Visualization for Tracking Users Navigations On Web Pages Trackca

Gustavo A. Peña Morales
ga.pena@uniandes.edu.co
Jorge Hernandez.
je.hernandez1@uniandes.edu.co

Abstract: As social networks keep growing at such fast rate, the traffic of data that they have becomes 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 be 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 through the use of visualization system.

1. 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.

Visualization tools and systems offer many qualities and advantages that can help companies understand and organize data more efficiently with no need of data bases knowledge. The main objective of a visualization system is to be an interactive tool that the user can use to navigate data and have a better understanding of it, and also create insights and new knowledge about the data presented.

ConnectAmericas is a company that understand the importance of these tools, because of the great amount of data that they have, they find very useful a visualization system that they can analyze and study to have a better understanding of what is the user navigation in the platform and how to improve the user satisfaction with its content and engage more users to the platform. Among the needs of ConnectAmericas are the lack of a system that allows them to have an unified dataset, the difficulty to explore and find key values about the actions made on the platform, and how to see what the users are doing on it.

Having in mind these problems that ConnectAmericas we will discuss different methods of analysis and possible solutions and propose a visualization system that can help them solve these problems and also create new knowledge about how the platform is being used. We will explain the process of analysis that we did in order to come to this solution with its advantages and disadvantages, and also having in consideration previous work related and recommendations to improve this system.

Throughout this paper we will explain different methods of data analysis that can be used to reduce the size of it, with the use of dimensionality reduction, filtering and deriving which were the models used to get the information needed from the data to be visualized and fulfill the tasks proposed that lie into 1) Present the distribution of users navigation in ConnectAmericas content, 2) Derive the connections that users have in common while exploring the content and 3) Explore the characteristics of navigation, to determine patterns of use.

2. METHODS OF ANALYSIS AND REDUCTION

In this section we will explore the methods of analysis and reduction techniques, and provide insights about their benefits and why they were applied in this system.

2.1 Dimensionality Reduction

A technique used to reduce high dimensional dataset into more simple datasets where the amount of attributes from tables is reduced based on the criteria that these don't contribute to get more information and knowledge about the dataset. This technique tries to improve the performance of the analysis algorithms, that whit less data realize a faster and better analysis.

There are different methods to applied to this technique, some more direct than other but all with the same purpose reduce the amount of data to analyze, in this case we assume that the data that is always growing and generated in an independent form where the users have k components of attributes but these attributes can be more or less in each

user. Moreover, because we don't have the specific statistic model that is used, and only have the data. That's why we will only explain the less statistic ways to use this technique.

Missing Value Ratios: Is a technique where the attributes of table data are drop according to a threshold that is usually 40 to 50% reduce. This technique intends to reduce the size of the data in order to made it less specific and detailed, but more insightful and faster to connect and read.

Low Variance: When a data has an attribute or variable that is constant, the information that this adds to the dataset is minimum, this because its value has zero variance, in this technique those variables are found and drop from the dataset.

Decision Trees: Method where the data is divided into two or more homogeneous sets. These sets must be categorical input and output variables; these sets are divided into attributes where the most important attributes are used in order to obtain the most informative sub attributes, reducing the amount of items according to which sets has the most attributes.

In addition to these techniques we used models of filtering and embedded in order to reduce the data, and create a new dataset that is smaller and more insightful that the original.

2.2 Derive and Reducing

Derive: The main purpose of these technique is to produce a new dataset based on existing data. This proposes in convert an existing type of dataset into a new that is more complex and more insightful. In this case we have a multiple tables of datasets that were converted into a tree of data with the most important information having a clear hierarchy about ConnectAmericas content in order to apply the techniques describe before.

Reducing: Reducing is a method used to navigate and explore the data, reducing the size of the data available in order to the user to have a better view of the visualization. This method includes techniques of filtering, adding and embedded. That can be used at the same time in order to fulfill the navigation tasks in the data.

Filtering: Through the use of widgets the data is filtered and reduce so that the can be more manageable to the user. This technique allows to reduce in an intuitive and straightforward way the data, so that it can be easier to navigate.

Embedded: Technique where two or more datasets are unified in order to reduce the amount of data and avoid repetitions, this technique can be used at the same time with filtering and addition models.

