Visualizing Web Archives of Moderate Size

John Berlin, Joel Rodriguez-Ortiz, Slobodan Milanko

Abstract—Visualizations offer a powerful approach to understanding large, or even moderately sized, collections of data. The ability to discover, explore, and capture the most important aspects of collections creates a powerful platform for understanding. For our research, we leverage this platform as a means of understanding web archives in a multi user setting. In this study, we propose a new approach towards identifying the general state of the archives, by using contextual data provided during the archival process. This meta data allows us to identify the most common domains, archived resources, times and tags associated with a web collection. We find this approach an adequate solution for multi user settings, where the designed tool outlines the most important areas of focus in web archives and gives users a more clear picture of what their collections comprise of, both in specific and general terms.

1 Introduction

Collecting data is amongst the most important steps in further understating a problem. While the amount of data that needs to be collected is dynamically driven by a problem, the goal of creating a collection is not. They primarily exist to give answers, or possible guidelines, to questions that might not have been answered. They stand as proof that a particular concept or characteristic lies in the underlying dataset. While it is easy to agree that collections are extremely useful, its important to note that in some cases, they can also be very challenging to understand. One of the biggest drawbacks of having large amounts of data, is the associated complexity of extracting useful characteristics. As the collection increases, so does the difficulty in understanding and viewing it as an entity. In other words, the quality of information we are able to retrieve in collections drastically decreases as the user becomes more overwhelmed by the amount of content. While this is true for a plethora of different categories and topics, it is also evident within web archive collections.

Archiving the web, or any other source of data, gives us the ability to replay an experience that a user had before [1]. During this time, the typical approach involves locating a resource of interest, referred to as a URI-R, and creating a memento of its existence. A memento, also referred to as URI-M, is the archival record of a resource at a particular time. Unlike the URI-R, mementos remain snapshots that represent a particular instance, whereas resources evolve over time. [2]. It is completely possible to get a good understanding of the web archive collection, inferring that the topic of interest remains consistent and the number or archivers participating is minimal. However, as the size of archives increase, i.e. hundreds of unrelated URI-Rs and Mementos added by a collaborative set of archivers, understanding the entire collection becomes much more difficult. The problem then becomes, how can we get an understanding of a moderately sized collection of web archives, by visually presenting it in a non-overwhelming manner? Is it possible to create underlying relationships and use them towards identifying patterns, trends, or similarities in our collection on demand?

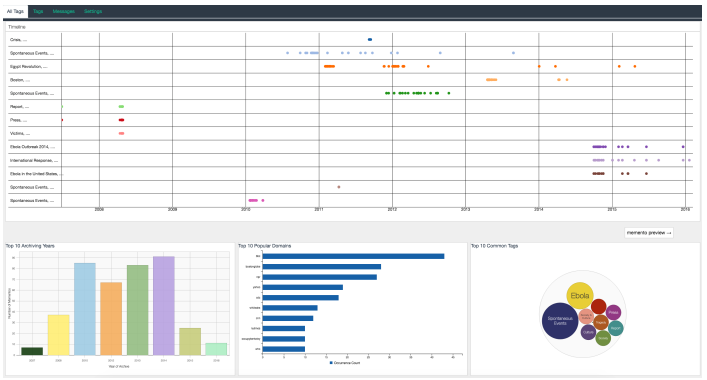

Fig. 1. Web Archive Visualizer

In our current solution, Web Archive Visualizer, we propose an approach that addresses those problems. The idea behind our solution is that contextual information created during time of archive, such as the name and keywords associated with a resource, can create a set of useful characteristics about the data. Using this additional information, we are not only able to relate sub sections of an archive, but also limit both the amount and the types of resources shown. While the cost of going through the archival process is noticeably higher, the potential increase in overall understanding far outweighs its limitations. Web Archive Visualizer can be seen in Fig. 1, where the system is separated into multiple views denoting particular points of interest.

