Navtree

Design Bachelor Thesis
David Peláez Tamayo
Universidad de los Andes
Bogotá

For three months now, I've been working on my bachelor thesis at Universidad de los Andes. As the final project in my design education until now, it's been the longest academic project I've been commited too. The process has been truely iterative, improving in small steps and building from scratch each of the tools required for the final outcome. This document contains detailed information from the project, that I grouped using the subtitle and the contents table to make it easier to scan.

Initial Concept behind Navtree

The important part in education, knowledge and entertainment that the web has today has open internet as a field of study on it's own. The technologies to be developed, the standards to be improved, the categorization and organization of the content available and the filtering of such content to make it relevant to personal interests are among the many studies, experiments and developments solely related to the web.

Given such concept I've become interested in the central question that gave origin to this project -how do we reach contents?- in other word how do we get in time from URLa to URLb. Inside that question various opportunities, hypothesis and technical and visual challenger emerge and this project is in essence and exploration through design of how to answer that question and, more importantly, how the information generated to answer that question can be an opportunity to generate relevant information that is enriching for the surfers and other stakeholders.

The concept or triggering idea gave birth to some questions that even though are not to be resolved in this project explain the relevancy of Navtree. More advanced versions of the project could lead to more insight into the answers of the following questions:

Triggering Questions

- How do we get to certain content online?
- How do we obtain new content?
- How do we surf, what are the differences?
- Could we define a universe of content?
- Can I know the universes of content where I move myself?
- Is it possible to identify user niches according to their content consumption?

What's Navtree then?

The way a user browses online on his desktop (initially) can be recorded and represented as a set of trees (since some areas may not be connected at all, more than one tree appears) with different branches where each node is a url and the connections represent clicks to go from one node to the next. Such data set can then be understood as a directed(???) graph represented on a time line that is the starting point of this project, thus the codename Navigation Tree. We have then variables like time, location and precedence in the navigation tree for a given node. More variables like tags, ratings/likes, publication date and other semantic could be potentially added in the future. This additional variables are however way to complex to take into account in the original brief of this project.

This project then focuses around the exploration of appropriate ways to gather the dataset and then the visualization of it in order to create information relevant to the online community interested in understanding online user behavior and if possible to the user providing the data.

In summary Navtree is the data gathering tool that records a person's web browsing inside the Safari browser (as time, url and previous position -origin-) and the applications used to obtain information from that data, visualize patterns and enable its users to understand and/or make decisions when provided with multiple datasets to compare. Navtree was initially considered as a enriched version of the browser's history, a multidimensional representation that included not only the variable of time, but also the relationships between the recorded urls. However, currently the complexity of the dataset has made it harder to achieve that and the research side of the application has gained more importance. That's why Navtree is thought for researchers and computers analysts to whom the data can help make choices and broaden their understanding in their own fields.

Some particularities of the project

Thanks to the support and continuous assitance of design Prof. Santiago Barriga, Navtree has been growing not only as a software and design project, but also as a philosophical and personal one. You will see that the project takes into account the information design and technical requirements to solve the main brief, but also ethical and philosophical ideas related to the media (Ruby on rails and processing) it's being created on.

The constant support and real opennes of Santiago has allowed me to create a project with some personal elements strongly connected to my intentions as a designer and as a person. That being said I consider important to mention three key points related to that personal side of the project that are important to understand it and judge it inside it's reality. The three point are related to the language of the documentation, intellectual property issues and lastly the experimental process, learn by doing and uncertainty of Navtree:

First, the project is written in English with only some personal notes written in Spanish that are mostly temporary and scattered around the repository. The language selection was important because it has been my intention to allow the project to be shared outside the university borders if someone finds it interesting enough. Therefore the information and documentation available are in English in order to broaden the possible readers and truly documenting the project for sharing purposes.

Second, the project is licensed under creative commons, giving anyone access to the source code and documentation and allowing them to build upon, share and distribute in a non-commercial context. This was important to build transparency, since the Navtree gathers sensible information from particulars' web browsers that could create commercial value somewhere in the future. The value of data, the ownership of it and the security under which it's being used and stored couldn't be made clearer without the creative commons license. This of course means that Uniandes doesn't hold any right whatsoever regarding the intellectual property of Navtree, reinstaiting through the project other philosophical issues that I consider strongly attached to education and design: Transparency, openness, freedom of access and flow, and fair use.

The last important remark on the project has to do with its experimental nature and the process of design. Rather than pointing a brief or issue to solve and creatively build representations or concepts of such information, Navtree is a real software, information, interactive and interaction design project. In the process, many topics outside my limited knowledge have appeared -Math in general, computer simulated physics, data encoding issues, computing power restrictions, complex information design, etc.- that has made the process harder but also rich and, I must say, it has enabled inside me the so mentioned idea of designers as naturally interdisciplinary professionals. The project has evolved by doing and it's aim is not to create wireframes or concepts of the possible results as long as they can be explained and presented through smaller experiments with the real data. This of course means, that I have been committed to programming since the beginning and that has brought entirely different problems than if it were made only using representation of the product. The experimental part of the projects exists in the process and it's slow pace, but also in the uncertainty linked to it. I cannot be sure of the outcomes of the project. However I can make the proper software, sketches and analysis but that doesn't assure that relevant data will emerge or that beautiful pictures and interfaces will be the result. It's has been clear since the beginning that it's likely to find patterns and beauty and to shape information through the data, but the uncertainty of it has been documented with many "failed" screenshots of the process (these can be found in the Github repository). In the explanation of the process and the evaluation of the project I kindly ask the reader to consider the complexities attached to that, the time restrictions and, at the same time, the extra value that "doing the real thing" provides.

The aims and reach of the project

Initially there was a basic idea that the representation of the dataset as a tree or directed graph would provide enough information to consider the project complete. Far from that, the complexity and huge amount of data proved harder to manage. To explain the process more clearly, I would like to present you with the initial hypothesis regarding the project. The hypothesis are organized in order of complexity:

1. The NavTree can be visualized to provide of the surfer's consumption of online content through time in more complex and meaningful ways than existing linear browsing histories.

2. The use of various NavTrees from different users allows us to create visualizations that represent different styles of web browsing in the form of patterns deriving from the basic variables considered in the dataset gathering.

3. The inclusion of the previously mentioned more complex variables, specially semantic ones, could enable the complexity of the visualization to increase, making visible universes of content through which the users moves. The relation between those universes could be made visible or (even more interesting) the separation of them inside a hypothetical online endless semantic universe may appear.

4. The ability to record from numerous surfers NavTrees can help gain insights on content consumption trends useful from a sociological perspective and for user research.

With all those hypothesis in mind, you should by now understand more clearly the uncertainty that was mentioned before. To clearly define the project, I used the concept, explain the technical feasibility of building the dataset and decided that at least hypothesis number 1 should be completed. The hypothesis strongly simplifies the project, reducing it to the data gathering tools and the representation of the data as to provide more meaningful ways to visualize browsing history and taking into account only an specialized target and leaving the user source of the data only as a potential user of the application, rather tan its main target.

Limiting the target, and understanding the different requirement between browser and researcher puts a hole in the project that I want to make clear right now: It removes any incentive from the user to provide the information besides the good will of cooperating with the project. In an ideal way, a more advanced Navtree would prove useful for the user, like a GPS tracking it's position on the web as an emotional recording of the personal journey across the internet. That emotional component would also make a nice feature for future browsers and could engage the users to record their data, creating as a by product the research material. I have taken seriously into consideration the commercial and practical implications of the project, but the current state of it makes more important to focus on obtaining information from the dataset using target users that can handle more complexity that a regular person with Safari. This is a way to accept the complexity of the end user and be real to the time constraints of the academic semester.

What has been completed until today

As of today, november 7th 2011, the following thing have been completed:

1. The documentation has been published and everything is localted in a single Github repository at http://github.com/davidpelaez/navtree
2. The website http://growanavtree.com was created. It's still not complete visually, but the server behind it successfully records browsing information. Currently

only my personal browsing information is there with a total of 6360 unique URLs recorded and 13108 connections between them.

