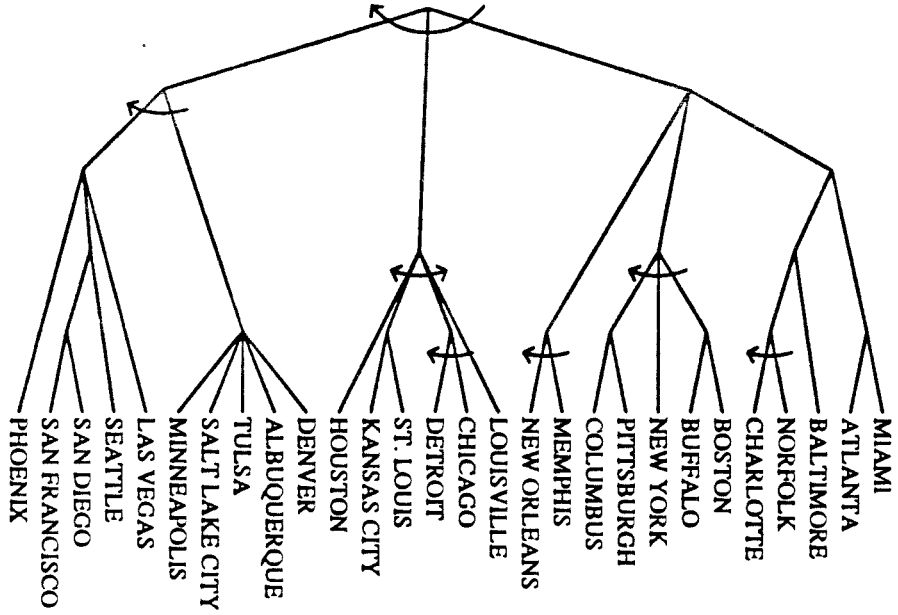

Figure 1. Ordered tree for subject 7.

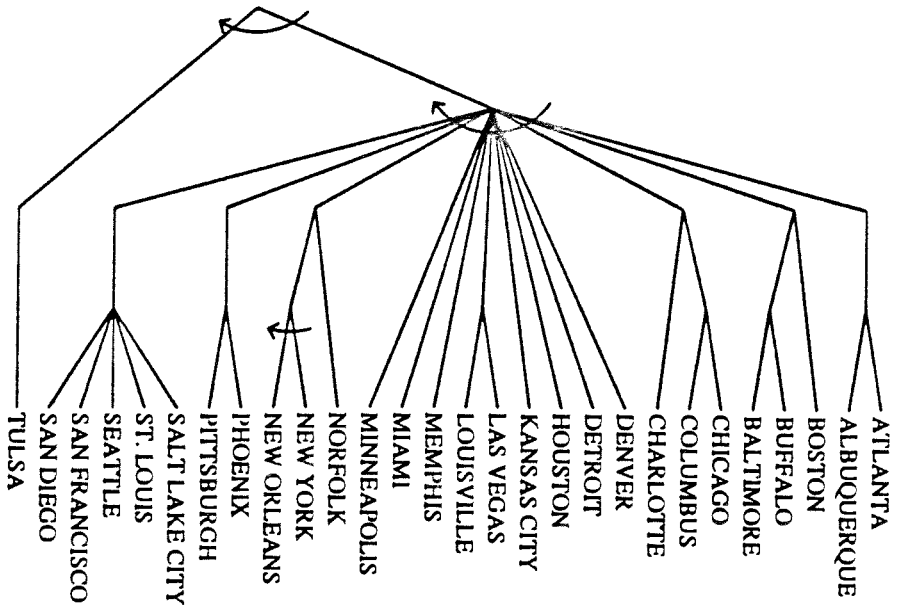

Figure 2. Ordered tree for subject 10.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

