Citation Cleaning Procedure

Top citations were extracted from WOS database with the following format: first author's last name and initials, year of publication, journal name or book name, volume, page numbers, and DOI.

Duplicate citations were merged using an R script, which clustered citations based on a measure of similarity, and then merged them. The similarity was calculated for each article name, after blocking by author name and year, as well as type of document (book or journal).

The measure of similarity was calculated using Jaro-Winkler method. This method is best suited for short strings, and is a type of edit distance that uses matching characters and the number of transpositions to quantify similarity. When looking at Jaro-Winkler distance, the lower the distance, the more similar the two strings. Jaro-Winkler similarity is equal to 1 - Jaro-Winkler distance. When looking at similarity, 1 is an exact match, and 0 is no match.

Initial cleaning

First, the DOI was removed from the citations because it was not consistently reported. Eliminating it increased the chance of matching based on similarity of text.

Second, punctuation other than commas was removed, as well as instances of "in press", and instances of numbers that were not dates or page numbers.

Lastly, common words in the journal or book names were abbreviated to reduce their weight so that more weight was put on rare words that distinguished citations better. See list of abbreviations in the raw data folder.

Author name standardization

To decrease the number of comparisons, the citations were blocked by author. Blocking the citations by first author also makes the clustering algorithm more specific since only documents that are likely to be similar i.e. ones that share a first author get compared. Before blocking, the author names were standardized because they were often misspelled or formatted differently across the citations.

Authors were standardized by merging similar author names using the same Jaro-Winkler similarity measure combined with the hierarchical clustering algorithm that merged them below a .20 threshold. Authors who got combined into large clusters were split up into smaller clusters.

This is enough to get a pretty close standardization. There are always a few authors that have names that make up other author names (e.g. Stern and Sternman) that get combined into a single cluster, but I accept that because I will be retaining the original author names anyway.

Another issue to note with the author blocking is that if the author names are exceptionally long (for example over 25 characters), which is typical to some organizations, those will stand out much more and therefore will make the clustering broader. Broader clustering means less specificity and more seemingly or relatively similar authors getting grouped together into a single group just because they are relatively very different from the authors with longer names.

To help prevent this from happening I recommend, cleaning up authors with exceptionally long names. The cleaning abbreviates the author names when possible and is now built into the standardization function.

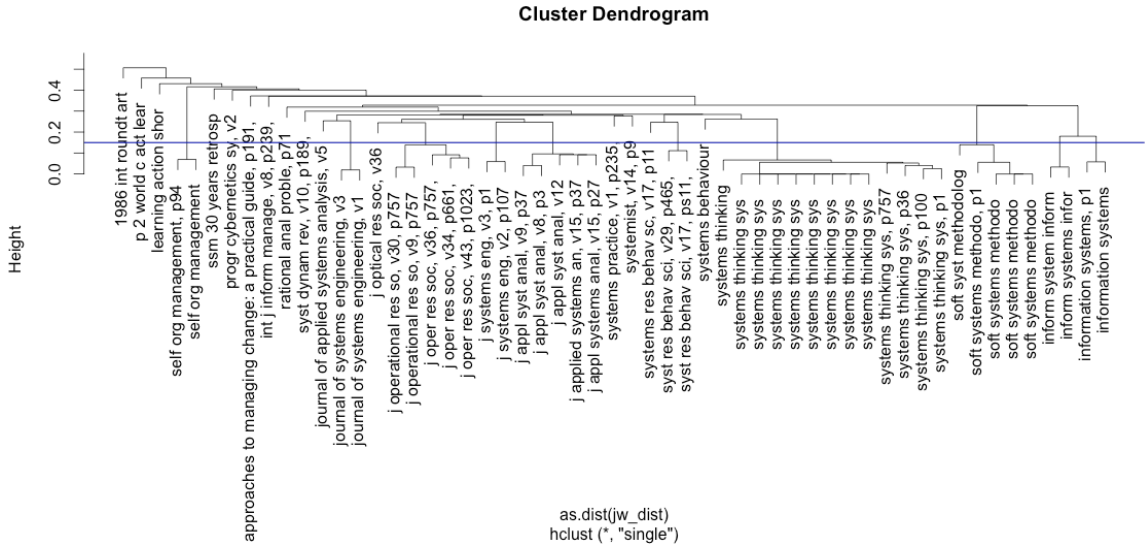
Citation merging:
For the comparison, I use the extracted journal name rather than the full citation to eliminate the influence of author names and year. This is okay because citations are blocked by authors. Ignoring the year of citation does merge citations from different years that contain the same words. This is good for books, but bad for journal articles that are submitted to the same journal.

