Analyzing the Structure of Knowledge, Finding Strongly Connected Components within Wikipedia

Robert Micatka
University of Michigan

December 20, 2013

Abstract

Analyzing multiple wikipedias by finding strongly connected components hints at the underlying structure of information within these collaborative online encylopedias. The number and size of the strongly connected components within this database of knowledge will allow the organization to be examined. The number and size of the nodes created by the page links will give a clear picture of the macro organizational structure of several different language wikipedias.

I. Introduction

Wikipedia is a collaborative, multilingual, open, online encyclopedia. Launched on January 15, 2001 by Jimmy Wales and Larry Sanger Wikipedia has grown exponentially over the years with the English Wikipedia containing over 4.4 million articles and has an estimated 365 million readers annually. [1] Bolstered by these impressive statistics, wikipedia has a very real claim of containing nearly the sum total of written human knowledge. Due to the success of the English language wikipedia hundreds of other languages have spawned their own versions of wikipedia based on the same foundations.

The ability of anyone and everyone to edit Wikipedia is both its greatest strength and its greatest weakness. The benefit is that experts from all over the world can edit and update articles in their speciality, allowing articles to be always be up-to-date and complete. The downside is the possibility of vandalism. However, there are many volunteers who keep track of page edits and will quickly revert them if vandalism is detected. In addition a 2005 investigation by Nature showed that the science articles had a level of accuracy similar to that of Encyclopaedia Britannica. [2]

In order to analyze this vast, complex web of data a graph theory technique of strongly connected components will be utilized. Strongly connected components are defined as a maximal set of vertices from a directed graph $G = (V,E)$ where every pair of vertices within the set can are reachable from each other. [4] Finding strongly connected components within the directed graph of wikipedia internal page links will allow the structure to be uncovered. Finding the strongly connected components will allow identification of the main nodes of knowledge.

In order to find the strongly connected components, after creating the directed graph, Kosaraju's Algorithm will be utilized. This algorithm uses the transpose of $G = (V,E)$, a directed graph, which is defined as $G^T = (V,E^T)$. This creates what is essentially the same graph but with the directions of the edges reversed. The generation of the transpose graph allows the strongly connected components to be computed in linear time using two depth-first searches. The pseudocode for the algorithm is as follows: [4]

Stongly-Connected-Components(G)

1. call DFS(G) to compute finishing times $u.f$ for each vertex u
2. compute G^T
3. call DFS(G^T), but in the main loop of DFS, consider the vertices in order of decreasing $u.f$ (as computed in step 1)
4. output the vertices of each tree in the depth-first forest formed in line 3 as a seperate strongly connected component

Kosaraju's algorithm is linear-time as the time to create the transpose of G is linear in time as well as the two depth-first search passes. This allows for relatively fast execution time which is needed when working with such large data sets. There are faster algorithms in practice such as Tarjan's strongly connected components algorithm which only performs one graph traversal, however, Kosaraju's is much simpler to implement.

II. Methods

In order to analyze Wikipedia effectively an offline version was required. There are datadumps and backups taken every couple of weeks for all of the different wikipedias which are available to the public for download. The backups consist of a large xml file that constains all of the articles as well as internal and external links within the articles. No images are included but are available for download as well, however for this project images wererequired. The xml file for the full English wikipedia is on the order of 44 gigabytes. Due to computational limits smaller wikipedia versions were chosen, namely the Korean, Esperanto, Danish, Hindi, and Simple English wikipedias. This contains only 97 thousand articles, compared to 4.4 million articles for the full English wikipedia, but is only 450 megabyes which allowed for easier computation. [6]

Table 1: *The Size of Different Wikipedias Used in Project and English Wikipedia for Reference*

Wikipedia	Article Count (thousands) [6]
English	4,400
Korean	254
Esperanto	188
Danish	183
Hindi	100
Simple English	97

The raw data from the xml file is not suitable to creating a directed graph. The internal links must be resolved in order to show what other pages you can reach from any given page. In order to accomplish this a 3rd party tool was utilized. Evgeniy Gabrilovich created a tool named wikiprep for his Doctoral thesis work on "Feature Generation for Textual Information Retrieval Using World Knowledge." [3] Wikiprep parses the xml dump and resolves the internal links as well as many other tasks that are irrelevant for this project. The new, extended xml file generated by wikiprep was then parsed and the link information extracted.