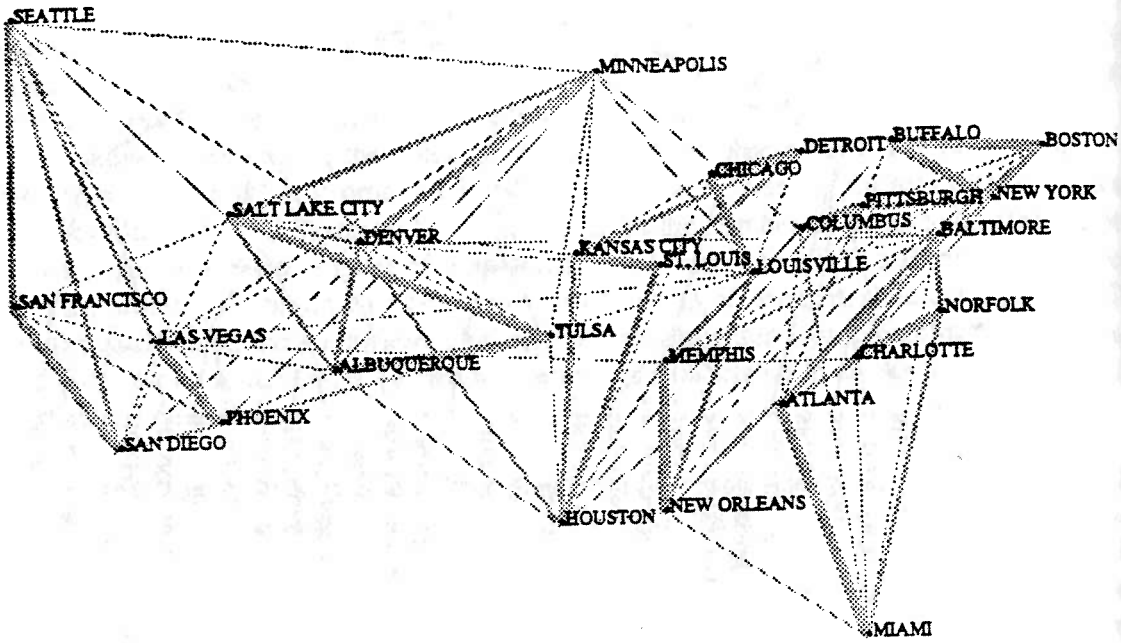

Figure 3. Map path graph for subject 7 showing evidence of spatial clustering.

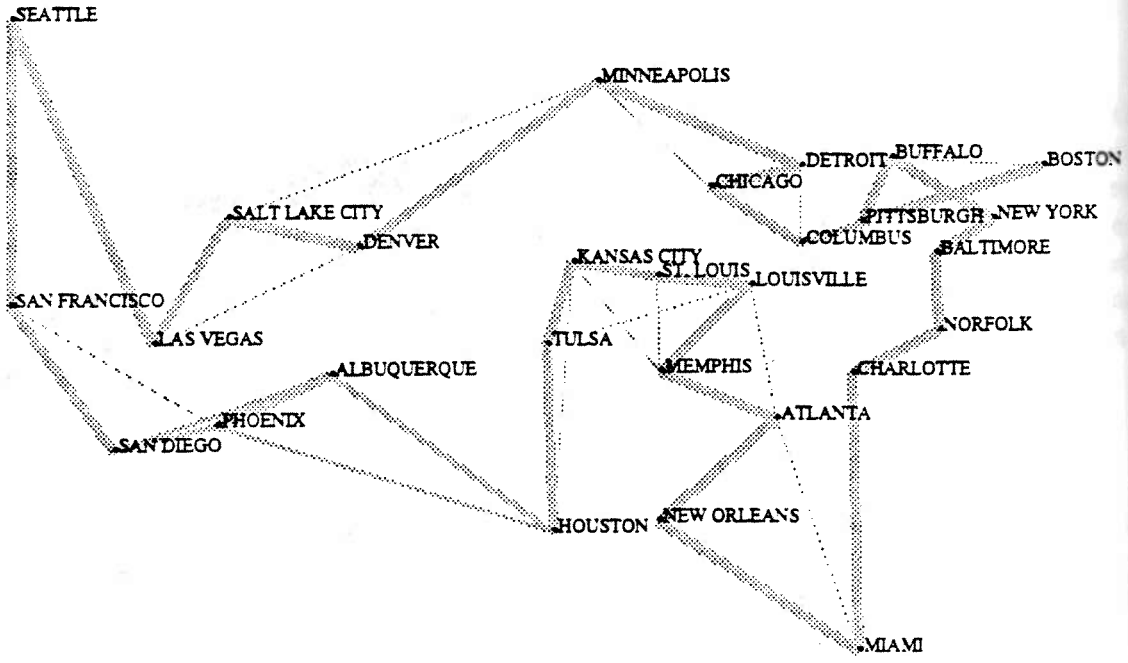

Figure 4. Map path graph for subject 25 showing a unidirectional recall pattern based on spatial proximity.