3. The structure of the dataset was defined: It's a database logical structure that records relevant information from the browser like the url, the time, the previous url, weather that url appears in a new window or is a tab in the background, etc. Understanding what information the Safari browser extension could provide explains clearly what variables can be used to create the visualizations, you will se more about this variables and the components of the dataset in the design section.
4. The server uses a Safari extension that I created from scratch that sends the information in the background without the user noticing. A private browsing mode that temporarily disables the URL syncing was created to ensure that the extension adapts to the minimum user needs to give him/her a sense of control.
5. Recently two more people have been asked to install the extension and enrich the database, to allow visualization from different datasets to be presented to the jury. They should provide at least 1000 connected nodes in the following three weeks.
6. A research of visual references and related work has been made, nothing closely similar was found. Several online documentations regarding the topics of directed graphs and data visualization were used. The book Visualizing Data: Exploring and Explaining Data with the Processing Environment by Ben Fry has been extremely useful in the process.
7. Some sketches have been made trying to represent the data. You will see more detail regarding this in the section dedicated to the process regarding information design. This processing sketches have been done in two directions: Some of them are representations of the data, the others are sketches integrating processing libraries to create a space that could be navigable (pan & zoom) because the size of the generated images is too big. These last sketches have to be adapted from Processing to Eclipse IDE because of the growing complexity of the applications.
8. Once the complexity of the actual processing sketches emerged, I decided to define what my final outcomes would be visually by the final presentation. Defining these has been important to close the project with structure and order regardless of its uncertain results. You will see more information about this in the design section.
9. Finally, some details about the contents and important elements to communicate the project are appeared during the meetings. You will see thoughts on that matter by the end of the document, on what to expect for the final presentation.

What still has to be completed before the presentation

As of today, november 7th 2011, the following thing have not been completed. However they will most likely be finished by the date of the final presentation in around three weeks.

1. The design of the website as the homepage of the project will be finished, opening the possibility of people not related at all to the project to understand it in a simple way and be willing to record their data.
2. Currently the Safari extension has some usability issues and the unfinished state of the website doesn't contribute, these should be fixed specifically in three points: a) The extension will be capable of knowing that it's missing the

public key required to send information to the server without revealing the user identity. This will simplify the extension setup. b) The private browsing mode and the feedback to the user indicating how much and what information has been sent will be completed. It'll be a simple floating screen that the user can toggle to shift between the browsing methods (private or recording) and see the most recent synced URLs with the Navtree Server. c) Privacy filter will be added. The user will be capable of registering his name, usernames, nicknames and some urls that he/she want to be ofuscated from the url. This will allow the user to keep his information private. For example, the user could add these wors: [david, pelaez89, xtube, pelaez, d_pelaez, pornhub] to filter all urls containing those words and keep the private while still recording the node in the database.

3. The current graphs are not interactive, therefore I will include some basic interactivity to them to allow the user to get further information from that he has on the screen. The most important of this addition is the textual information connected to mouse events, since currently the user cannot get more detailed information about specific points of interest in the displaying image.
4. As of today, the nodes only use color to differentiate them by depth in the tree. A root node (a url that the user typed and the address bar that wasn't reached by clicking a link) has a different color that a node that the user visited after clicking a url on that root (that's be level 1, and root depth 0). As you will see in the design section, more variables can be included in the graphics. Therefore by the final presentation other variables will be included in the graphics to relate the complexity of the dataset with several visual variables like type of line, node size, color or distance from some reference.
5. Connected to the addition of visual variables, I expect to finish three or four visualizations for the final presentation:

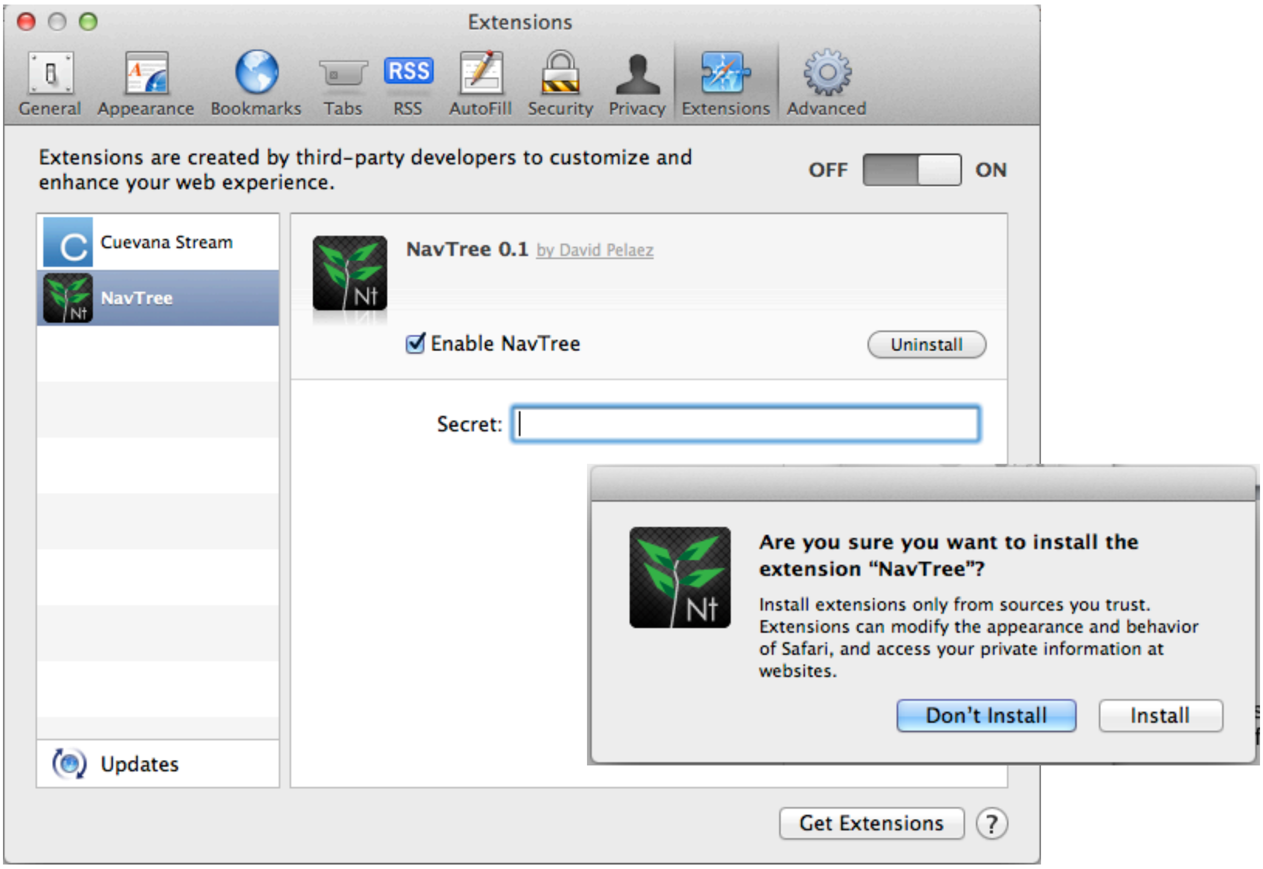
    a. **Linear time vs node depth:** In this graph the time grows as a line and the nodes appear over that line depending on when they were visited. This graph should contain no more that 4000 nodes (One week for a heavy internet user). This is a improvement on the current tree visualization (images in the design section).
    b. **Radial time:** This image only takes into account time that grows radially from a center. This experiment intends to show that there are patterns in the density of the rings that grow from the oldest point in the dataset as the center of the image. This is an experiment yet to be done.
    c. **Textual/Numerical statistics:** Basic information like the number of nodes being displayed, the average branching factor of the tree, the size of the dataset, the time span being visualized, the most visited domains, the most visited node, the average node density at any moment, etc. All those numerical and textual data will be included as a single visualization that also provide information that is harder to show in other ways.
    d. **Intensity wave:** Very similar to a sound wave in it's digital representation, I will represent every node as a bar placed on a line of time. The height of the bar will be related to the depth of the node. This will allows us to see the differences in time of a given dataset, similar a heartbeat line graph or a seismograph. The hypothesis in regard to information design, is that different kinds of users will have totally different intensity waves in their browsing.