Using the link information generated by wikiprep a directed graph can be constructed. The links information contains the node, the article, and its corresponding edges, links to other pages. This graph was created using a program written in C++ and stored internally, displaying the graph would not provide useful context to this examination. Using Kosaraju's algorithm on the created directed graph produced a set of strongly connected components and what they consisted of. The list of strongly connected components was reduced by removing all components that consisted of a single node, this occurs when a node is not part of a strongly connected component as it inherently is connected to itself. Removing this does not influence the analysis as nodes linking to themselves are not useful or interesting in this project.

After the list of strongly connected components was created the size of the components was then found. Examining the contents of the components would not provide useful contex for this project. Finding the size of the compenents allows finding the number of components of a similar size. This allows the overall network structure to be seen as it shows the number of small, medium and large strongly connected components within wikipedia. This elucidates the overall structure showing many large, more centralized nodes with more smaller, less connected components on the outskirts.

Note: All code used by this project can be found on my github at `https://github.com/s1syphus/ComplexSystems511FinalProject`

III. Results

Table 2: *The Frequency of the Size of Strongly Connected Components within Different Wikipedias*

Wikipedia	2	3	4	5	6	7	8	9	10	11 - 20	>20
Korean	725	198	68	24	9	5	8	2	8	6	1
Esperanto	428	91	31	13	5	1	4	4	5	4	2
Danish	725	114	34	26	6	4	1	1	4	2	1
Hindi	282	62	20	8	2	4	0	1	5	5	2
Simple English	293	62	20	10	9	3	3	1	4	3	1

Table 3: *The Normalized Frequency of the Size of Strongly Connected Components within Different Wikipedias*

Wikipedia	2	3	4	5	6	7	8	9	10	11 - 20	>20
Korean	68.79	18.79	6.45	2.28	0.85	0.47	0.76	0.19	0.76	0.57	0.09
Esperanto	72.79	15.48	5.27	2.21	0.85	0.17	0.68	0.68	0.85	0.68	0.34
Danish	78.98	12.42	3.70	2.83	0.65	0.44	0.11	0.11	0.44	0.22	0.11
Hindi	72.12	15.86	5.11	2.05	0.51	1.02	0	0.26	1.28	1.28	0.51
Simple English	71.64	15.16	4.89	2.45	2.2	0.73	0.73	0.24	0.98	0.73	0.24

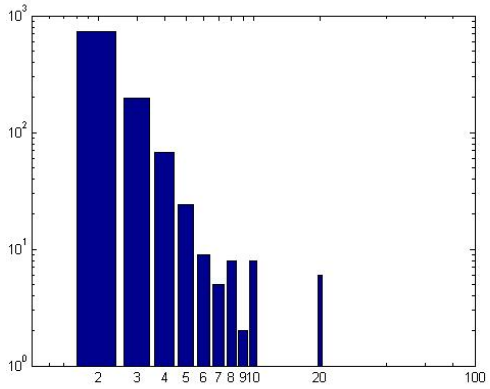

Figure 1: *Histogram of Frequency of Strongly Connected Component Size for Korean Wikipedia*

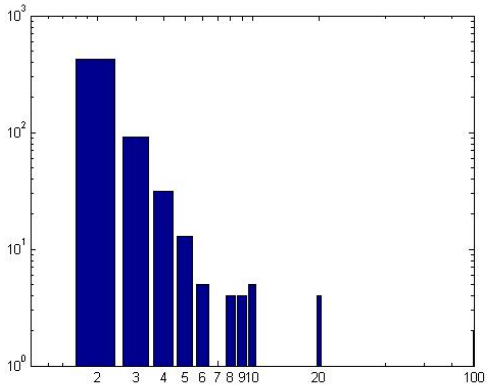

Figure 2: *Histogram of Frequency of Strongly Connected Component Size for Esperanto Wikipedia*

