Student Expectations for Library Web Site Organization

Charity B. Hope, San Jose State University
Shannon M. Staley, San Jose State University

Available at: https://works.bepress.com/shannon_staley/8/
1. We submitted IRB documentation and received approval.

2. We came up with a list of items to create cards for based on major pages and sections of the library web site. We then narrowed the list to a reasonable number that could be sorted within one hour.

3. We developed a recruitment tool and scripted instructions to read to each participant during testing.

4. We developed short questionnaire to capture information regarding web use, library use, and library instruction.

5. We created card piles using 3 X 5 index cards and computer generated labels, one on each card, representing each item to be sorted.

6. We recruited and tested 201 participants from February 2002 – June 2002. We approached students randomly, using the recruitment tool. We did not accept any student participants who worked in the library. Originally, we began by scheduling appointments and testing at a later time, but had better luck recruiting and testing on the fly. We tempted students with free candy. Testing sessions ran about 45 minutes to an hour. We gave a thank you gift after testing (library mug with treats).

7. We had participants sort cards manually. We wanted to make sure they had plenty of room to spread out. We also didn’t want them to have to learn a software program to do the exercise. We wanted to represent all participants, rather than computer savvy ones.

8. In testing session, we had them sign a release form, then asked them to sort cards into groupings that made sense to them. We asked them to set aside cards they were unsure of (these were omitted from the results pile) and asked them to sort each card only once (cards were not repeated in more than one grouping). We then asked them to label the groupings with a post it. Once they did this, we asked them if any of the groupings could be merged together to form higher level groupings. If so, we asked them to label these new groupings. Afterwards, we stapled all the groupings together, being careful to preserve the original, lower level groupings within the larger ones. We then administered a questionnaire and gave them a thank you gift.

9. After the card sorting exercises, we entered results into a card sorting software program, EZCalc, keeping track of participants by accession number (not name). We kept a separate spreadsheet that logged each participant number along with information we gathered from the recruitment questionnaire: sex, age, level in school, ethnic identity, SJSU major, web use, library use, library instruction.
Some Aspects of Hierarchical Clustering as Demonstrated by the EZCalc Program

Distance

The first step in any cluster analysis is determining the distances between all pairs of the entities being clustered. In your case this is the 59 x 59 matrix that is displayed by clicking the Matrix button. The values in this matrix are all between 0 and 1. The formula is not given anywhere, but almost certainly, the distance between concepts $i$ and $j$ is computed as:

$$d_{ij} = 1 - \left( \frac{n_{ij}}{N} \right)$$

where $n_{ij}$ is the number of participants that put concepts $i$ and $j$ in the same cluster and $N$ is the total number of participants. If nobody put the two concepts or items in the same group, then $n_{ij} = 0$ and the distance $d_{ij}$ is 1. If everybody put items $i$ and $j$ in the same group, then $n_{ij} = N$ and $d_{ij} = 0$. You should be able to verify this by looking at the raw data for few pairs of items (select pairs where $0 < d_{ij} < 1$).

Single Linkage Clustering

The simplest form of a hierarchical clustering is the single linkage hierarchical clustering. The Single button in the program produces this. The results are displayed as a tree diagram. The scale across the top (and bottom) goes from 0 to 1. These are the distances between pairs of items. To understand what is going on, run the single linkage clustering on the female.rec file that you sent to me.

The cluster process begins by joining the two (or more) concepts that are the closest in distance. You will see a few pairs such as Beethoven Center and Steinbeck Center which are connected at the lowest level. In fact the vertical line connecting them is at the zero point of the distance scale. If you look in the distance matrix where row 39 (Beethoven Center) intersects column 57 (Steinbeck Center), you will see that the distance between the two is exactly zero meaning all 10 participants put these two items in the same cluster.

After all clusters are formed where the distance between items in the cluster is zero, the program looks for the item closest to each cluster. In the case of the Beethoven-Steinbeck cluster it is the Multicultural Resource Center item that is 0.2 in distance from each of these items so it is joined to these two with the vertical connector showing the 0.2 distance. Even if Multi-Cultural had been 0.2 from Beethoven and 0.3 from Steinbeck, it would have been jointed to the Beethoven-Steinbeck cluster because that cluster has the item closest to Multi-Cultural. The process continues until eventually each item is jointed to the existing cluster to which it is closest and all items have been connected at some level.