3. RELATED WORK

The use of visualization tools to present data is not new, as we can see in **Interactive Visualization (Bill Ferster)** is a method that has been used for the last century. With the rise of computer technology and improvement in computer graphics, the use of more advance visualization systems has become one of the most important tools for companies and industries.

Visualization systems allow in interactive ways to study big sets of data and to create knowledge that can help companies to take better actions and understand better current situations. These contributions go beyond TI departments, visualization tools can help find KPI (key performance indicators) and understand them better. In **Applications of Data Mining Techniques to Identify Relevant Key Performance Indicators (Jesus Peral, Alejandro Maté, Manuel Marco)** they explain how the use of data mining and representations can help to get better KPI's and automate this indicators according to current user data.

Other authors and researcher have study what is the better way to represent such indicators, these because of the importance that these can have in the business, in Key Performance Indicators: Developing, Implementing and Using Winning KPI's (David Parmenter) They explain the real importance of kpi's in the 21st century and how these can make a company grow or fail. In work such as **Visualizing Key Performance Indicators Using GKPI Procedure (Brian Varney)** we see different methods to represent these indicators, so that the companies can understand them in a better and easier way,

These works are important because in ConnectAmericas the way they can see and how they analyze the impact of their platform is by the use and representation of how many views and users each content of the platform has.

In addition to these works the most relevant thing is to see how people have represented navigation paths on the web, for example in **Visualizing Web Navigation Data with Polygon Graphs. (Jiyang Chen, Tong Zheng, William Thorne, Daniel Huntley, Osmar Zaïane and Randy Goebel)** They developed a model where they take the data from a web browser and represent through the use of polygons how the user navigates the platform and what have in common with other users.

Other people have interested in see how people from different places connect and have analyze the way social networks allow people to interact with others on many topics and subjects. The project **Evaluating Social Navigation Visualization in Online Geographic Maps. (Yuet Ling Wong, Jieqiong Zhao, Niklas Elmqvist)** have as objective to see what places are the most active and represent them in map.

Other investigations have take even further interest in how people navigate the web and has research in ways to indentify patterns of use like **Visualizing and Discovering**

Web Navigational Patterns.(Jiyang Chen, Lisheng Sun, Osmar Zaïane and Randy Goebel) they study this, because they understand how important is the modern age.

Finally in **Model Based Clustering and Visualization of Navigation Patterns on a Website. (Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, Steven White)** We see how researches and companies work together in order to maximize the impact of different contents. In this paper they use a model based clustering to analyze each content of the platform of msnbc.com, to see what are the more popular contents and then reprensted in a bar chart where they can see the views of the contents.

These papers and system visualizations have helped us to decide what is the best way to represent and solve the tasks that we had stablished.

4. SOLUTION

First 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 tree 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 some scented widgets which are the first filter of the visualization where the user we will choose the respective content that wants to visualize. According to what the user chooses the user will interact with the treemap. The treemap is a representation that allows to differentiate through the use of color a specific country and with size what country is the one that is more active in that 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, were the size and color will be the representation of the different variables inside the treemap. After this the user will explore the specific opportunity by choosing it, and it will go to an icicle.

The reason behind the use for the icicle is that because we wanted to express a sense of vertical navigation similar to a tree, which is the form of the new dataset. As said before the treemap is the main view of the system that will lead to an icicle under the treemap.

