Making new public policies from metadata of public statistics thanks to a thematic classification

Pierre Raimbaud, Camilo Espitia, John Guerra, Andres Mauricio Clavijo Abril
Universidad de los Andes, DANE Colombia

Abstract

Nowadays, all the citizens of one country expect that policymakers use the statistical data from making decisions. But they need for that purpose to have some tools for visualizing the statistics and that these tools present a thematic classification adapted to the typical public lines policies. In this paper we present our approach for deriving from metadata a thematic classification and also for building visualizations that use it.

Keywords: statistics, thematic classification, public policies.

Index Terms: visualization, design guidelines, interactive exploration

1 Introduction

The DANE (National Department of Statistics) is one of the most relevant organizations regarding data in Colombia. The public institution gathers information about all the major topics in the country periodically. Its information ranges from population statistics to technological literacy, access to services, among others. Nowadays, the institution is concerned because Public Policy in Colombia should be data driven. Currently, most of the time public institutions may not take in account all the data as per it is not available or not easy to access.

The DANE is aware of the need to improve the data availability and to provide tools for decision makers. For this reason, it is looking forward to develop several Visual Analytics applications to deliver appropriate tools to its stakeholders.

This public institution owns data both from *Statistical Operations* and from *Administrative Registers*. Although each dataset has different attributes and purposes, many datasets share characteristics and are related to several relevant topics for the country. The main objective of this project is to build a tool to understand which topics and keywords are present among different groups of *Statistical Operations* and from *Administrative Registers,* ultimately allowing decision makers to overview major topics and find which *Statistical Operations* and *Administrative Registers* are related to a specific topic or keyword.

2 Related work

2.1 Tamara Munzner's Framework

To carry on this project, we implemented Munzner's Visualization framework to abstract and understand the data, the user tasks and choose the best idioms to allow users to complete the tasks.

Munzner's framework has three dimensions: 1. the WHAT or the data, 2. the WHY or the user tasks and 3. the HOW or the visual representation.

WHAT: It refers to the information available to visualize. The basic abstractions of the dataset arrangements are tables, networks, fields and geometry. Within each dataset we can find items, attributes, links, etc. Data can be static or data can be dynamic and finally, the attributes of data can be ordered or categorical [4].

WHY: It refers to the tasks abstraction that must feature mainly one action and one target. The main objective is to clarify what is the main purpose of each visualization made. Actions can vary from high to low level, and range from exploring data to identify, compare or summarize targets. On the other hand, targets can be values, outliers, trends topologies, and others depending on the given dataset [4].

HOW: It refers to the design decisions taken to visualize the data and perform the required tasks. The objective of the final part of the analysis is to decide which visual channels like size, color, etc. will represent the data. In this section, we also choose the marks, or the visual representations for the data. In this stage we choose the visual encoding and the Idiom or representation that best suits the WHAT and WHY, to develop the visualization accordingly [4].

2.2 Examples

Once described Tamara Munzner's framework, we could explain visualizations similar to our work and that follow this framework. Each visualization uses an idiom: it is the representation (HOW) used for showing the data (WHAT) according to the task (WHY).

First we want to present the first idiom that we have used in this paper, which is the Bar Chart one, one of the most famous idiom in the literature. Typically it permits to summarize distribution, and show extremes if it uses order (ascending or descending). Elzer et al. [1] have already presented the efficiency of the bar chart idiom, going further in the reflection, by proposing a way to automatically understand bar charts. Note that another possible idiom for these tasks is the stacked bar chart but as Indramoto and al. [3] explained it, its focus is more on combining single-attribute and overall-attribute comparisons rather than making only single-attribute comparisons for one or more dataset (which is our case, see next section).

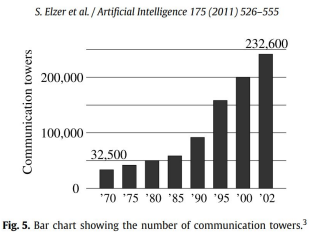

Fig. 5. Bar chart showing the number of communication towers.[3]

Figure 1: Bar chart visualization

Then, in this paper, we derived the original dataset (table) to network and tree datasets. In this case, following Tamara Munzner's framework, the dataset is composed by nodes and links - note that these links can be shown or not, depending on the task. About the data, in this project, the focus was on discovering a new thematic classification. As Ochs et al. [5] showed it, ontologies manipulations and representations are crucial nowadays but required much work: derivation, clustering and visualization as network or tree map. They presented a software framework for deriving, visualizing, and exploring abstraction networks. In this paper, we propose another approach using Tamara Munzner's framework and d3 technology, but also for thematic classification. In our case, we used the tree map representation and the radial one. Note that in both cases, one of the most critical point is the usage of forces in order to separate the nodes, depending on one attribute or relationship. Hilbert et al. explained the usefulness and importance of the forces in a network visualization ; they even went further, by proposing evolution of forces [2].