The previous 5 points can be summarized as: Visual design of the website, usability of the extension, interactivity of the sketches, addition of visual variables and set of final visualization.

What has been tried so far regarding design

In this section you will se mostly pictures with some information regarding details about design. I hope to show what my understanding regarding information design (and some other minor details in other areas) has been so far in order to make clearer how this one is deeply a design matter more than a technical one.

Let’s start by the most trivial elements, the Safari extension and the website.
The extension is installed and after that the key of the user has to be added inside preferences to begin syncing. That’s all that required. The logo was created but as mentioned before, usability issues are to be fixed.



Extensions are created by third-party developers to customize and enhance your web experience.

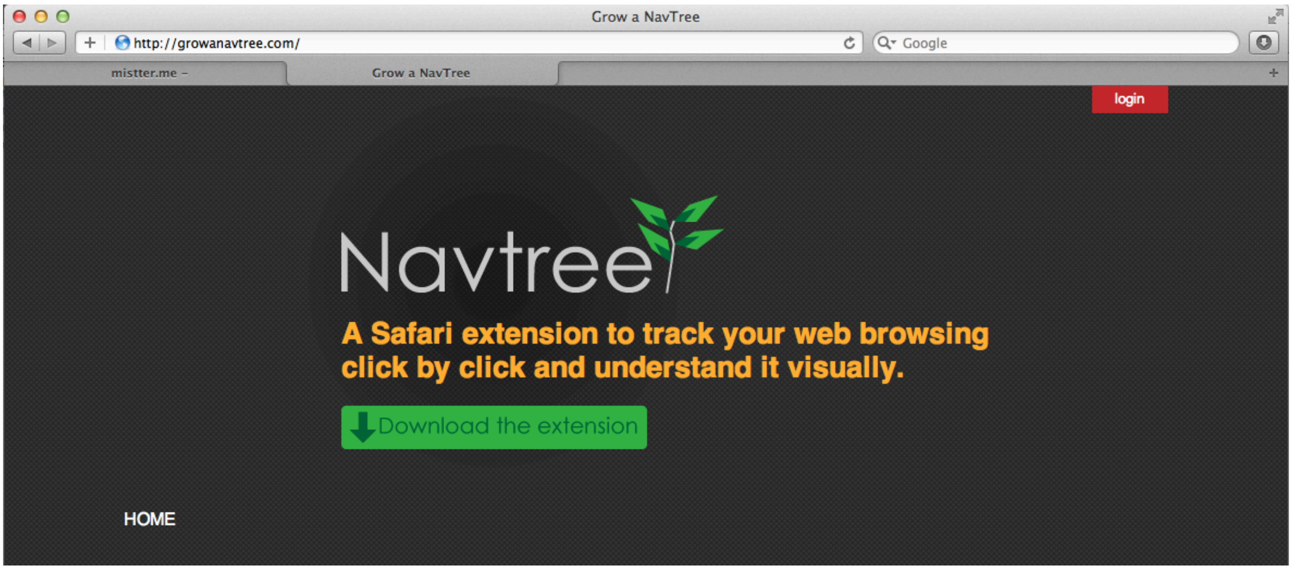
NavTree 0.1 by David Pelaez

☑ Enable NavTree

Secret: |

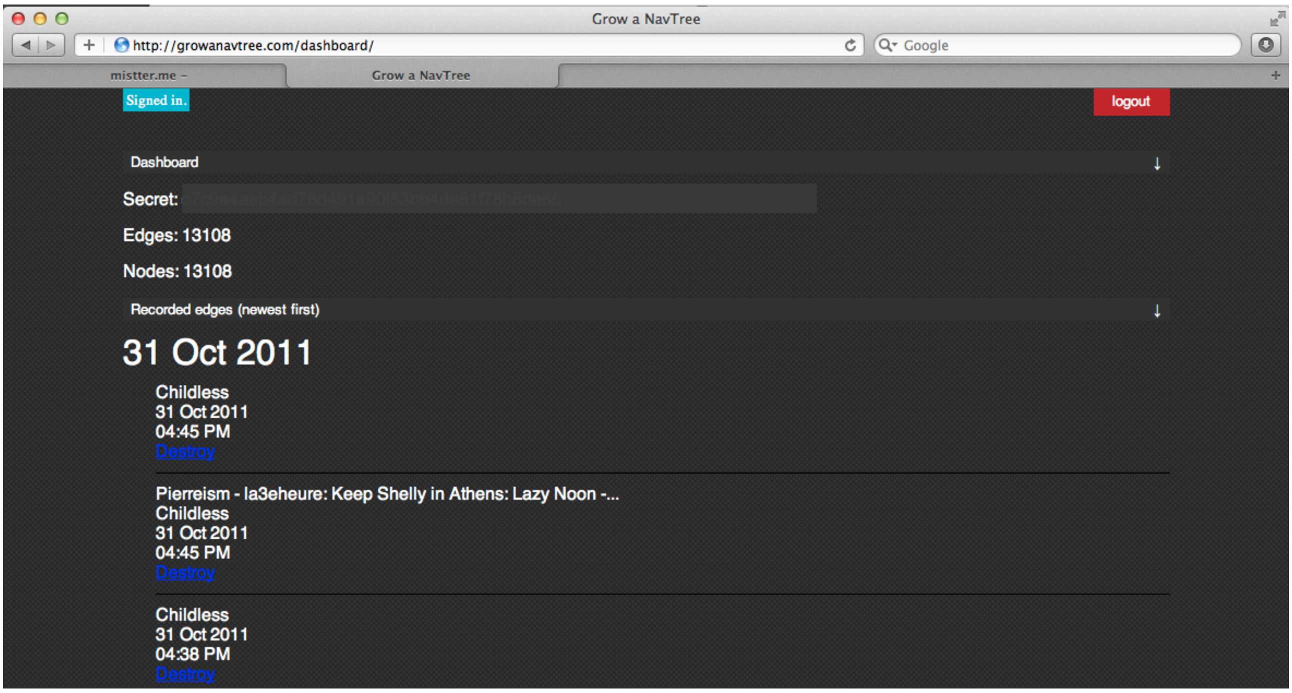
Are you sure you want to install the extension “NavTree”?

Install extensions only from sources you trust. Extensions can modify the appearance and behavior of Safari, and access your private information at websites.



A Safari extension to track your web browsing click by click and understand it visually.

Currently the homepage and the dashboard of the web server is pretty bad, provides almost no information and has to be fixed. However, the server works ok and as the screenshot below indicates, the date of the nodes appear and a list of all nodes can be seen page by page. There you can see that 13108 connections are currently recorder in my account.



Secret: [redacted]

Edges: 13108

Nodes: 13108

Recorded edges (newest first)