There are three different approaches that I tried. First, I merged just based on the author names. Second, I merged based on author name and year, and third I merged based on author name, year, and document type. The third approach gave us the best results.

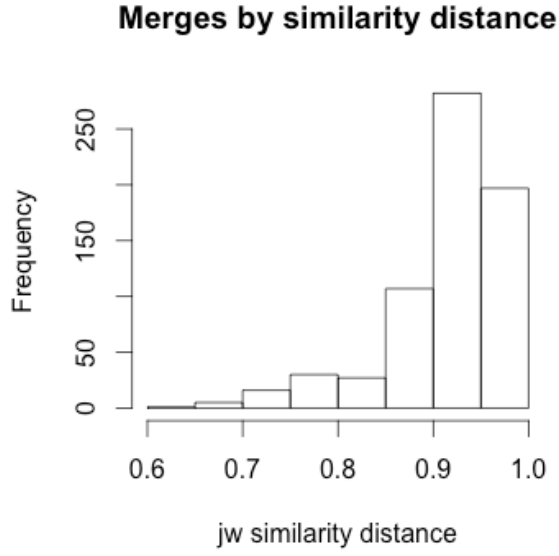
Approach 1

Within each author block, I cluster anything below a .15 jw distance threshold. Then I calculate the jw similarity to compare the original clean citation to the new merged citation. Again, similarity of 1 means the two are identical.

Below is an example of how Checkland clusters. There are about 55 different journal documents in which Checkland appears as first author. You will see that only the name of the journal or book appear in the clusters. This is because I eliminated the author names to make the matching better. The blue line represents the .15 threshold, so anything below the blue line gets merged. Click on the image link to view a larger image.

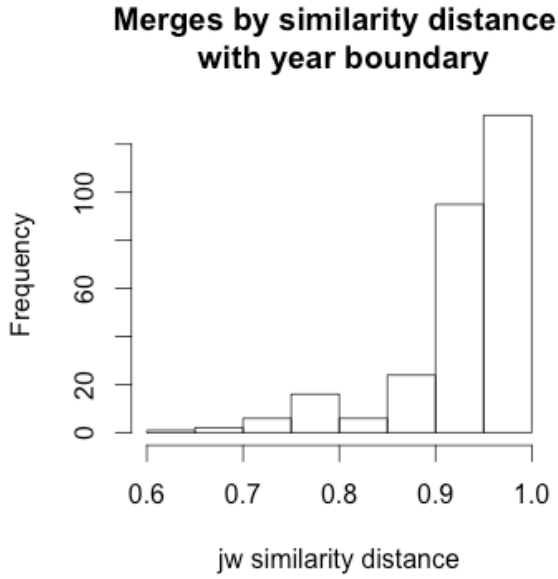


There were about 688 merges. Below is the count of the number of citations that get merged at each value of jw similarity. The SCI 2 program typically merges at .95 similarity.



