Chinese Character Frequencies

Beining (Jenny) Zhang*

Massachusetts Institute of Technology

1 INTRODUCTION

With over 50,000 characters and no alphabet for these characters, Chinese is often considered a difficult language to learn. Pinyin was introduced as a system for defining pronunciations of characters using the Latin alphabet, but learning how to identify and write the characters still remains an essential part of learning the language.

This project demonstrates the frequency of different words and characters in Chinese, focusing on the ones that are used most often. Very broadly, this project aims to help the user learn something new about the language, regardless of their prior experience with it. This project highlights commonly used component and breaks down characters into more easily understood pieces to assist, making it more accessible, while also providing additional means for the user to explore the data to gain deeper insights.

2 BACKGROUND AND RELATED WORK

This section introduces some background on the Chinese language as well the datasets used for this project and other sources of inspiration.

2.1 Background

In Chinese, words can be made up of one or more characters. Characters themselves are defined as an order of strokes. They often contain a semantic radical, a subset of strokes that can lend the character meaning. It can also containg a semantic component, another subset of strokes that can influence the character's pronunciation. Figure 1 shows the character "zhāng," which can mean "to expand, stretch." It has 7 strokes. The left side (red) is the semantic radical, meaning "bow." The right side (blue) is a character that can be pronounced "zhǎng." This character is an example of how the phonetic and semantic components combine in a character

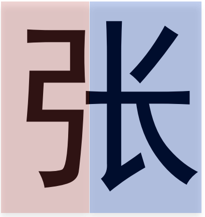


Figure 1: Character "zhāng," meaning "to expand, stretch." The red section indicates the semantic radical, and the blue section indicates the phonetic component.

2.2 Datasets

The main dataset used is the SUBTLEX-CH dataset, which details the number of times a Chinese word appeared in the analyzed set of movie subtitles [1]. The dataset also included data on frequencies of characters and words along with their semantic role, but these were not used in this project. The authors of this dataset argue that the frequency of words in subtitles is a good indication of their usage frequency in general.

Radicals and their meanings were taken from the Make Me a Hanzi dictionary [2]. Definitions for both individual characters and words and Pinyin pronunciations were taken from the CC-CEDICT dictionary [3].

2.3 Other Related Work

The decision to use a force network graph was based on my experience seeing similar tree-like graphs for English words, where words were connected by similar roots, prefixes, or other characteristics. Yan et al. also developed a similar node-based figure for Chinese characters as part of theirwork for developing a more efficient method of learniing Chinese [4]. In their graph, nodes are grouped by radical or other character components.

3 METHODS

The work for this project can be divided into data parsing, creating the graph, and developing user interactions.

3.1 Data Parsing

In the first step, I developed a Python script to parse the SUBTLEX-CH data and convert it into data in the form of nodes and links, the format needed for force network graphs. I decided on having each node represent a character, and for links between nodes to represent two characters appearing in a word. Since a word can contain more than two characters, a word in the graph may then be represented as multiple nodes and links.

I decided on this encoding for nodes and links after considering using links to represent the same semantic radical or the same phonetic component. I decided to use radicals or character components as links as it would result in clusters of strongly connected nodes, with no connections between any two clusters. The graph would then be fairly predictable. Grouping by shared words can result in more unexpected configurations, due to characters very different in meaning becoming connected, or certain characters having many connections.

I generated one data file in the format of nodes and links and another with detailed data on characters, radicals, and words created by combining all three datasets. Character data included Pinyin pronunciation; definition; total frequency/number of uses; a breakdown of usage based on word or with another character; and all words it appears in. Radical data included definition and words with the radical. Word data included Pinyin pronunciation, definition, and frequency. The two data files were linked by unique IDs given to each character.

* beining@mit.edu

3.2 Force Network Graph

The core part of the visualization is the network graph. Node radius is proportional to $f^{.4}$, where f is the number of occurrences of the character (either by itself or as part of a word). This value was determined after some experimentation. Due to the most frequent character occurring 4x as often as the tenth most, a directly proportional equation would result in most nodes past the 50th most common being virtually invisible next to the largest node. On the end, a logarithmic scale resulted in the sizes too similar, making it difficult to distinguish smaller differences. This proportionality does well in that it makes the largest nodes obvious while not reducing the smaller ones to nothing.

