Visualizing Collections of Archived Webpages

John Berlin, Joel Rodriguez-Ortiz, Slobodan Milanko

Department of Computer Science, Old Dominion University,CS725/825

In this document, we will first address the initial description of the selected project, consisting of what we know and what we aim to propose as a solution. This is followed by a brief discussion of the dataset that we are anticipating, and how we will mitigate risk if the dataset type changes. Lastly, we will address some questions we aim to answer with our visualization, which will be refined as we learn more about the data.

Initial project description

As a group, we have decided to propose an idea for visualizing collections of archived webpages. With general examples presented by Dr. Weigle, we believe that we can make a solution that will help users navigate large archive collections more easily. One of our plans is to make the visualization clutter free, where we aim to ensure users can manage the amount of information displayed, no matter the size of the collection. Our next goal is to give the visualization compatibility features, providing accessibility over many browsers, and lightweight features, the power of dynamically filtering data to increase performance. Lastly, as we approach further milestones in the project, we aim to continuously assess the effectiveness of our proposed solution by surveying our advisers and clients.

Dataset

Per our email conversation, we will first reach out to Yasmin, in hopes of understanding her already developed collection. This will give us a good idea as to what the client seeks to see, and what data is ready for us to use. If the provided material is not sufficient or lacks in quantity, we will follow a general format for Google Docs in creating sample collections. Lastly, we also aim to talk to Michigan State clients with Dr. Weigle, to ensure the dataset we use will meet their needs.

On a more specific note, we anticipate the dataset being presented to us in a table form. Each item within the data, at a minimum, will contain an attribute for a target webpage and an archived webpage. In addition, we anticipate categorical attributes for tags, which will give more context to the webpages we are differentiating. While uncertain, we forecast clients will desire several other categorical, temporal or quantitative attributes to filter data easier.

Questions

The current list of questions, in no particular order, includes:

- How does the archived webpage differ from the currently live webpage?
- Is there a relationship between specific tag combinations and the ratio of difference between current and archived webpages?
- Are some webpages better archived than others?
- Is there a specific period where archived pages were damaged the most?
- Will the user want to navigate the archived network spatially?