Approach 2

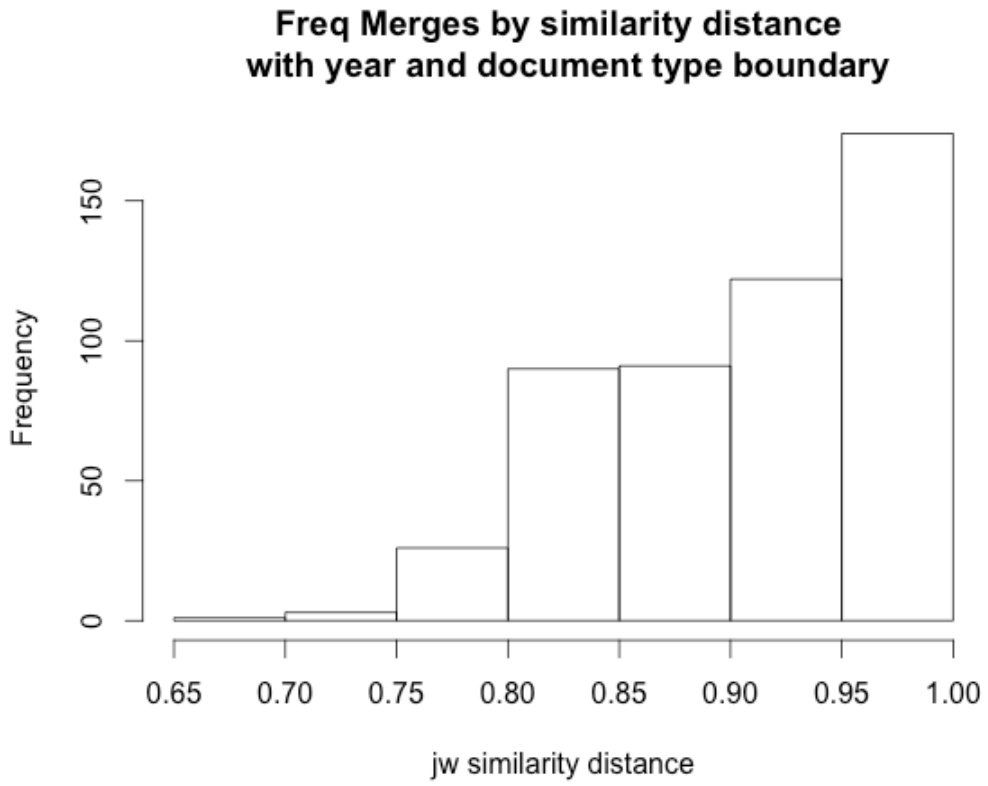
Second approach is to restrict the merges to only those that share the same author and year of publication. When I use the year boundary to additionally restrict, there are 285 merges.



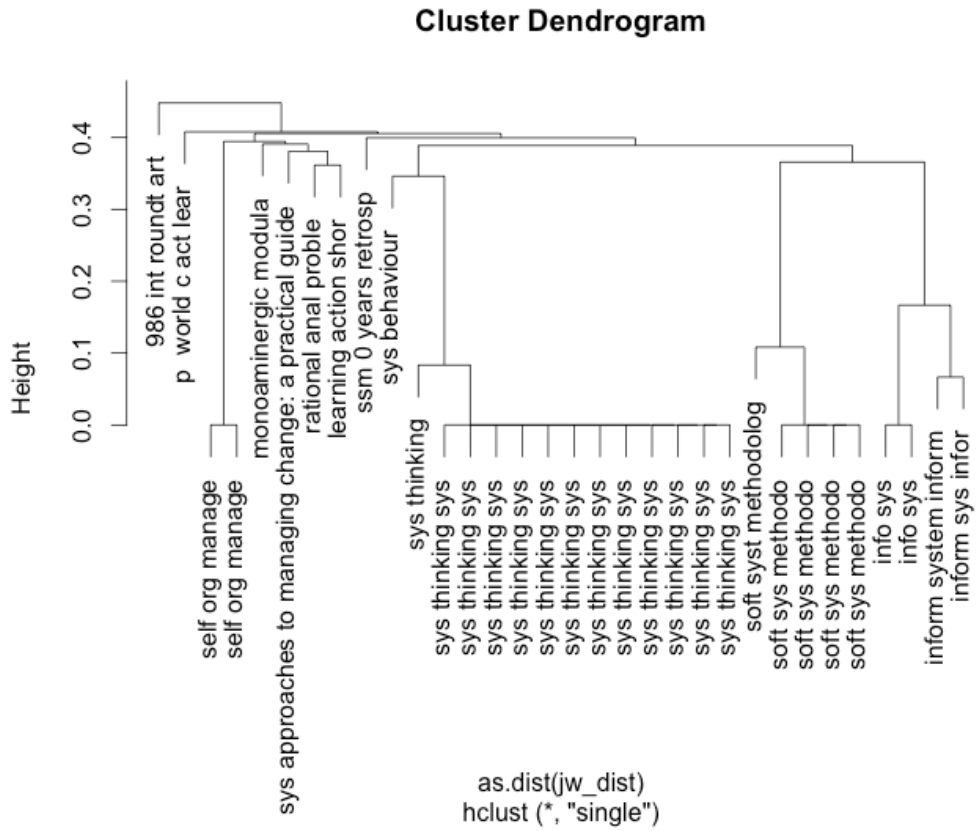
Approach 3

Third approach is to create two different journal information types. One for journal articles which would include anything that has a volume or page number, and another for books which would not have volume or page number information.

Using this method, there are 507 merges.



Cluster for Checkland’s books only



Sources:

Record Linkage - Wikipedia (https://en.wikipedia.org/wiki/Record_linkage)

Jaro-Winkler distance - Wikipedia (https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance)