DEVELEOPMENT OF A VISUALIZATION FOR TRACKING WEB TRAFFIC IN CONNECT AMERICAS.

Gustavo A. Peña Morales
Jorge Hernandez.

Abstract: As social networks keep growing at such fast rate, the traffic of data that they have become more valuable every day. Connect Americas is a company who connect companies around Latin-American and the Caribbean, and has a very powerful database who needs to be unified and easier to understand. In this project we try to exploit this dataset to increase the knowledge of how the users behave and what they are interested so that Connect Americas can improve their experiences.

INTRODUCTION

Arguably the analysis and study of data is the most important trend right now in technology, the amount of data that is available because of web growth in the last two decades has been unexpected and very useful. This data is the biggest asset companies have nowadays. Data can help companies predict new trends, find ways to be more persuasive and get to know clients and users faster than ever. This data growth mixed with the advances in computer science and computer graphics are the main reason why visualization is one of the most helpful tools to understand, analyze and present the data that companies have.

ConnectAmericas is a social network that have a great amount of data from clients across the world, but doesn't have a visualization system that can help them reduce and unify the data that they have, in order to get the most out of it. In this project we will study the different datasets that ConnectAmericas have and create a visualization that allows them to understand what the users of the platform do, and what are the more important trends inside the social network.

DATA CHARACTERIZATION

ConnectAmericas is the biggest social network for businesses in Latin-America and the Caribbean, have 70000 registered users and have different information about this users, this information is what forms the dataset of the platform. At the moment ConnectAmericas counts with two main databases:

- Drupal: A MYSQL database on drupal that have the information of each of the users that are registered in the platform.
- MongoDB: A database that have the navigation records of the users inside the contents of the platform.

These databases are stored in tables with quantitative and qualitative attributes and items, the items are the users. These tables are the primary form of the data, in addition to this the dataset types have cluster classification in which the data is separated according to the contents of the platform. So that the information of each action of any user is stored and classified according to the content related and not save as a new item.

Having this information, the company wants to create a visualization tool, that is able to show the navigation patterns of the users inside the platform and be able to see what are the biggest and more trending contents of the platform. Discovering performance indicators for the departments of management of the company.

STATE OF THE ART

There has been many research and works related to the use of visualizations that can help show the tracking of the user's navigation and kpi presentation. Here we present the articles and papers that we found more helpful with our research.

Applications of Data Mining Techniques to Identify Relevant Key Performance Indicators. Jesus Peral, Alejandro Maté, Manuel Marco. (2014):

Nowadays the most common tool to monitor business performance are Dashboards. This include visualizations of many techniques, where we can find Key Performance Indicators (KPI), which are very important information in order to compare the current performance of many businesses, but there are many problems defining which KPI's are relevant. In this paper they explain different techniques using Data Mining to obtain specific KPI's for semi-automated objectives in a business, this approach is better because doesn't use existing KPI or test them over cycle and rather use existing data.

Visualizing Key Performance Indicators Using GKPI Procedure. Brian Varney (2010):

This paper presents the new GKPI procedure, which can be used to create graphical key performance indicators charts including new marks and channels which can help identify these indicators. This paper is an introduction to the integration of the GKPI procedure to the SAS environments.

Key Performance Indicators: Developing, Implementing and Using Winning KPI's. David Parmenter (2007).

In this book the goal intended is to help minimize the risks that working with a KPI and balanced scorecard projects have. It is designed for the project team, senior management, external advisors and team coordinators whose role is to succeed with this projects. In this book it will be explain the most efficient ways to obtain KPI's according to the needs of the projects and how to implement them in order to achieve the success that could have a legacy inside an organization.

Visualizing Web Navigation Data with Polygon Graphs. Jiyang Chen, Tong Zheng, William Thorne, Daniel Huntley, Osmar Zaïane and Randy Goebel.

Data Visualization is connected with the world of machine learning in order draw the inferences from large datasets located in WWW. In this paper they describe the process of creation of a WebViz system, that allows to use visualizations in order to have better navigations using data mining techniques. They created a Polygon Graph which is a tool who helps to discover knowledge patterns of implicit relations among different data variables.

CZWeb: Fish Eye Views for Vizualizing the World Wide Web. Brian Fisher, Makrina Agelidis, John Dill, Paul Tan, Gerald Collaud, Chris Jones.