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

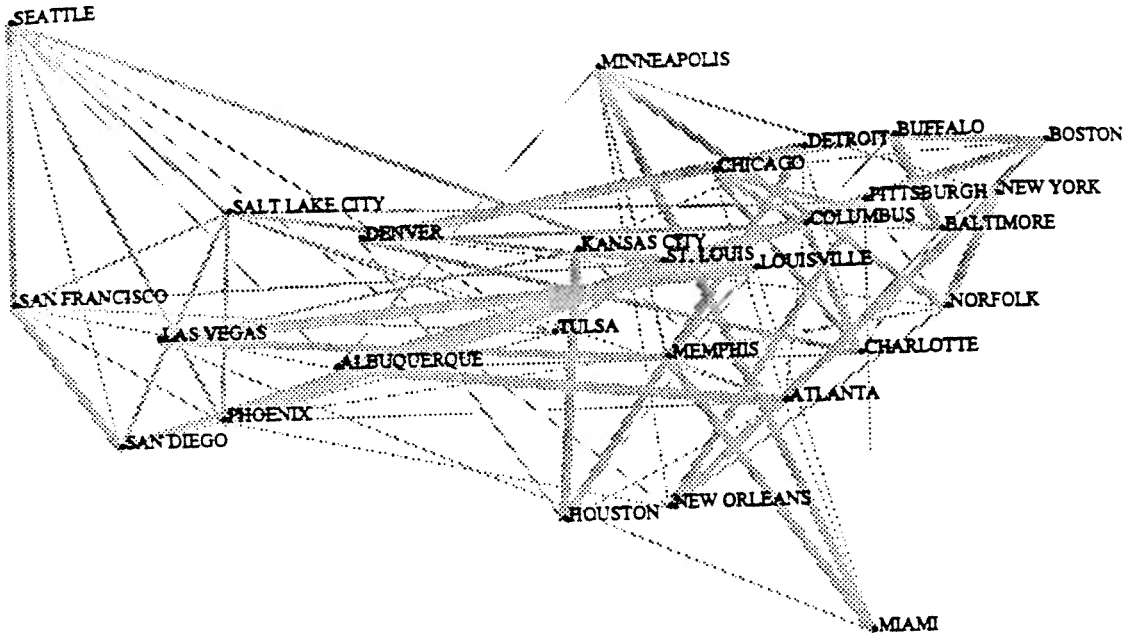

Figure 5. Map path graph for subject 10 showing no evidence of spatial clusters.