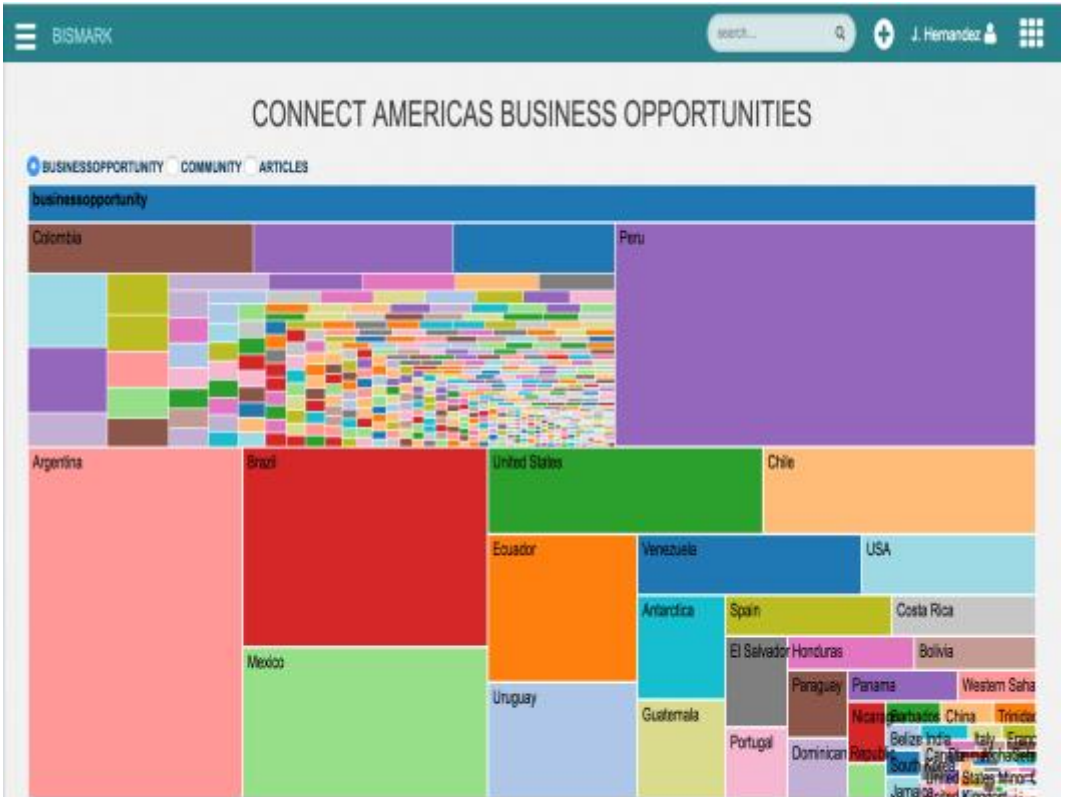

Figure 1 Main View TrackCA, treemap with the distribution of navigation.

In the icicle, the specific content will become the principle node or top of the tree, 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 part of the visualization will allow ConnectAmericas to find and specific user and identify what are his actions, and how many time he has visited that content, as well other contents. Also identified what are their status with the platform if they are just visiting or if they are members of the platform.

5. CONCLUSIONS

After being reviewed with ConnectAmericas we conclude that the system developed is an interesting tool that serves as the first part of a plan for the company to centralized the entire data of the social network. This system allows to navigate the distribution of what are the countries that visit more the respective contents in the platform.

In addition is a tool that allows ConnectAmericas to discover what things have in common the users that visit specific contents in the platform, also allows to pin out a specific user that can be studied in order to see the patterns of use that this user have while using the platform. These tasks that can be done with the visualization system TrackCA allow us to conclude that the tasks stablished at the beginning were accomplished.

Americas countries visit the most the contents of the page, they don't sing up and use it to interact like it is intended in a social network.

We can conclude too, that the company should have more campaigns and methodologies to engage more users in Latin-american countries that is the intended target of users of the company but also where they are having less impact.

These things were concluded with the use of the visualization system implemented, in order to see the distribution of web navigation that users had on ConnectAmericas.

6. FURTHER WORK

This visualization is a first step in the work of ConnectAmericas to explore and unified the databases, and allows them to have a first system where they will be able to explore the most relevant contents in the platform, but not all the contents.

The next step is to take all the data availbale to have a view of all the contents that are in the ConnectAmericas page. This will help them to create more global reports in a faster way and in a more concise way.

This system allows them to see a specific user this will help them to see how the user behaves and how he navigates in the platform, creating new information that can be visualize in a trend chart to see how many time he interact with each content of the web page.

In addition, the availability to integrate data that have a bigger time line will help them to have a more detailed view of the state that the company and its contents have, allowing them to have better KPI's established.