With the growth of the interconnected information in the web, is necessary to find new forms to organize these connections, that's why it is implemented CZWEb a system that traces the user's web paths of the using a Fish Eyes technique. A technique that allows the user to zoom different parts of the map that the user is creating in order to see in more detail the path of navigation. To create this maps the system include clustering techniques and creates a network that interconnect nodes.

Evaluating Social Navigation Visualization in Online Geographic Maps. Yuet Ling Wong, Jieqiong Zhao, Niklas Elmqvist.

"Social navigation enables emergent collaboration between independent collaborators by exposing the behavior of each individual. This is a powerful idea for web-based visualization, where the work of one user can inform other users interacting with the same visualization. We present results from a crowdsourced user study evaluating the value of such social navigation cues for a geographic map service. Our results show significantly improved performance for participants who interacted with the map when the visual footprints of previous users were visible."

Model Based Clustering and Visualization of Navigation Patterns on a Website. Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, Steven White.

They present a new methodology for exploring and analyze the navigation patterns on a web site. They analyze the patterns of sequences of URL categories that the users explore, where they placed similar paths into a same cluster and then they present the users that are inside that cluster. The cluster approach is a model-based and portions the users in the way the requested the URL, with this cluster technique they arrange to explore thousands of items and they visualize in a system that they called WEBCanvas.

Visualizing and Discovering Web Navigational Patterns. Jiyang Chen, Lisheng Sun, Osmar Zaïane and Randy Goebel.

"Web site structures are complex to analyze. Cross-referencing the web structure with navigational behaviour adds to the complexity of the analysis. However, this convoluted

analysis is necessary to discover useful patterns and understand the navigational behaviour of web site visitors, whether to improve web site structures, provide intelligent on-line tools or offer support to human decision makers. Moreover, interactive investigation of web access logs is often desired since it allows ad hoc discovery and examination of patterns not a priori known. Various visualization tools have been provided for this task but they often lack the functionality to conveniently generate new patterns. In this paper we propose a visualization tool to visualize web graphs, representations of web structure overlaid with information and pattern tiers. We also propose a web graph algebra to manipulate and combine web graphs and their layers in order to discover new patterns in an ad hoc manner."

WebOFDAV – Navigating and Visualizing the Web Online with Animated Context Swapping. Mao Lin Huang, Peter Eades, Robert F. Cohen

"This paper presents a novel navigation approach that helps the user, not only by providing a visual aid to guide the Web journey, but also by preserving the user's mental map of the view while the user interactively navigates the Web by swapping of views. This approach does not predefine the geometry of whole visualization at once; instead it incrementally calculates and maintains a small local visualization on-line corresponding to the change of the user's focus. This feature enables the user to explore the Webspace without requiring the whole Web graph to be known."

Interactive Visualization. Bill Ferster.

This book present different techniques to create interactive visualizations that can solve questions and create knowledge about the dataset presented. In this book the author shows the importance of knowing what is the type of data that it will be used and how to approach the best solution for the problems that this data present.

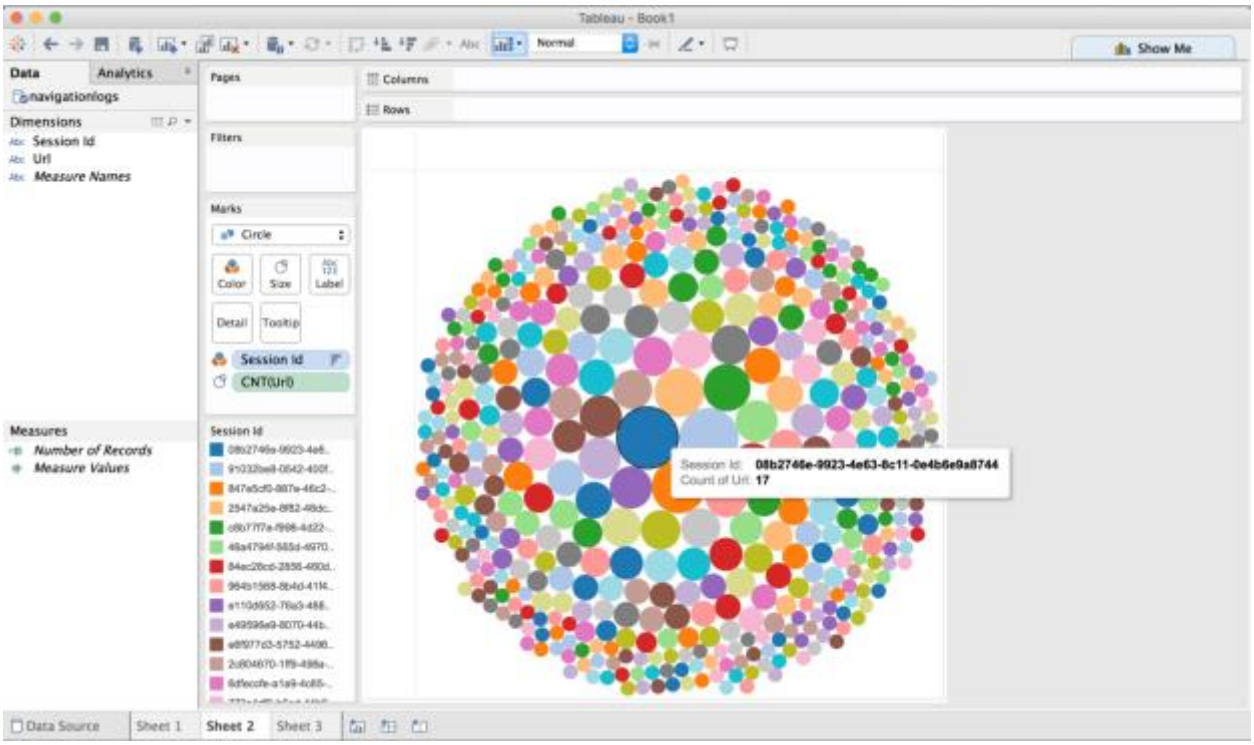
SOLUTION

According to what was talked with the company and feedback from a meeting with experts it was concluded that the solution most suited for the problem was this a two window interface with a treemap as a main view and as a secondary view an icicle with linear charts and a statistical box.

It was identified three main tasks that the visualization tool must do:

1. Present the distribution of the users navigation on ConnectAmericas.
2. Derive the connections that users have in common while exploring the content.
3. Explore the characteristics of navigation, to determine the patterns of use.

According to what was talked with the company and experts, these are the most important tasks that need to be solved with the implementation of the visualization. With this in mind the first idea of visualization was a bubble chart.



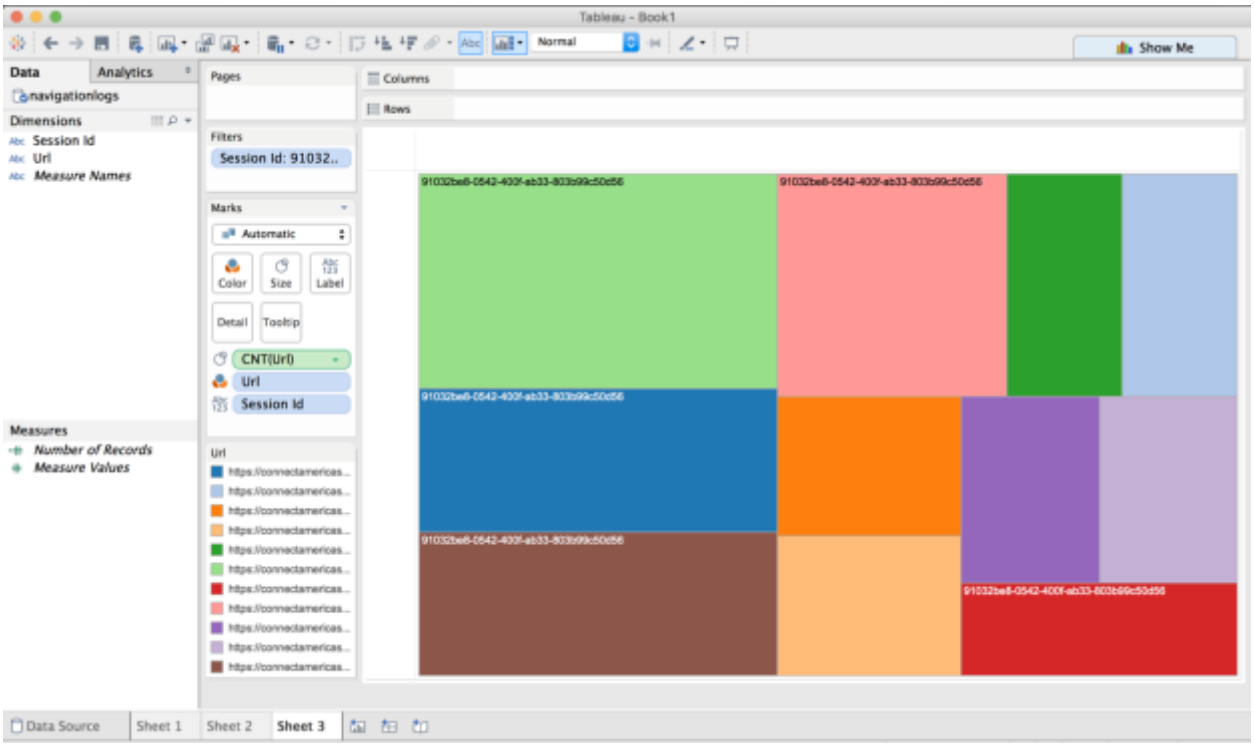
A bubble chart where the colors will identify the different contents of the platform, also the size of these bubbles will show what is the most viewed content of the platform.