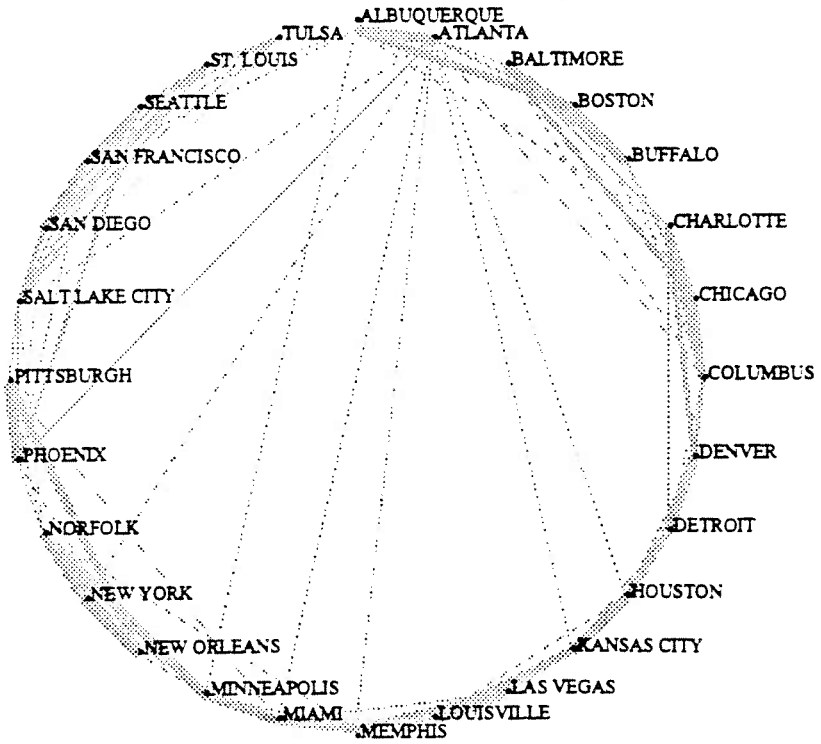

Figure 6. Alphabetic circle graph for subject 10 showing evidence of a first-letter alphabetic mnemonic.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

FREE RECALL

The first set of data come from a memory experiment where subjects were asked to recall one of three sets of U.S. cities (Hirtle, 1991b). Each set consisted of 28 city names. The sets varied in terms of homogeneity of both prominence and spatial location. 48 subjects participated in the study and each subject recalled the city names, 20 times, after memorizing the set of names. The ordered tree analysis produced one ordered tree for each subject. Two common strategies emerged from this analysis: 27 subjects used an alphabetical strategy, whereas 15 subjects used a spatial strategy. The remaining 6 subjects used either a combination of these two strategies, or a separate, idiosyncratic strategy. Two ordered trees highlighting a spatial and alphabetical strategy are shown in Figure 1 and 2, respectively.

Of course, it is somewhat difficult to infer the exact spatial nature of the clusters in Figure 1. An alternative representation which captures the spatial distribution might be more informative. As one approach, we considered the possibility of constructing a path graph, were the width of the paths between two cities reflects the number of trials in which two locations were recalled contiguously. Figure 3 is the path graph for the ordered tree in Figure 1. There is clear clustering of certain locations, such as Seattle, San Francisco, and San Diego. Figure 4 shows the graph for another subject with a strong, unidirectional strategy. That is, the recall strategy was not based on clusters, but rather on chaining the 28 locations together in a string.

If a subject used an alphabetical recall strategy, then the path graph based on city location shows up as noise, as seen in Figure 5. However, if an alternative underlying distribution based on an alphabetical pattern is used, such as shown in Figure 6, the path graph once again highlights the clusters. Careful inspection of Figure 6 shows the use of a first-letter mnemonic, which is shown by the cords connecting the cities with same first letter.

SPACIAL CHOICES

As a second application, we consider an experiment in which subjects were asked to solve a variant of the traveling salesman problem (Hirtle & Gärling, 1992; Hirtle, Gärling, & Ghiselli-Crippa, 1992). For each of six maps, 24 subjects generated 4 solutions, which may or may not be identical. Two of the six maps contained 6 points, two contained 10 points and two contained 18 points. The purpose of the study was to uncover heuristics used by subjects to solve these problems.

Order is particularly critical in this experiment, so we modified the path graphs to include order information. In this version, an arrow indicates the subjects consistently moved from one location to the next, whereas a solid line indicates each direction occurred in some solution to the problem. Overall, there was no single dominant strategy. Figure 7 shows the directed path graph for one of the 10 point maps.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