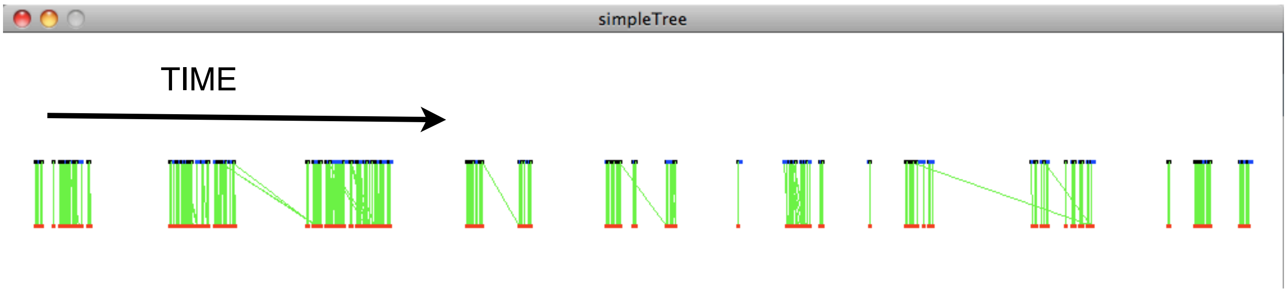
31 Oct 2011

Childless
31 Oct 2011
04:45 PM
Destroy

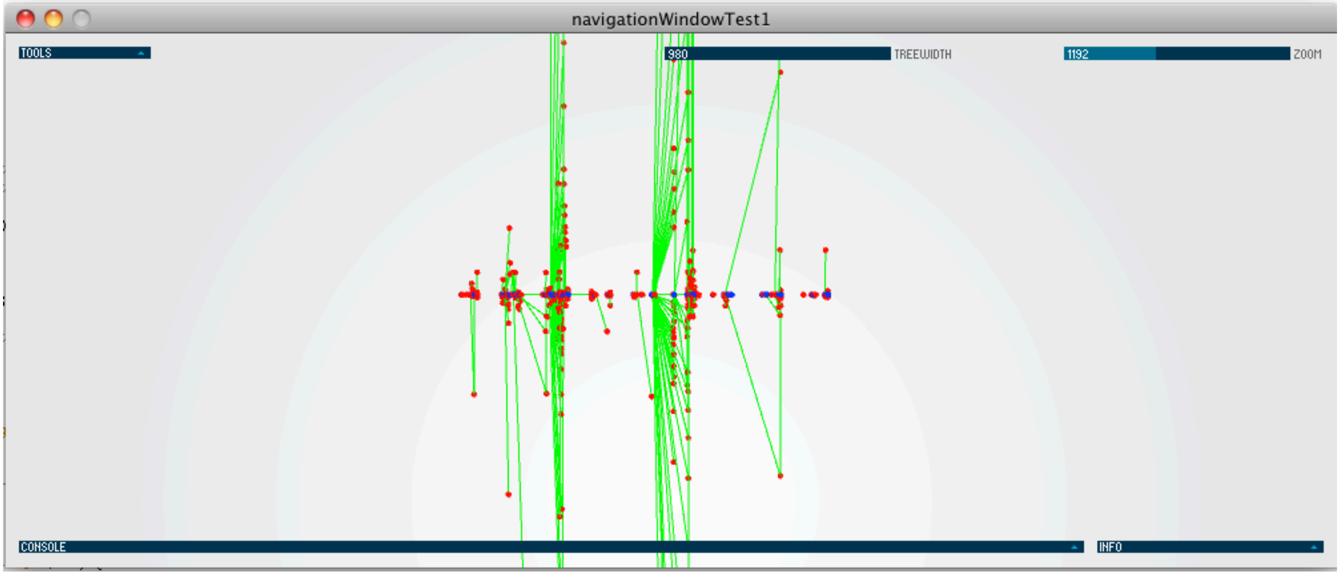
Pierreism - la3eheure: Keep Shelly in Athens: Lazy Noon -...
Childless
31 Oct 2011
04:45 PM
Destroy

Childless
31 Oct 2011
04:38 PM
Destroy

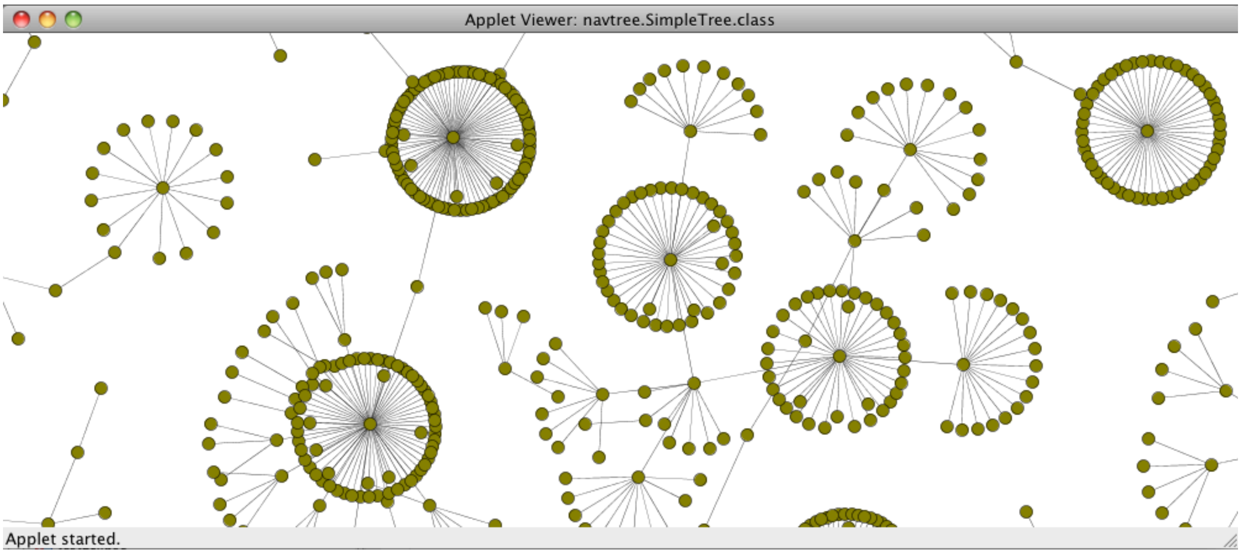
Regarding the visualizations that have been created so far, I have images that show the progress of what has been done. Most of the images shown use 4000 connections from the information that I recorded of my own browsing using the extension. In most graphs all the root nodes (those without a parent which url was typed by the user on the address bar) are on the center of the screen creating the darker line.



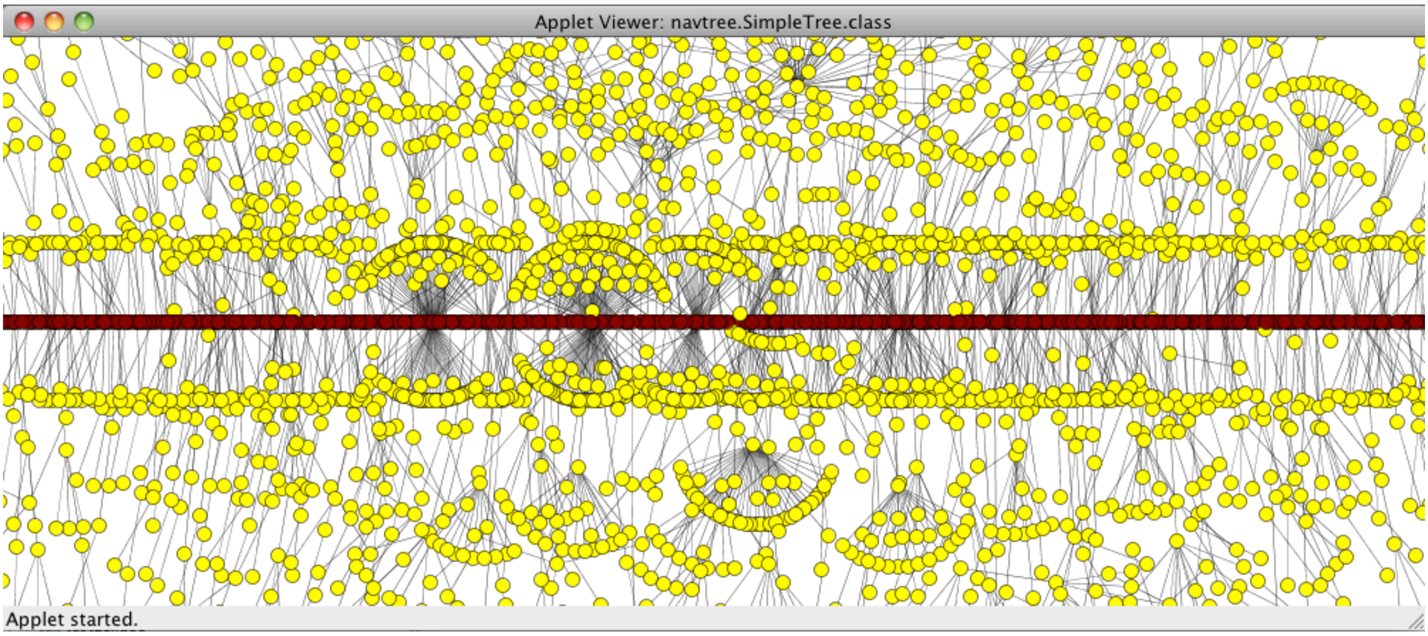
The image above if the first image that I created. Roots are blue and all deeper nodes are red, green lines connect the nodes. Even though this image is very poor, the time axis goes from left to right and it's clear that the wholes between big blocks of nodes are nights or other times where almost no browsing was made. I checked and the first three block are days of the week, for example: Wednesday, thursday, friday.



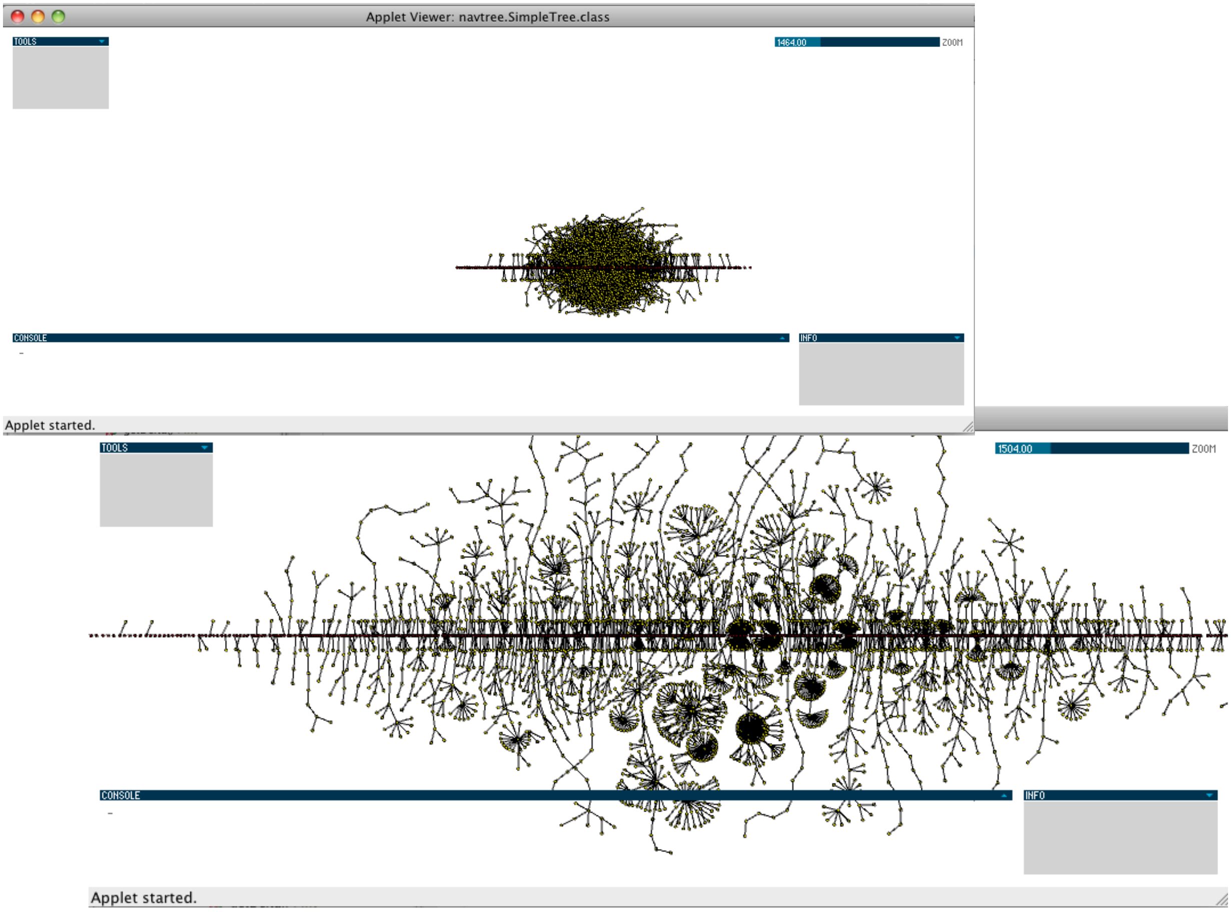
This improvement over the initial poor picture tried to improve in creating an equally distributed time line (again from left to right on X) but separating the nodes on Y according to it's depth. The graph is about 4500px wide, therefore I created the space in the background that allowed me to zoom and pan using PEasycam and ControlP5, both Processing libraries that I've been using. At this moment I began to more deeply reserach about mathematics to create this networks and found out the book by Ben Fry. I had to move to Eclipse IDE to continue the development and I had to rewrite some of the code for that environment.



Using the examples in the book by Ben Fry and already working in Eclipse, I was capable to see something more like a tree where the relations between nodes and their parents was visible. Here, the nodes try to move away from other nodes until they become stable. The big circles show that from one specific tab many links in the background were opened. This shows initial information about the behavior of the user that shouldn't appear in many other datasets because opening content in the background is not common to all surfers. In this graph only ancestry of the nodes was added an no further variables were included. For instance, there's nothing indicating depth, time and a total lack of textual information and interactivity or responsiveness from the application.



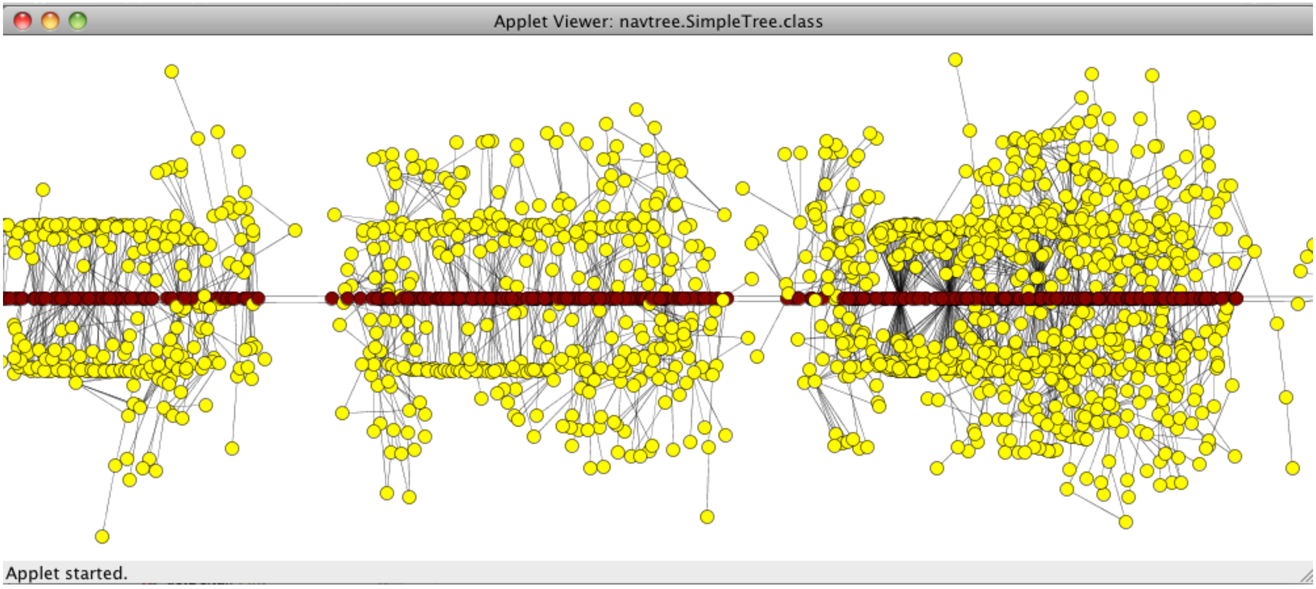
In the image above, root node are constrained to the center of the screen and color are added. This is an intent to add color differentiation and hierarchy, but it's for from a finished result. Here again, node respect time partially since they are attached to the roots that are distributed by time from left to right, however a node can be place more to the left than its roots, breaking the time line since a node cannot appear before it's parent.



Applet started.

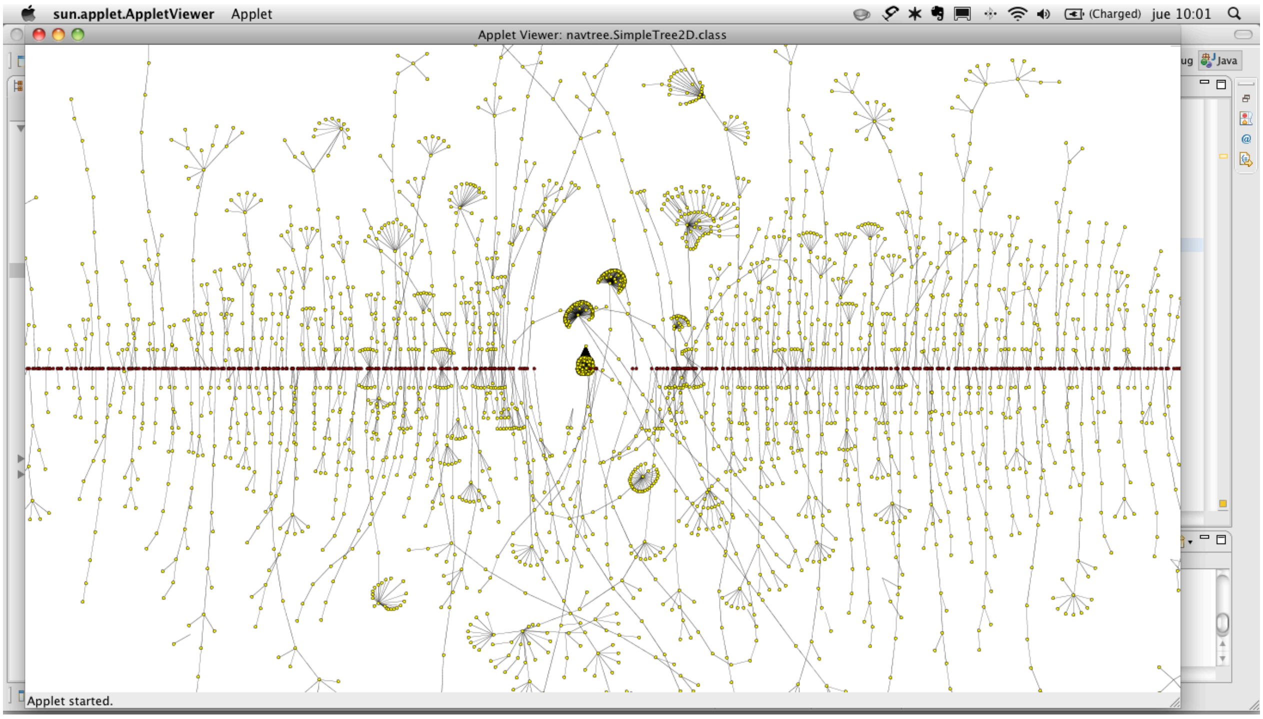
Applet started.

In the first two images the evolution from chaos to something clearer is visible. The tree is cleaner since more constraints have been added and roots without children are not being mapped. The long lines show very deep navigation path: Going to a website, click a link inside that page, click a link inside that page and so on...

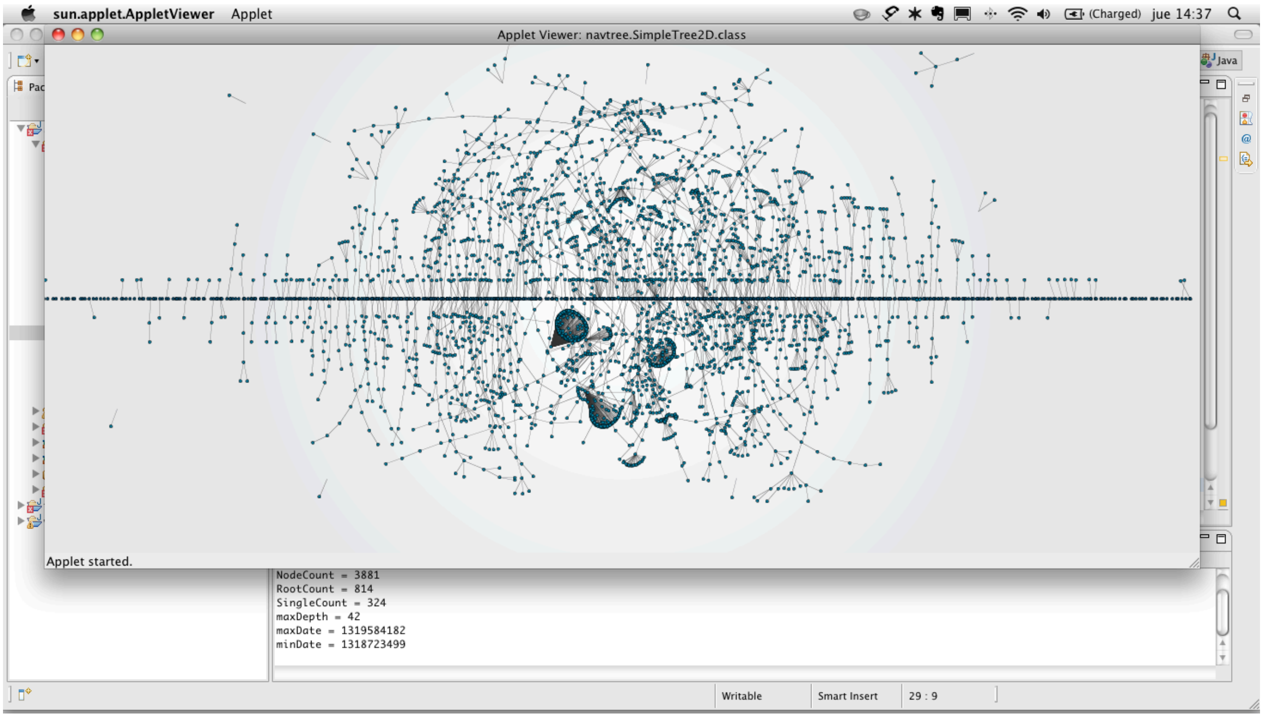


Applet started.

Since all the graphs so far are time vs depth, time was better mapped in the image with the yellow nodes where you can see wholes in the root's line, those are moment where almost no browsing was made.



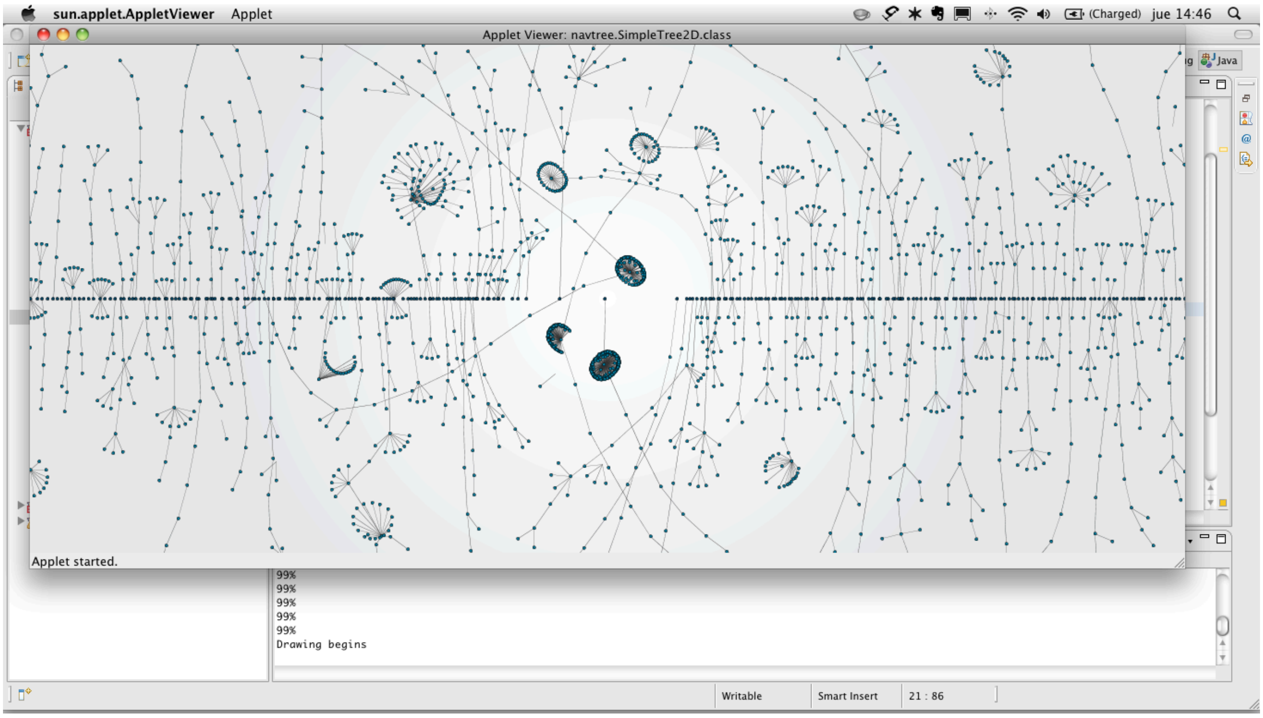
Given the limits in computational power, I decided to start mapping 10 days with around 3000 nodes. In this graph the lines go a lot to the top and bottom of the screen, that's because there are many long lines of browsing with a lot of depth. This same graph compared to a user with low use of tabs and background link opening should reveal that there are almost no lines going to the bottom or top of the screen. This sort of comparisons show how the sketch is beginning to approach the generation of more relevant information for comparison when an analyst uses the Navtree app.