The problem with this visualization is that the bubbles don't allow to have the clearest visualization about what is the bigger size, also didn't allow to infer more information about what was the reason why it was viewed that content.

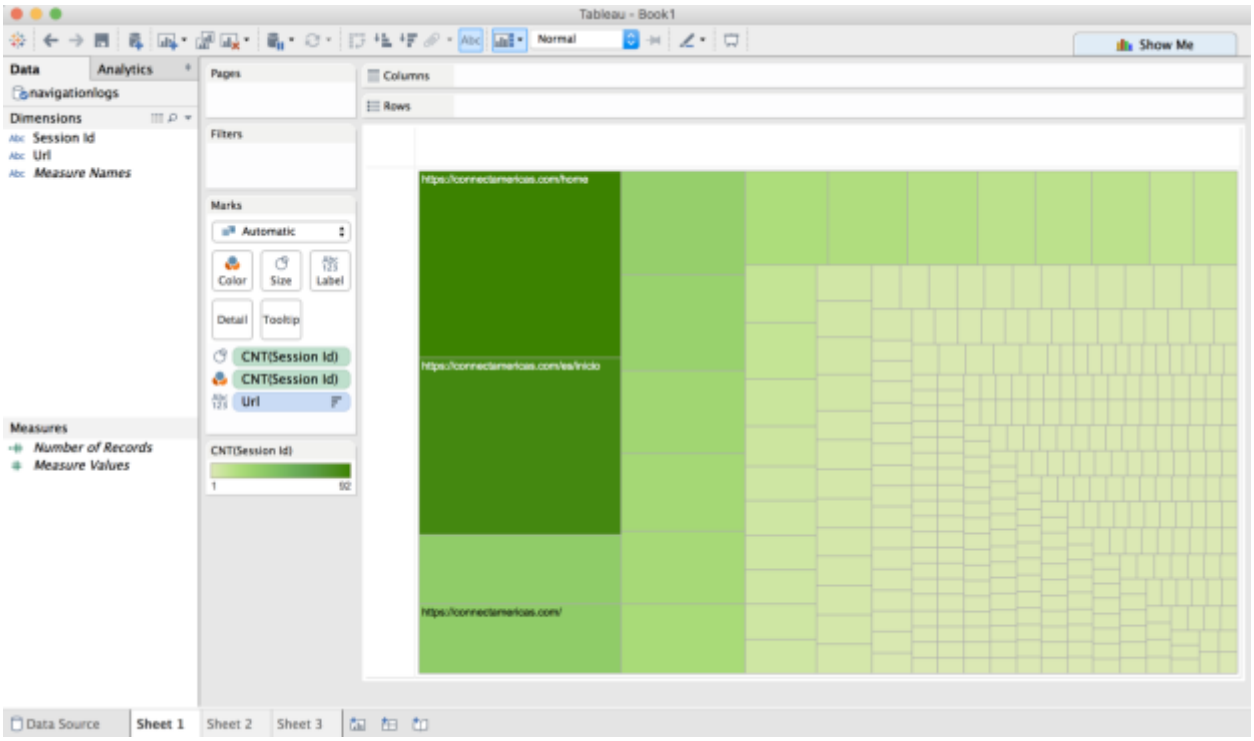
Taking note of the expert review it was decided to change the bubble chart, to the visualization proposed earlier that after being review with the client and expert advisors was choose as the most suitable. As describe before the solution will mixed different views that combine will accomplish all the tasks proposed.