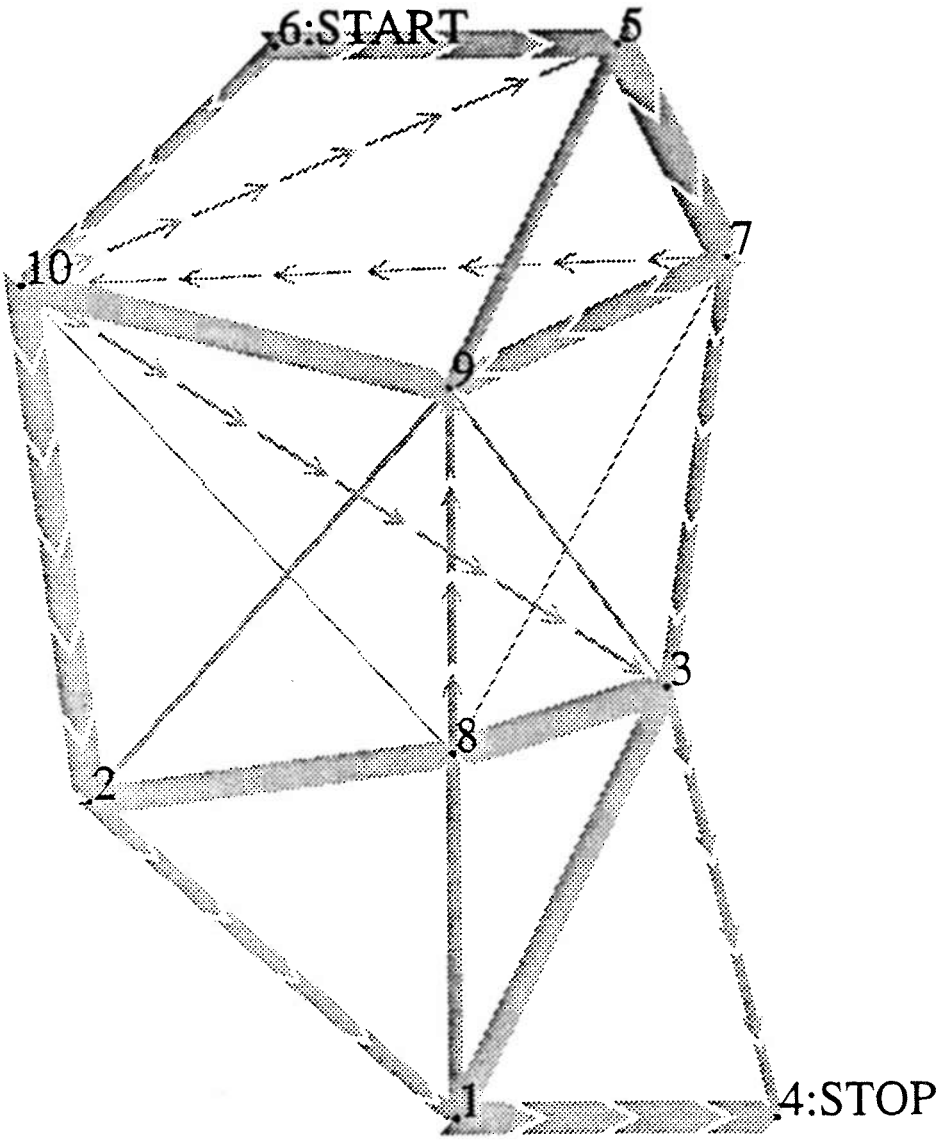

Figure 7. Ordered path graph for Map 10a.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