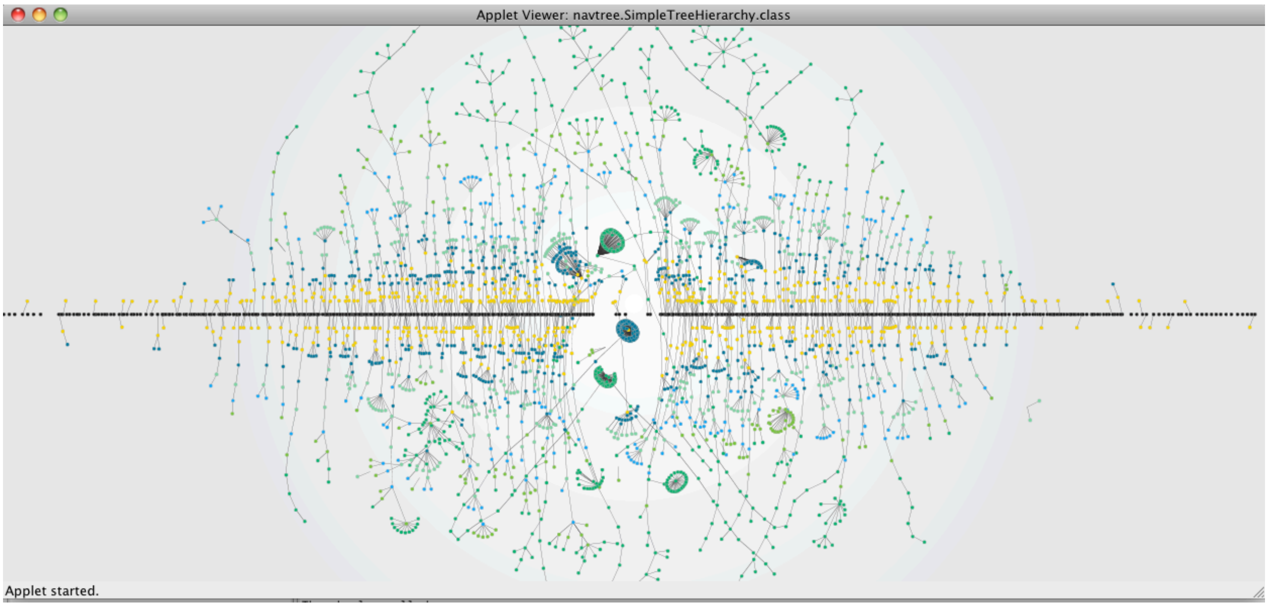
```
NodeCount = 3881
RootCount = 814
SingleCount = 324
maxDepth = 42
maxDate = 1319584182
minDate = 1318723499
```

The image above and below are the same sketch. In this sketch and evolutive algorith was used to represent every node and a line connecting it to its root. This graph has the pecualirity of being processed 1000 times before appearing on the screen, that means that the graph evolves before being shown to reduce the notion of chaos.

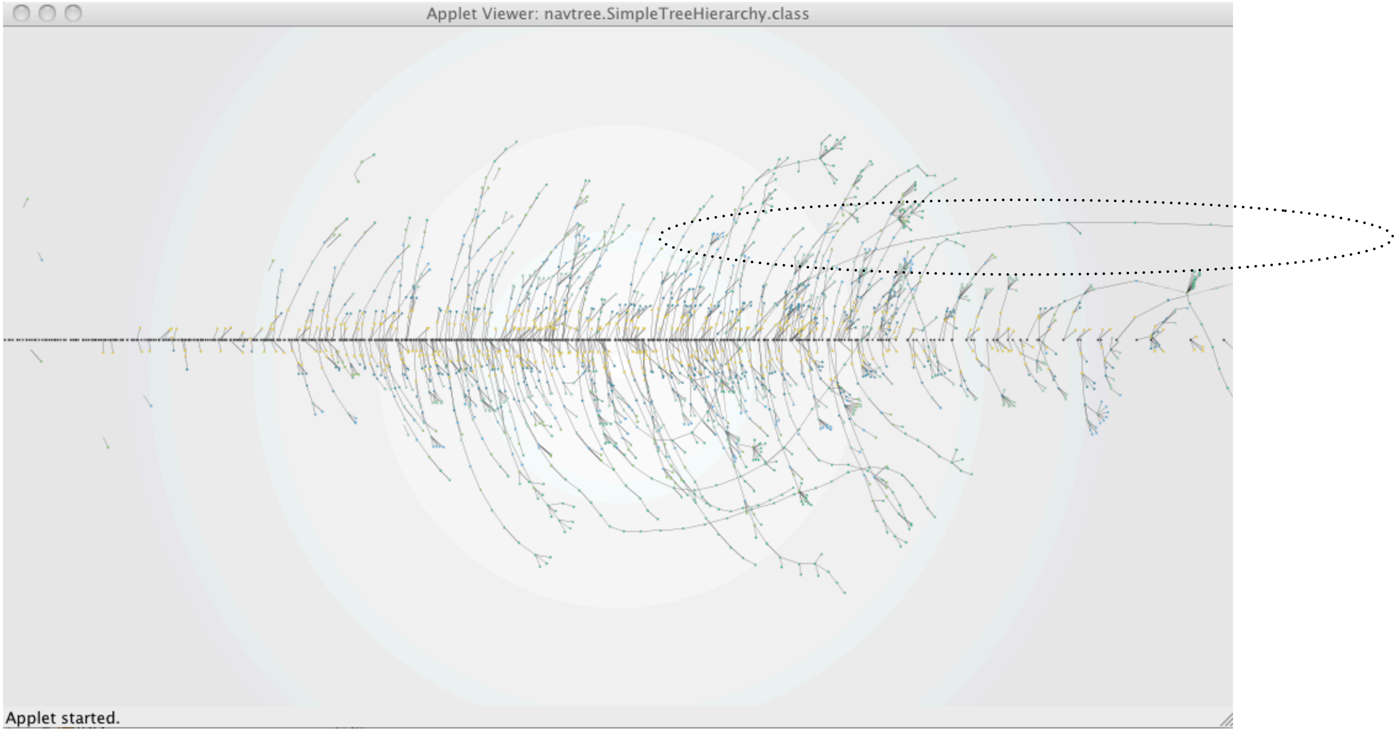
The graph below had been running for longer than the one above, that's why there not so much chaos. Notice that nodes move to the right and left, not respecting the axis of time.