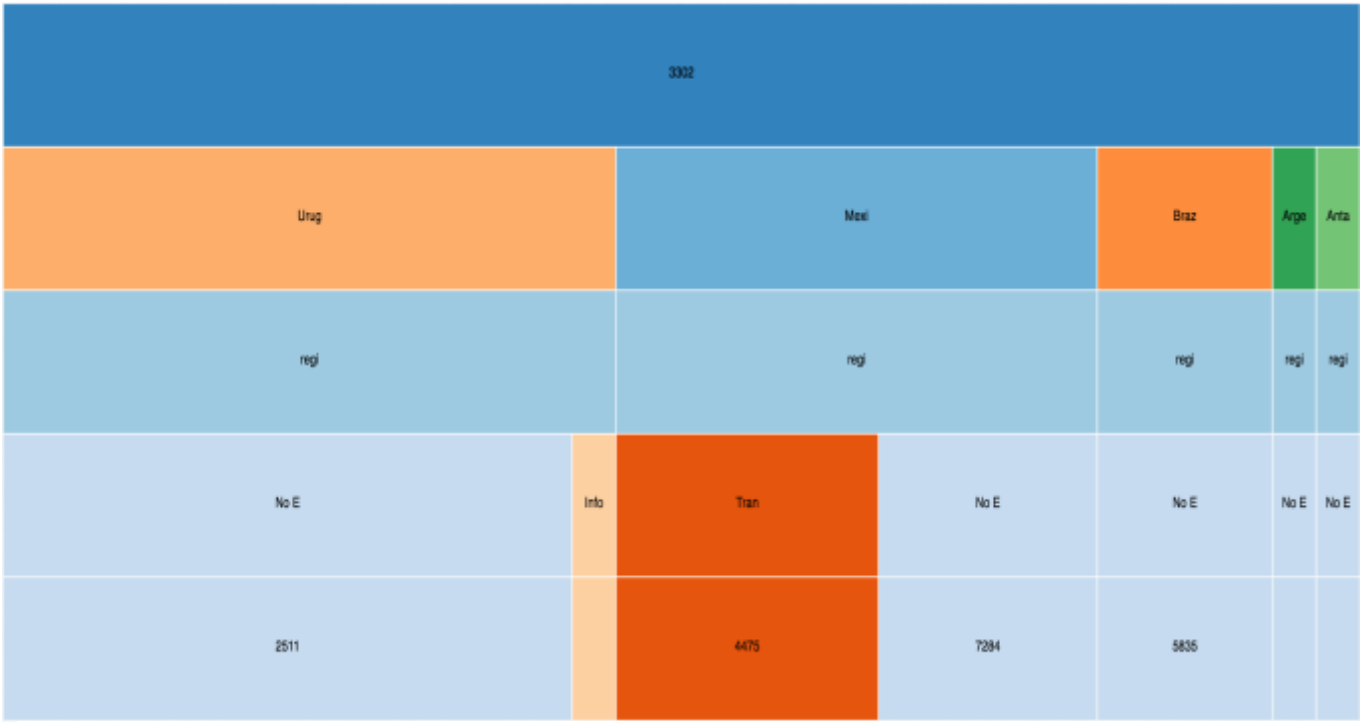

Figure 2 Icicle, representation of the characteristics that have the users while navigating a content.

M

countries are the most active in the platforms and use it as social network such as the USA or the European countries, while Latin

BIBLIOGRAPHY.

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

Interactive Visualization. Bill Ferster. (2011)

Applications of Data Mining Techniques to Identify Relevant Key Performance Indicators. Jesus Peral, Alejandro Maté, Manuel Marco. (2016)

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

Visualizing Web Navigation Data with Polygon Graphs. Jiyang Chen, Tong Zheng, William Thorne, Daniel Huntley, Osmar Zaïane and Randy Goebe (2007)

CZWeb: Fish Eye Views for Vizualizing the World Wide Web. Brian Fisher, Makrina Agelidis, John Dill, Paul Tan, Gerald Collaud, Chris Jones. (1997)

Evaluating Social Navigation Visualization in Online Geographic Maps. Yuet Ling Wong, Jieqiong Zhao, Niklas Elmqvist (2015)

Model Based Clustering and Visualization of Navigation Patterns on a Website. Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, Steven White. (2003)

Visualizing and Discovering Web Navigational Patterns. Jiyang Chen, Lisheng Sun, Osmar Zaïane and Randy Goebel (2004)

WebOFDAV – Navigating and Visualizing the Web Online with Animated Context Swapping. Mao Lin Huang, Peter Eades, Robert F. Cohen (1998)