2 Related Work

The vast majority of current and previous works of web archive visualizations places a heavy focus on their evolution. Dating back to 2005, the developers behind WebRelievo [3] propose a visualization to monitor the change of resources within a web archive over a particular period. Aside from a known archival date, the goal of the system aims to identify relationships between different web resources

- *John, Joel, and Slobodan are all students at Old Dominion University.*
- *The archiving process described was provided and outlined by Michigan State University*

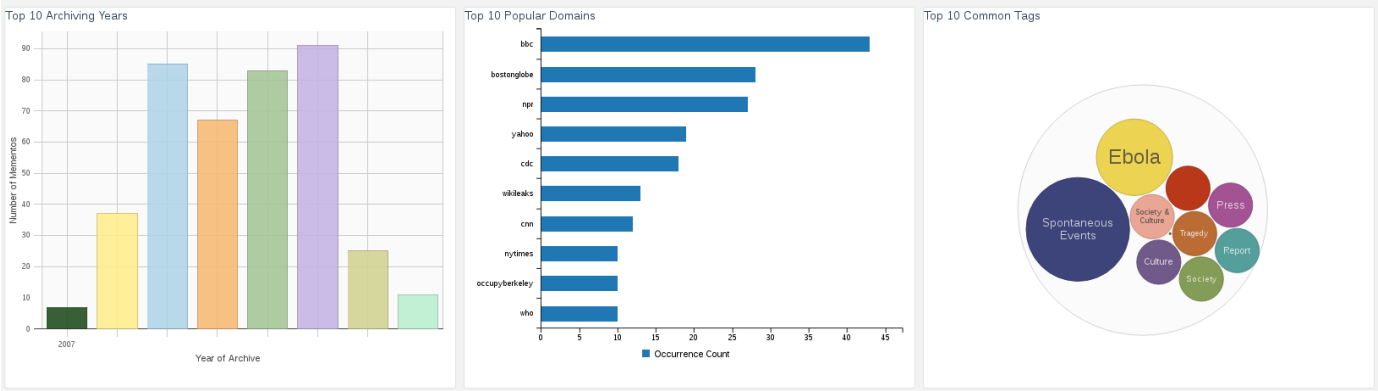

Fig. 2. This figure shows the general idioms that apply to the entire collection.

without any contextual data. In another words, they aim to see the interconnecting of the web and how it differs over a span of years, without taking any additional characteristics from the original archivers. Even though there is a lack of contextual data, WebRelievo proposes a good step towards relationship identification of resources within collections, regardless of the collection size.

Older works, such as those in [4], show relationships between resources located within similar communities. In this context, communities are clusters that all relate to a particular topic. While this approach also uses only temporal attributes of the archival process, the authors show the importance of identifying points of interest within a web archive. The general idea of this work directly supports ours, in which we leverage relationship identification of like meta data, i.e. tags and temporal attributes, to help classify and discover potential areas of interest.

3 Design

The real estate of the viewers attention is split into multiple views of interest, as seen in Figure 1. In both the top and bottom portion of the screen, referred to as the main content, users are given the ability to view specific characteristics of all data within the archive. The main content is complemented with arrows to help navigate between different specific views. In addition, text of the navigation links dynamically changes to alert the user about the next and previous idioms in line. We find this a helpful feature, where the element of surprise is limited, allowing the user to anticipate and possibly expand their exploration towards more specific charts.

On the contrary, the bottom portion outlines multiple views for more general characteristics that can summarize the entire collection. We refer to this section as the complementary content. Note, the split of interest between these two sections is important, as it gives users more control on the amount of data that they see at once. In essence, this is a strategically enforced measure to ensure cognitive memory does not become overloaded. Unlike the main content, views within the complementary content are always shown, as a proactive measure enabling comparison of data between the more specific content and the overall message sent my the web archive collection. Its centered location on the screen complements the use of comparison as it stays within the real estate bounds of different views.

To make the boundary between main and complementary contents more clear to the user, we separate generated visualizations by the area of categorical influence. This causes resource specific views to fall into several categories, including: timeline of resource archives, thumbnail views of mementos, resource clusters, and keyword groupings. They are granular enough to ensure that the targeted audience can fork out and create relationships of the low level details surrounding their data. On the contrary, idioms showing general characteristics primarily focus on the higher level summaries, such as outlining the most active archiving years, popular domains, and the most common tags.

3.1 Resource specific idioms

One of the primary areas of focus, alike most recent work, is time. Knowing when a particular resource was archived, is the first step in identifying differences over a lifetime and drawing conclusions as to how temporal periods affected particular resources. For this reason, Web Archive Visualizer places primary focus on the Timeline view, as seen in Figure 3. The purpose of this idiom is to let users view keywords associated with archived resources, over time. Aside from the obvious, keyword dominance over time, this chart also allows us to see the levels of activity as the collection years grow. When zoomed out, the scatterplot can outline this trend easily, leading us towards identifying temporal trends and outliers within our collection.

Another resource specific idiom presented in our vis is wordplay, as shown in figure 4. The purpose of this idiom is to outline the keyword clusters and its associations between resources and mementos. Its use is primarily evident when one wants to identify areas that are created by different keywords. Take for example a keyword “cat”. By using this view, the user can see mementos and resources that are centered around “cats”. In addition, the view allows us to create a cluster of like words, that are also found within points of interest. In the above example, this would create a closer distance between keywords like “animals” and “cats”. Lastly, if the categorical number of keywords becomes exponentially difficult to manage due to its size, filtering and zooming will create a controlled setting for many prospecive users.

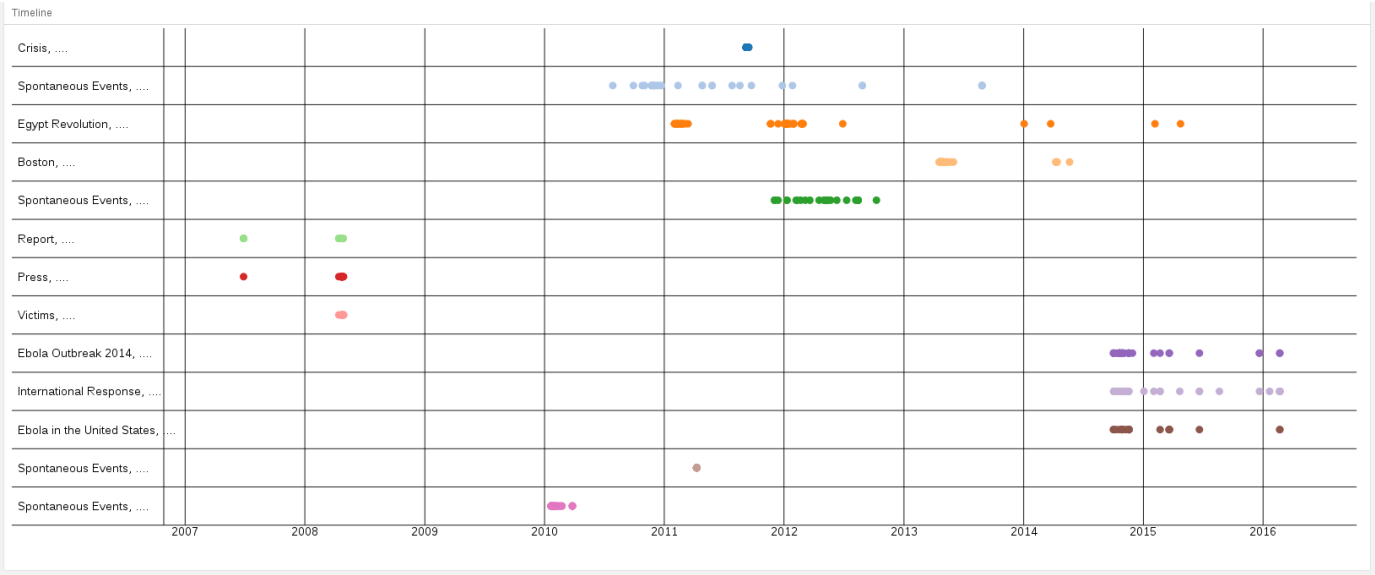

Fig. 3. Timespan chart showing archives associated with particular tags

Web Archive Visualizer also contains a keyword barchart, which allows us to view the associated counts of different tags, and how they relate to the rest of the collection. Depending on the area of focus in the archive, this particular idiom can be a great identifier of primary associations in the archiving efforts. In other words, it can help answer questions such as: "What words are primarily evident in our archives?" and "Is there a shift in categories of those words within the entire collection?"

For those who find bar chart counts limiting, the system also provides temporal tree maps. This idiom takes the barchart results a bit further, in which they identify the spread of keywords over time for both resources and mementos. While a bar chart showing keyword counts can help answer many questions, temporal tree maps can provide reasoning and awareness as to why those counts have their respectful values. This idiom allows for navigation through the years, and resources, allowing the user to pinpoint exactly what time is primarily responsible for an increase or decrease in keyword associations. Temporal tree maps also have another large benefit, in which they present size with importance per year. This makes it especially easy to compare how particular keywords relate to the remainder of that within a year.

Lastly, the system concludes resource specific views with the inclusion of keyword timegraphs, and their association with resources and mementos. This idiom aims to show both overlaps in archiving and which domains are responsible for the basis of our entire collection. Note, while there is no single correct chart to use when answering questions, it is the goal of these specific views to help answer as many diverse questions as possible. In some cases, many views can answer several similar questions, but some will make it much easier as compared to others.

3.2 General idioms

Figure 2 shows three different general views. The left most view, memento activity over a set of years, denotes the number of archives present for a particular resource in some time. Understanding the most active years has many advantages to archivers and general users exploring the collection. First, it allows them to conceptually visualize what periods of activity are responsible for the peaks within the collection. If the collection has a specific context, such as an archive consisting of hurricane events, particular years could outline the most influential years. The same is also true for identifying trends or outliers in data, where particular years might show a steady incline, decline, or abnormal levels of mementos collected.

Most popular domains, middle view in Figure 2, shows where most of the resources are being collected from. This is especially important to know when understanding domain reliance or influence of content of the web archive. Alike the idiom denoting years, understanding domain dominance can also help outline the skewness and bias within our collections. This is especially helpful, when the archiving team is collecting mementos relating to one particular category from a limitless scope of domains.

The last view in Figure 2 shows the most popular tags, whose occurrence count is represented in an increased size. Once a tag dominates others, we can gain more contextual awareness regarding our data. This can help us answer questions like: "Are we focusing on archiving particular areas?" and "Are there particular keywords associated with our collection subset?" Having this general view is also especially helpful when attempting to compare results found in the views of main content. By having two different keywords for comparison, we can begin to see dominance between the words of interest and those that are most popular.

3.3 What Why How Framework

Web Archive Visualizer was created as part of the Information Visualization class taught at Old Dominion University. In it we read through Tamara Munzner's "Visualization Analysis and Design", in which she presents a framework

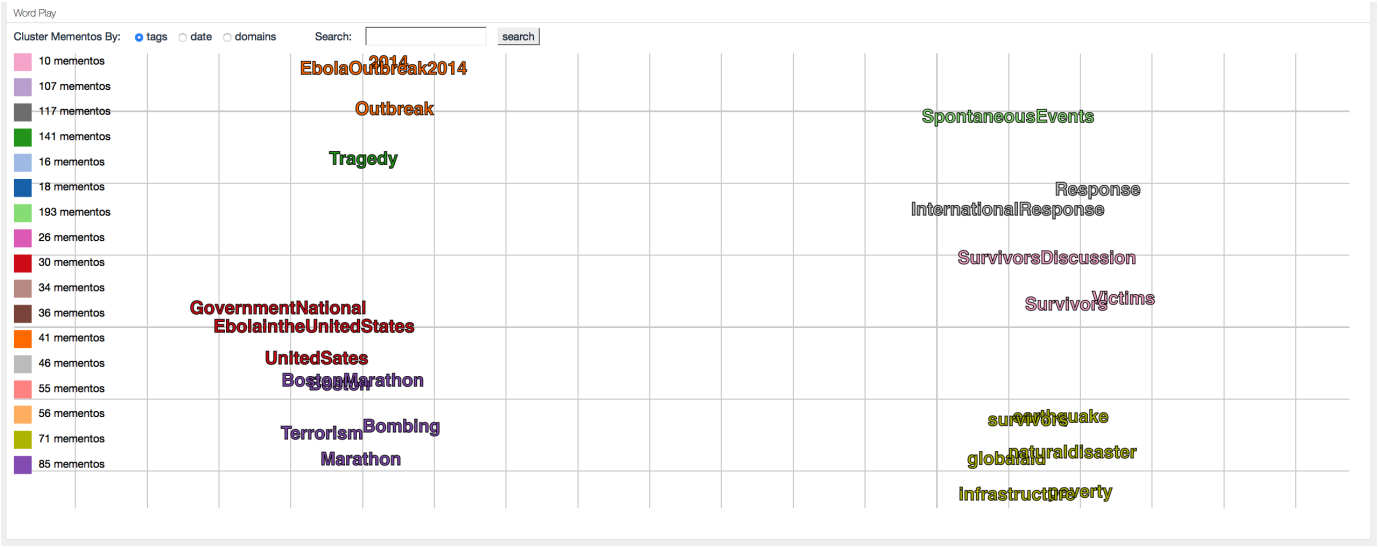

Fig. 4. Word play showing clusters and tag associations

for the systematic analysis of visualizations and a common language for their evaluation; the What, Why, How Framework. The What, Why, How Framework provides abstractions for validating visualizations at four distinct, nested levels of design. These are the Domain, Data/Task Abstraction, Encoding/Interaction Idiom, and Algorithm levels. This project's scope and duration did not allow explicitly performing the different validation techniques discussed by Munzner at the different levels. However, we justify the decisions made at each level with respect to the What, Why, How Framework next.

Our domain situation is explicitly characterized by the work and research performed by web archivers. Web archivers digitally preserve records so that we can later replay and study them. In the course of doing so, they amass collections which they need to share and query for facts in general. These problems of search, analysis, and query directly benefit from our vis tool support.

At the domain situation level we characterize URI-M collections from the digital preservation domain. A collection of mementos includes URI-Ms associated to different attributes such as tags, dates, archiver, and URI-R. We then produce a design that comes from requirements formalized by several interviews performed with the course Instructor.

At the data/task abstraction level we represent the collection of URI-Ms as a table with a number of items made up of a number of attributes. We did this to facilitate identifying a varied number of abstract tasks. Tables can be easily sorted and aggregated enabling a number of actions and targets. Attributes can also be easily appended or removed from a table without necessarily altering the views that present the data.

One of our primary goals was supporting a wide number of abstract tasks. Because of this we produced a system with a number of varied views, each supporting a unique set of actions and targets.

We support the discover action via views that provide knowledge that was not previously known. Our top 10 views are examples of this. We support the present action via a number of timelines which can be used for expressing how the data evolves over time. We also support general enjoyment; our visualization makes it much easier to digest the collection overall in comparison to its spreadsheet form. We support the lookup action via our word map view in which a specific (or incomplete) item or attribute can search for (or looked up), and then explored as a table for a number of individual targets. Our treemap, view supports the browse and explore action, allowing the dynamic navigation of item sets aggregated by different attributes. Finally, our visualization allows compare and summarize via our juxtaposed top 10 views. Most of our abstract targets involve individual items, individual attributes, and trends over time.

At the encoding/interaction level we made three primary decisions. The first is more of a rule of thumb, rather than an explicit encoding. It is the rule of thumb discussed by Munzner as "Overview First, Zoom and Filter, Details on Demand". A large chunk of our screen real estate is dedicated to summarize oriented idioms. These idioms start the user off with a general awareness of the entire information space and allow requesting incrementally specific item subsets categorized by multiple attributes.

The second was the use of juxtaposed views. Given a varied number of abstract tasks, we felt that overall delivery efficiency would be maximized by juxtaposing a number of smaller navigable views, each subject to its own secondary encodings. Some of these include ordering, selection, navigation, and the encoding of different attributes via color.