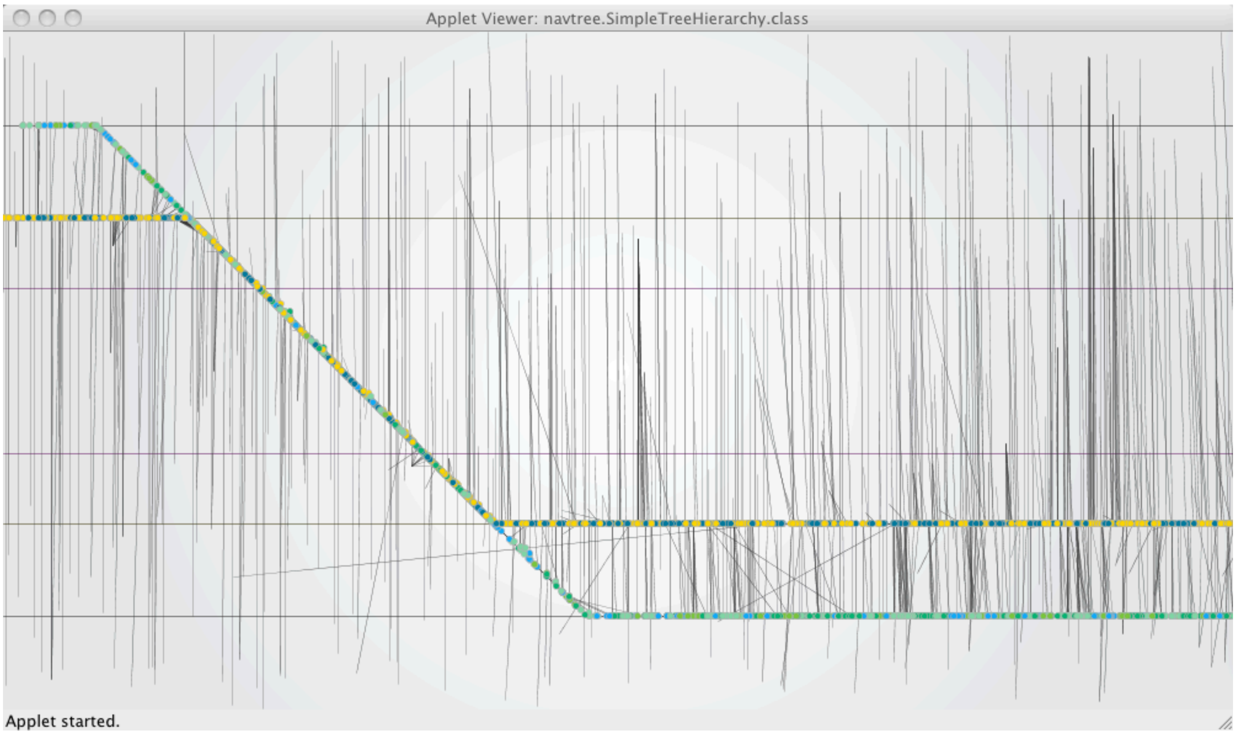
```
99%
99%
99%
99%
99%
Drawing begins
```



In the quest of building meaning to the viewer, I added color to the nodes according the the depth. You can see the roots line, then a yellow area and then other color (soft blues and greens). It's been part of the process to find the right colors to generate contrast and enrich the visualization from an information point of view. This graph show the data as some sort of living organism and maybe the simple option to compare different organisms can prove to be a strong enough source of information to the viewer once he knows what the structure of the graph is.



This visualization shows all lines leaning towars the right of the screen since it respects partially time in the X axis: Every node if more to the right that it parent. Notice around the dotted ellipse a big line going out of the screen. That line has more than 15 nodes which is very interesting because shows a great deal of continuity in the browsing. When textual information and responsiveness is added to the nodes, the viewer will be capable to see more information about those nodes and gain further understanding about that branch.



Many mistakes have happened in the process too. This one shows a sketch gone bad, I was trying to more strongly create lines to the nodes in different depths, but the connections between them just failed. I was unable to find the mistake and had to start from scratch.

The last element that I'd like to explain in detail is the structure of the dataset. Creating the dataset was contrained by the information that the Safari extension could obtain from the browser. I want to mention as a simple list, because the variables that every node has can be directly correlated to visual variables in the final designs. In the following table you will see what information is gathered and what possible variables I have taken into account that can be related when building the representations:

Dataset Variable / Info	Possible visual elements
• Is Node a root? • Node depth. • Was opened in background? • Is new window? • Is blank or has an URL? • Node's parent. • Date & Time. • Is direct child or a child in the background? • How many node siblings?	• Color of node. • Size of node. • Shape of node. • Type of line (solid, dotted, dashed). • Node position on X & Y. • Direction of time.

I hope that making concrete the list of variables that I've encountered you will recognize the design process as a filtering and correlating in meaningful ways, and more clearly show the information component of the project.

Restrictions found and other issues

I'm using an Unibody Macbook with 4Gb of RAM and a 2.4 Ghz Intel Core 2 Duo, it's not that hard to hit the limits in memory of the computer or have an unresponsive user interface. I expect to test the final application in a more powerful computer, however that restrictions have also proven that with today access to computer power by the average user, it's unlikely to provide an end user version of Navtree that can successfully run on an average computer while showing as much information as available.

To be more concrete, I reached the 10000 connections a couple of weeks ago, however it's been impossible to graph more than 4000 of them at the same time and 2500 remain to be an ideal number to still have enough computer power to add interactivity to the software. That means that it's hard to view at the same time a big amount of information, once again implying the need for a more specialized target and limiting my personal dream of having a browser history more advanced for common users.

The tools used have also presented some challenges. I have to choose between using Processing (that I already knew) or C++ with OpenFrameworks. The former options is simpler and faster to program, however theres a trade off in comparison to the latter: Speed. The development was already complex enough to change to a language that is unfamiliar too me and more complex. In the other hand, choosing Java made me changed from the comfortable processing environment to a professional IDE, Eclipse, and I also had to readapt and even rewrite some of the code to more adequately fit the new environment. There's one last thing related to the language selected that has to do with the reach of the project and it's nature. Being strictly web related, Java makes it easier to move the software from the desktop to the browser as a Java Applet. I even considered the Javascript of Processing as an option to embed the application as part of the Safari extension to really use the extension to add features to the browsers besides the syncing capabilities already built in it.

Other issues have to do with my current knowledge of mathematics. Dealing with creating the right scales to represent time, all the formats and frameworks to create the applications and the complexities involved in generative data visualization hasn't been trivial. Even though the book Visualizing Data by Ben Fry has been of great help, some of the algorithms included are not totally clear to me, an understanding and adapting them has been a great challenge. Given that mathematical complexity some simple problems, like making the representation of a time lime proportionate, remain unsolved. I have made approximations and they should be good enough to contain enough value to work as validating prototypes. Nevertheless, such "malfunctions" must be acknowledged in this document for anyone interested to be able to improve them.

What it will be in 4 weeks: Expected results

The final presentation with the juries has produces an unsettling feeling for the last weeks. Personally, I think the project is very clear and well defined, but communicating is a harder task. Linked to that you already know from the beginning of this document that this is an experimental project and that the time required to develop it is an important factor when measuring results in a limited time scope. In that order, please don't expect a finished application for the final presentation, but a working simple prototype with the projected visualizations.

I expect to show graphics using at least three datasets from three different users in the hope of evidencing differences in the nature of the graphs to prove the first and central hypothesis: "The NavTree can be visualized to provide of the surfer's consumption of online content through time in more complex and meaningful ways than existing linear browsing histories."

I will also present the restrictions that the software currently has, the hardware limitation that we face today and the complexity issues that I dealt with during the project. In connection to that I will explain how the project was more clearly scoped towards an academic target rather than a regular web surfer.

As I mentioned before, changing the target to a group that could handle more complexity to simplify the creation of the graphics creates an impact on the motivations of a common person to record the browsing data. I will therefore speak about the concept in a broader sense, what lead me to create this projects and what could be expected in the future if this project were to be continued.

In summary I will present these points:

- The triggering idea and how it was defined as a project.
- The problems faced and how the project had to be narrowed and the implications
- The results as a set of projected visualizations in a working app that is a medium resolution prototype.
- The results will include data from at least 2 users, but three is expected.
- I will speak about the current problems that are yet to be solved and the hardware constraints that I faced.
- I will close explaining the prospective achievement if the project were to be continued and what projects like these can mean for design research, human understanding and it philosophical impact to sustain the relevancy the project has.