The green and red “criterion” bars in the tree display are used to define two levels of closeness. The red bar (on the right) might be thought of as defining the degree of closeness required for certain web pages to be thought of as a single unit. The green bar (on the left) might be thought of as defining the degree of closeness required for items to belong on the same web page. Again,
these are just two levels of closeness for which the program gives you visual aids in seeing the results. There is nothing about the clustering process that restricts you to two levels – or to what each level means. That is up to you.

To see how the criterion bars might work, suppose we set the green bar at 0.14 and the right bar at 0.47. I’m picking arbitrary numbers that don’t coincide with any actual distances. This is just an example and in no way meant to be a “meaningful” solution.

If we use the criterion of 0.47 to divide items into groupings of web pages, then we would have 19 subgroups. The 19 comes from counting the number of alternating color sections. Note that all items within each section will be no greater than 0.47 apart from each other.

If we decided that all items on a single web page had to be within 0.14 in distance to each other, then we could use the green bar to see what items would go together and to determine the number of pages in each section. The first 13 sections identified by the right criterion bar would have one web page each. Section 14 would have one devoted to the Multicultural Resource Center and one devoted to the combined Beethoven – Steinbeck centers. Section 15 would have 4 pages, Section 16 would have 14 pages, Section 17 only one page, Section 18, two pages and Section 17 would have 17 pages. The number of pages comes from counting the number of alternating color patterns within each section.

The single linkage cluster algorithm has been criticized for producing unbalanced clusters as the number of clusters goes down. In our example, set the right criterion bar to about 0.62. This produces the fewest number of clusters (other than the trivial case when everything is put in the same cluster) which in this case is three. The top cluster has a single item, the second cluster has two items, leaving 53 items in the bottom cluster. This is a clear example of this problem.

**Complete Linkage Clustering**

With single linkage clustering an item could be connected to a cluster if there was one item in the existing cluster that was the closest to the item under consideration. For example, suppose the process starts by putting items A and B together. Now when the program looks at item C it examines the distances between item C and all other items. Suppose the smallest distance is 0.2 for the distance between A and C ($d_{AC}$). Then item C is attached to the AB cluster even though the distance between C and B might be 0.5.

With complete linkage clustering item C would be attached to the AB cluster only if both the distances $d_{AC}$ and $d_{BC}$ were 0.2 (and 0.2 was the smallest distance from C to anything). According to Aldenderfer and Blashfield (1984) “complete linkage has a tendency to find relatively compact, hyperspherical clusters composed of highly similar cases.” Using the example data set with the same 0.14 and 0.47 criteria we get 37 sections with no more than three pages per section and many with one page per section. The clusters do appear more balanced in terms of items per cluster as you move the right criterion to generate fewer high level clusters.
Under single linkage clustering, item C is connected to cluster AB if one of A or B is the closest item to C. Under complete linkage we require both A and B to be the closest to C. Average linkage is a compromise between single and complete that connects C to the cluster whose average item distance to C is the smallest. For example, if AB is an existing cluster and DEF is the only other cluster, the program would compare \((d_{AC} + d_{BC}) / 2\) to \((d_{DC} + d_{EC} + d_{FC}) / 3\) and choose the cluster with the smaller average distance.

All three clustering methods are examples of hierarchical agglomerative methods which are discussed in the Aldenderfer and Blashfield reference (also cited in the EZCalc description) on pages 35-44.

**Strategy**

Cluster analysis is clearly more art than science. The success depends on the number and interpretability of the final clusters. A lot of trial and error is involved as there is no real basis to eliminate any of the three possible solutions and each solution involves many possible ways to define the levels.

The best approach to your task would be as follows:

1. Analyze the entire sample of participants.

2. Decide ahead of time on a reasonable number of “Sections” or high level groupings (call this \(N_H\)) and reasonable number of web pages or low level groupings (\(N_L\)). Clearly, you must have \(1 < N_H < N_L < 59\). These are not hard and fast numbers, but approximate values.

3. With each clustering method adjust the distance criteria to produce something close to \(N_L\) and \(N_H\). Print out the best couple of solutions with each method. It looks like a color printer than can use legal size paper would be ideal.

4. Get feedback from other library science professionals. See if you can get a consensus on the one best solution.

5. Compare the best solution to that generated using the same method and approximately the same criteria but restricted to a particular demographic subgroup. Hopefully, differences will be small. In the event of large differences you will have to decide if some groups interpreted your concepts differently than others, in which case the whole cluster analysis approach is in doubt.

**Reference**