The third is extensive use of filtering and aggregation. A consistently driving goal was providing access to a dataset that lacks semantics in its raw form. Many of our individual views aggregate data items by attribute and filter for items of interest in attempts to mitigate cognitive load on the user.

Our primary goal was providing access to a dataset that is otherwise inaccessible in its raw form. Via the use of multiple idioms (summarized below) we present the dataset from multiple, varied perspectives that enable effective analysis, search, and query.

(Then you need to add the table in our presentation.) (And then to our table in the presentation we need to add this if it doesn't have it). **How: Encode — Timeline, Time Graph, Bar Chart, Histogram, Bubble Chart, Word Map, Treemap** (I think that's all of them in our vis?) (Is top 10 circle one called bubble chart?)

4 Evaluation

NOTE: Talk to Dan about note for this section. To perform our evaluation, we ran a performance analysis for load times of our visualization. The system used for this validation comprises of the following characteristics: AMD A6-3650 Llano Quad-Core 2.6 CPU, 4GB of RAM, and a 5400RPM 100GB hard drive. The proposed machine is running Windows 7 and Google Chrome Version 50.0.2661.86 (64-bit).

Looking at Figure 5, we are able to conclude that the performance of the Web Archive Visualizer is able to handle several hundred mementos. The system becomes much more strained as the memento counts get into several thousands. In respect to our goal, several hundred resources are the aim of moderately sized collection, whose times of loading fall within 2 seconds. For large collections, which were out of our scope, there is an exponential trend in load times.

5 Conclusion

Understanding and creating relationships within web archives are tasks that can quickly become difficult to manage. When the size of the collection is relatively small, those tasks can usually be easy to accomplish. However, as the collection grows into several hundred results, the difficulty of completing our tasks becomes much more difficult. For this, we propose a new system towards understanding moderately sized collections called the Web Archive Visualizer. Using contextual data provided during the archive process, such as keywords and temporal attributes, the proposed system outlines and answers questions about a collection that can be helpful to a wide variety of audiences; simple users looking to explore data and researchers discovering particular trends.

Web archive Visualizer is able to accomplish its goals by creating resource specific and general views. This split of information is a strategic approach to conserving cognitive memory, while allowing attention and focus towards particular areas of interest. The specific views allow for more low level characteristic analysis, spanning results in the entire collection. This includes idioms such as: bar chart showing counts of all keywords, timespan analysis of mementos, and the temporal evolution of resource archives. On the other hand, general views give us a higher level analysis of our collection, in which they summarize our collection. Having a combination of both is crucial, as the number of possible answers and comparison analytics can promote a larger understanding within our audience.

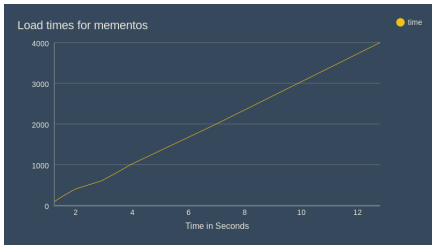

Fig. 5. Duration of loading mementos over time

6 Final Thoughts

NOTE: Phrase as: "In the future it would be useful to perform validation analysis on the system, instead of retrieve feedback?" In the future, it would be ideal to sample out and retrieve feedback on how useful the developed system was. We could use this analysis to improve, or reshape, the views that the users are looking to see and actually have. In addition, as more like tools arise, tests can be performed to identify which tools are the most useful in understanding moderately sized collections, and for what sort of questions they are most applicable.

References

[1] M. Kelly, M. L. Nelson, and M. C. Weigle, "The archival acid test: evaluating archive performance on advanced html and javascript," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE Press, 2014, pp. 25–28.

[2] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar, "Memento: Time travel for the web," *arXiv preprint arXiv:0911.1112*, 2009.

[3] M. Toyoda and M. Kitsuregawa, "A system for visualizing and analyzing the evolution of the web with a time series of graphs," in *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*. ACM, 2005, pp. 151–160.

[4] ——, "Extracting evolution of web communities from a series of web archives," in *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*. ACM, 2003, pp. 28–37.