First the we stablished that the way to accomplish the tasks, was to reduce the amount of data and then derive from the data the connections in order to get a new dataset who was more complex and insightful. To do this the data was filtered and reduce using dimensionality reduction were we took the attributes that were the most important for us, reducing the amount of data, then with this new data we derive points in common that different tables had in order to create a network that will be less in amount but bigger in meaning.

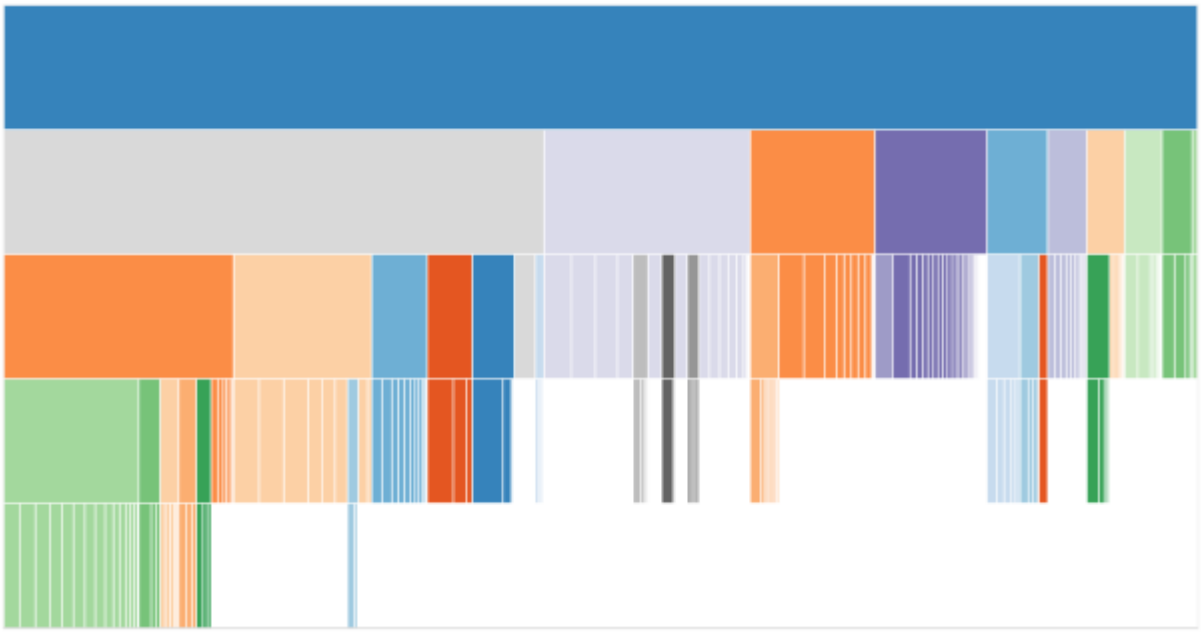
This lead us to the representation of the new data using a treemap as main view, this because it was the favor choice of the client due to the easy view. On our part we decided to use the treemap because was the idiom that allows to use more variables in order to filter the information and navigate the user path.



In the treemap we will see the different contents that exist in ConnectAmericas and will be differentiate by color, and by size it will be noticed what is the content more visited and relevant of platform. With this we will solve the first task and then continue exploring the treemap. By choosing a square of the treemap we will access to a more detailed information about the content.



In the next part of the treemap we will see what is the content inside the content of ConnectAmericas, for example if the user goes into the part of Business Opportunities, he will see the different proposals and business opportunities that are available in the platform. After this the user will explore the specific opportunity by choosing it, and it will go to an icicle.



In this icicle, the specific opportunity will become the node father, where it will be displayed the different characteristics that have the users when they navigate in the platform and visit that contents. This will help to discover the kind of users that use that content and see where they are from and what industry sector and different attributes. Leading to a specific user where it will be able to see what kind of pages, he uses in the platform and the way he gets access to the content. With this ConnectAmericas will be able to see what are the connections between the users and determine the patterns of use accomplishing the tasks 2 and 3.

This visualization was exposed to the client and receive in a very positive way, because in their opinion satisfy the three tasks that were selected and will help them see what is the public that visit the different content of the platform. In addition they consider that a line chart will be a good addition in where the company will review the metrics of visit of the contents over time and also they will like to add a statistical box with the averages of views of the contents and other metrics.