Link width is constant for all links and does not vary based on the frequency of that word corresponding to the link. This was decided for two reasons: first, a link could correspond to multiple words if the source and target of that link both act as parts of multiple words. This either results in the link perhaps being misleading about the frequency of words or the encoding becoming confusing with a distinction between what percent of the link corresponds to what word. Second, since the links are already relatively thin relative to the nodes in order to not block the nodes themselves, differences in link width would not be obvious enough to effectively convey information, and would possibly just confuse the user more as they try to understand this additional encoding, taking away from other exploration that could be done.

Text is shown on larger nodes to increase attention to the most frequent characters and provide starting points for any exploration.

3.3 User Interactions

3.3.1 Graph Interactions

Users can hover over nodes to highlight the node, the node's neighbors, and the node's links. This highlight results in a darker opacity. Users can see summarized information about the node when they hover over it; this includes pronunciation and definition. This information can help those who are not as familiar with Chinese to understand the connections better and more quickly (Figure 2).

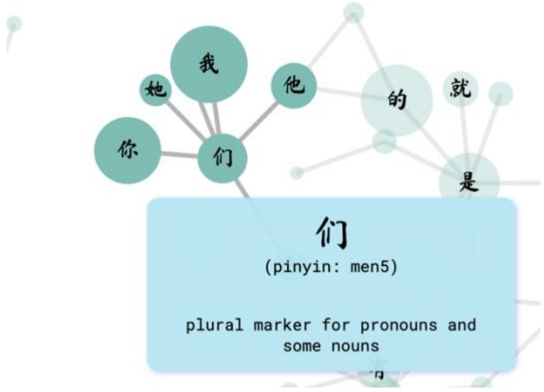


Figure 2: Hovering over a character

Users can also select a node to see more information about it in the information window on the right side of the page. Selecting a node performs the same highlight as hovering, but the selected node also changes to a brighter color (Figure 3). This choice was made because I wanted a distinction between a node that was just being hovered over and a node whose detailed information was currently displayed. Additionally, a user can still hover over other nodes while they have some node selected, so the encoding for hovered and selected nodes should be different. The color change for the selected node makes it the focus, as it should be. Since the user's cursor will be over the hovered node when it becomes highlighted, the "hovered" encoding does not need to be as severe, and a change in opacity is enough.

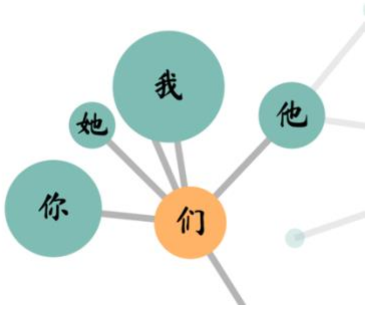
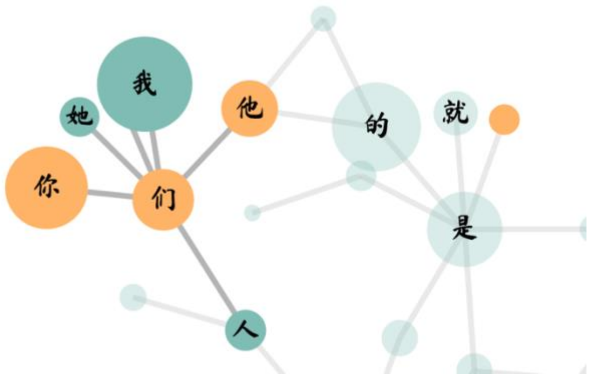


Figure 3: Selecting a character highlighting

When a character is selected, the right-side information window contains the character's pronunciation, definition, radical, radical meaning, and other words with the same radical. Users can choose to highlight characters with the same radical as the selected character. As mentioned before, words that share radicals often have similar or related meanings, and identifying radicals in characters can greatly assist in learning characters. If the option is turned on, same-radical words are highlighted in the same color as the selected node. Since these words are not necessarily going to appear right next to the selected node, changing their color to make them stand out will make it easier for the user to identify them (Figure 4).

Figure 4: Highlighting characters that share a radical



Additionally, the right-side information window contains a breakdown of how the selected character is used. It displays percentages of when it occurs by itself as well as when it occurs within each word currently included in the graph. When a user hovers over the word item, the characters in that word become highlighted in the graph. The highlight involves a change in color to make them stand out against other neighbors that are currently also opaque.

Finally, a user can drag and drop nodes to see connections more clearly or uncover overlapping nodes.