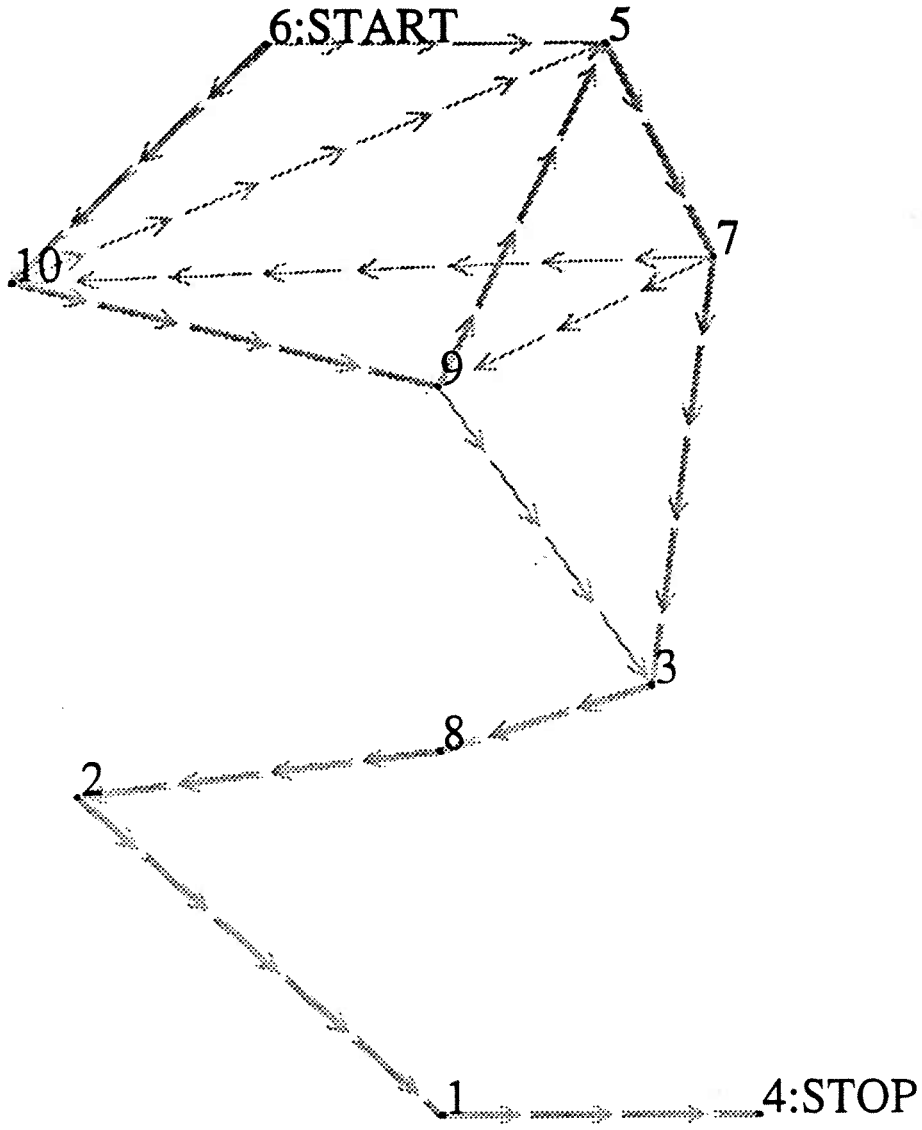

Figure 8. Ordered path graph showing the cluster heuristic on Map 10a.

