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 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

What?

Each item within the dataset, at a minimum, will contain a URI-M (memento) attribute for a URI. In addition, we anticipate categorical attributes such as contributors and tags, which will give more context to the webpages we are differentiating. While uncertain, we forecast clients will desire several other key identifiers, temporal or even quantitative attributes to filter data easier. This leads us to believe that the dataset will be presented in two distinct forms, a table or a tree layout. The attributes for each table entry could contain a specific tag, time date value of when a URI was archived, or even a set of mememots that construct a story. On the contrary, a tree layout could provide us with a set of nodes where each link is a specific tag. It is up to the user to define the primary attribute, and relationship, they seek to focus on.

Why?

Since the goal of our visualization centers around story telling, our users will be presented with a wide variety of analysis goals, such as discovery, presentation, and enjoyment. They can discover new knowledge or insights about each story within a web archive. In addition, the presented material can help narrate the same story of how particular archives eventually transformed into something distinctively different than its origin. Which ever the case, full enjoyment is achievable to any user getting a glimpse of what was once collected.

When it comes to searching, a quest towards understanding can simply begin by exploring the entire collection. If more data is known, such as a categorical tag and a specific URI, the users can even perform a lookup or a location search. Browsing will also be possible, in which a user might not necessarily know what they seek to see, but rather a time frame of when it might have happened. A simple example of this can include searching for mementos with tags, by only knowing a specific temporal value.

Finally, when querying for data, users have a choice of identifying, comparing, or summarizing the presented dataset. Identifying will allow a shift in focus towards a particular element, such as the URI and its associated tags over time. Comparing can help further differentiate multiple selected attributes, where finding key areas of differences is critical. With this, users can find answers to questions such as: how has the content or role of the resource changed, from the original starting point? Has the story shifted significantly over time? Taken even further, addition of summarizing queries, can help outline a specific generalization about stories, such as the most common set of tags presented within the collection.

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?