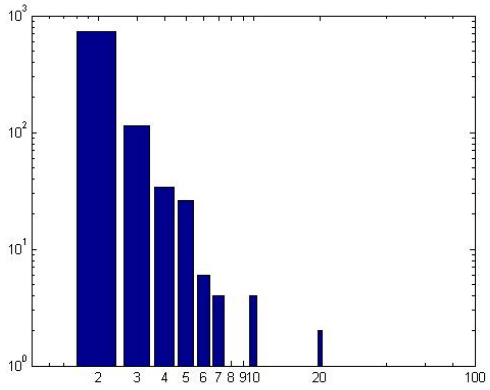

Figure 3: *Histogram of Frequency of Strongly Connected Component Size for Danish Wikipedia*

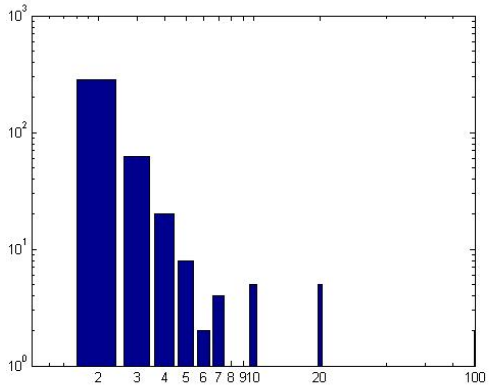

Figure 4: *Histogram of Frequency of Strongly Connected Component Size for Hindi Wikipedia*

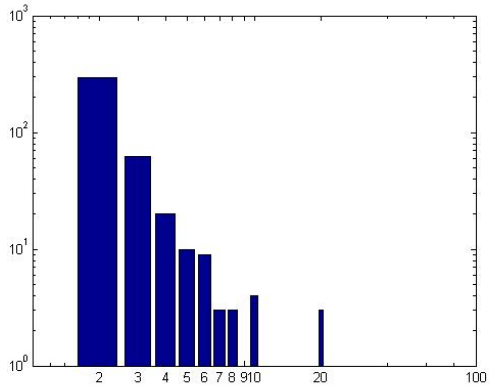

Figure 5: *Histogram of Frequency of Strongly Connected Component Size for Simple English Wikipedia*

IV. Discussion

The number and sizes of the strongly connected components mean...I don't know yet

V. Further Study

Analyzing several smaller wikipedias demonstrated... However, due to computational limits the larger wikipedias, those over sevaral thousand articles, have not been analyzed. Improving the implementation of the algorithms utilized in this project would allow these larger and more interesting databases to be analyzed. An improvement that could be made would be to parallelize a strongly connected components search algorithm. This optimization has been proven to exist with speedups up to 24x demonstrated. [5] This optimization would allow for faster and more efficient computation allowing the larger databases to be examined in a timely manner.

Once the strongly connected components are found, the next step in understanding the structure of the knowledge within wikipedia would be to look at the context of these components. This could allow the components to be compared to the category tag that is assigned to them within wikipedia. This comparison would be interesting to see which seemingly different categories are connected as well as how they are connected.

References

[1] "Wikipedia." *Wikipedia.* Web. 19 Dec 2013. <http://en.wikipedia.org/wiki/Wikipedia>.

[2] "Internet Encyclopedias Go Head to Head." *Nature* (15 Dec 2005). <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>.

[3] "Wikipedia Preprocessor (Wikiprep)." (2 November 2010). <http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/>.

[4] "Cormen, T., Stein, C., Rivest, R., and Leserson, C." (2009). *Introduction to Algorithms.* (3rd ed).

[5] "Hong, S., Rodia, N., and Olukotun, K." (2013). *Technical report: On fast parallel detection of strongly connected components (scc) in small-world graphs*. (Pervasive Parallelism Laboratory, Stanford University, Stanford, CA). Retrieved from <http://ppl.stanford.edu/papers/techreport2013_hong.pdf>.

[6] "Wikipedia Statistics." *Wikipedia Statistics.* Web. 20 Dec 2013. <http://stats.wikimedia.org/EN/TablesArticlesTotal.htm>.