ISSN: 2324-977

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

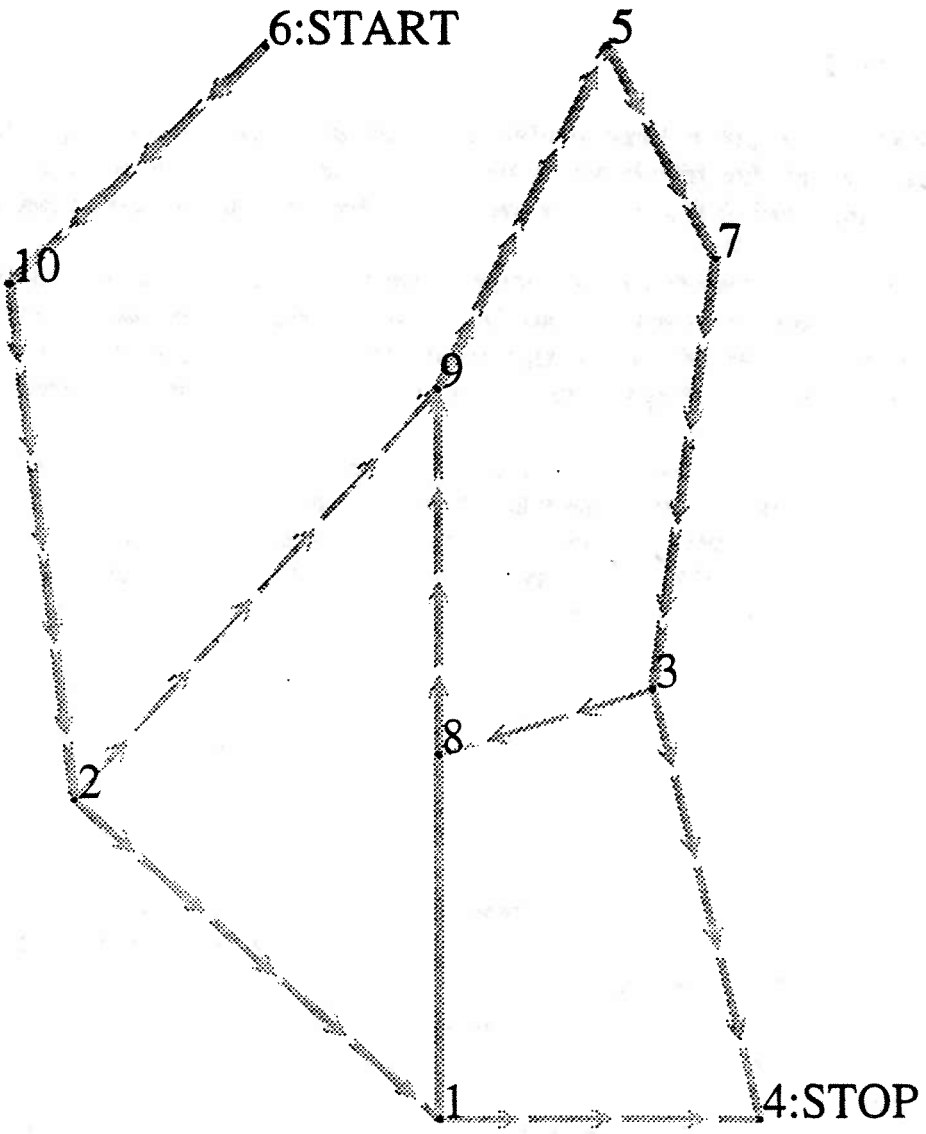

Figure 9. Ordered path graph showing the zig-zag heuristic on Map 10a.

ISSN: 2324-9773

Hirtle, S., & Ghiselli-Crippa, T. (1992). Visualization Tools for Clustering, Trees and Ordered Trees. 3rd ASIS SIG/CR Classification Research Workshop, 37-46. doi:10.7152/acro.v3i1.12594

However, we were able to extract different strategies used in solving the problem. Using a technique based on lattice theory that is described in detail elsewhere (Hirtle & Gärling, 1992), we were able to subdivide the list of all possible paths into distinct sub-strategies. For example, one common strategy was to treat locations {5, 6, 7, 9, 10} as a cluster, and to visit those locations in some order, then complete the path by visiting locations 3, 8, 2, 1, and 4, in that order. This is most clearly shown in Figure 8. Figure 9 highlights a quite different substrategy which consists of parsing the space vertically.

SUMMARY

The use of path graphs, with or without directionality noted, can highlight the structure of ordered data. There are few restrictions on the use of the technique, as opposed to the ordered tree algorithm, which is rather strict in requiring complete orders across a fixed set of items.

However, in contrast to the ordered tree algorithm, the use of path graphs does require an underlying space to map the points. In the cases presented, there was a clear a priori choice, either the fixed spatial location or the alphabetic circle of points. An alternative might be to use multidimensional scaling to give the coordinates, as is done with PathFinder (Schvaneveldt, 1990).

There is still much to work to be done. One interesting question is that of size. Trees become difficult to inspect for structure when the number of leaves grow. Even with as few 50 terminal items, it can be hard to interpret a tree visually. In contrast, the path graphs are a visualization technique that should be able to accommodate, with proper scaling, a much larger set of data, where the exact limits are yet to be determined.

REFERENCES

Hirtle, S. C. (1991a). Ordered trees: A structure for the mental representation of information. In *Advances in Classification Research*, ASIS monograph series. Medford, NJ: Learned Information.

Hirtle, S. C. (1991b). Knowledge representations of spatial relations. In Doignon, J.-P., & Falmagne, J.-C. (Eds). *Mathematical psychology: Current developments*, (pp. 233-249). New York: Springer-Verlag.

Hirtle, S. C., & Gärling, T. (1992). Heuristics rules for sequential spatial decisions. *Geoforum*, 23, 227-238.

Hirtle, S. C., Gärling, T., & Ghiselli-Crippa, T. (1992). The similarity of paths in a traveling salesman problem. Paper presented at *Distancia '92*, Rennes, France, June 22-26.

McKeithen, K. B., Reitman, J. S., Rueter, H. R., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, 13, 307-325.

Reitman, J. S., & Rueter, H. R. (1980). Organization revealed by recall orders and confirmed by pauses. *Cognitive Psychology*, 12, 554-581.

Schvaneveldt, R. W. (1990). *Pathfinder associate networks: Studies in knowledge organizations*. Norwood, NJ: Ablex.

ISSN: 2324-9773