3.3.2 Additional Interactions

The left-hand sidebar has additional options for customizing the graph, the goal of which is to make the interaction experience more suited for the user's specific needs and interests.

In the left-hand sidebar, users have additional options for customizing how the graph looks. First, they can turn off the summary information that appears on hover. This can be useful if there are many nodes close together and the window potentially blocks other nodes when it appears. Second, they can turn off the text on the larger nodes. This is also helpful if there are many nodes close together and the text becomes hard to read. Third, users can adjust the number of nodes in the graph. This is helpful if the user just wants to see the most frequently used words and other nodes interfere with viewing. Finally, the user can reset the graph view and settings.

4 Results and Discussion

This section expands upon the visualizations described in Section 3 with possible use cases and findings.

One potential user is someone with no background in Chinese. The instructions page provides some background. They can reduce the number of nodes to view a smaller version of the graph first, and they can start by looking at the larger nodes that have text labels. One larger node is "们," the plural indicator in Chinese, which they can learn from the summary in the hover tooltip. They can hover over the neighboring nodes or click on "们" first and then hover over neighbors to learn that they neighbors include the words for "he," "she," "I/me," and "you." They could try to guess what word the link represents, knowing the meaning of the sources. Alternatively, they can read the meanings in the right-hand side information window, highlighting over the words to see exactly which characters are part of it, and discover that the links correspond to "they," "they,", "us", and "you all," respectively. From this, they can perhaps learn something about plurals in Chinese – rather than having a different form of the pronouns like in English, in Chinese, pronouns are just combined with the plural indicator to generate the plural version of it.

For a user with more experience in Chinese or for the same user wanting to explore more, they may choose to select a character and explore what other words share the same radical. If they're also examining "们," they can see that a lot of the commonly used characters (such as "he" and "you") contain the same radical as "们" – the "person" radical. This perhaps confirms the hypothesis that pronoun-related words are generally commonly used words, but there are also non-pronoun words that have the same radical – words like "信" ("to believe, trust") or "做"("to do"). While "to do" might be expected as a common character, "to believe, trust" might be more a surprise. The user can click on that character to find that even though it does appear quite frequently, it isn't used by itself – with the meaning "to believe, trust" – as often as it's used as part of "相信" – "to believe something is true."

In both cases, the user learns something about the language – whether it be simply what words are commonly used, how the language is similar to other languages they may know, or how common characters are used.

5 Future Work

There are a number of ways in which this project can be further extended to allow for a greater variety of user interactions and analyses.

First, there are many ways to describe and categorize Chinese character that were not explored in this project. For one, this project focused on radicals that usually provided semantic meaning, but characters' phonetic component is also an important part. Characters sharing the same phonetic component often have very similar pronunciations, sometimes only differing on tone. Understanding both the phonetic and semantic radical connections between characters can help one understand some of the ways that Chinese characters are constructed and defined, and in this way, help one learn the language more easily. An extension to this project would be to add on to the infrastructure already in place for exploring semantic radicals in order to create functions allowing the user to explore phonetic radical connections.

Along with radicals, Chinese characters can be categorized by how they are broken down into components – whether there is a left and right piece, top and bottom piece, and so forth. They can also be categorized based on how many strokes it takes to write the character. Integrating this data into the visualization could add another interesting layer to explore.

In general, a useful extension would be to allow more user control over the graph. Currently, there is no grouping aside from the ones that result from the links, and the (x, y) position of nodes does not encode any information. Allowing the user to group by radical or by any of the characteristics mentioned above could allow for new patterns to be discovered and more interesting analyses. In addition, user controls over how close nodes are

Finally, as mentioned previously, the original dataset of word frequencies contained over 100,000 entries, but this visualization only used .4% of these for performance and aesthetic purposes. More work can be done to discover a way to use more of the data efficiently. One possible way would be to aggregate multiple characters that are related in some way into one node and allow the user to explore the "sub-nodes" by zooming in or through some other action.

References

[1] Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *Plos ONE, 5(6), e10729.*

[2] skishore (2018). Make Me a Hanzi. Github Repository. https://github.com/skishore/makemeahanzi

[3] MDBG (2016). CC-CEDICT. https://cc-cedict.org/wiki/

[4] Yan, Xiao-Yong & Fan, Ying & Di, Zengru & Havlin, Shlomo & Wu, Jinshan. (2013). Efficient Learning Strategy of Chinese Characters Based on Network Approach. PloS one. 8. e69745. 10.1371/journal.pone.0069745.