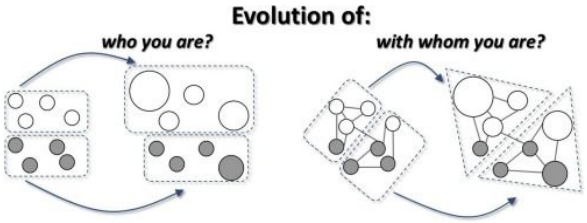

Figure 2: Network visualization with forces for clustering

3 Contribution

After implementing Munzner's framework we developed a visual analytics tool with three main components: 1. Context visualizations to analyse the state of the art of the information held by the DANE, 2. A treemap visualization to understand the results of the natural language processing undertaken to understand better the major topics around the DANE's datasets and 3. A visualization to navigate the identified topics and provide a tool for policy makers.

3.1 General and context

The first set of visualizations aim to represent the inventory of *Statistical Operations* and *Administrative Registers* held by the DANE. The main task is to **summarize** the **distributions** of both datasets according to different criteria, to answer the following questions: How many *Statistical Operations* and *Administrative Registers* does the DANE have? What is the proportion of *Statistical Operations* and *Administrative Registers* in the three major topics (Economical, Social and Environmental)? What is the proportion of *Statistical Operations* and *Administrative Registers* in each of the 30+ specific topics (e.g. Health, Education, Infrastructure, etc.)?

The datasets that we used were 2 inventories of *Administrative Registers* and 1 inventory of *Statistical Operations* provided by the DANE. The inventories dataset type were a table.

The best visual encoding to provide this overview were bar charts were *Statistical Operations* and *Administrative Registers* are differentiated by colors. With the chosen encoding is easy to provide the user with the summary of the DANE's information inventory.

Figure 3 shows the result of the first three general and context tasks and the visualizations developed.

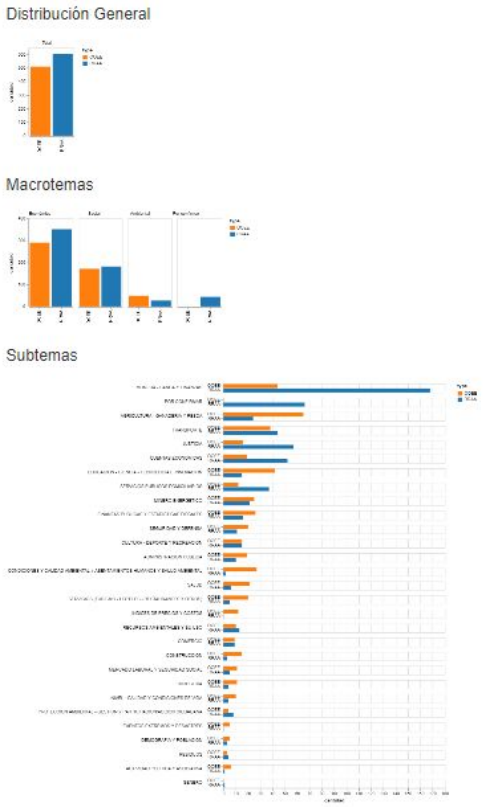

Figure 3: Bar chart visualization on already known themes

Figure 4 shows the result of the fourth and fifth visualizations developed, coming from the same dataset used in Fig 3 and with the same visual encoding, because it is still the best one for summarizing the distributions and also to identify extremes (because here we used the technique "separate order and align). Here the visualization presents the **distribution** of the attribute "Sub theme", allowing to know the global repartition of these subthemes for all these registers (in blue, left) and operations (in orange, right). It also allows to see that for registers, the subtheme most present is "moneda" whereas that for operations it is "agricultura".

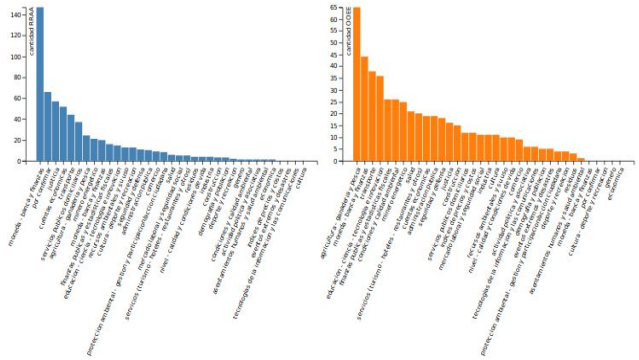

Figure 4: Bar chart visualization on already known themes

3.2 Treemap

Here we want to present the first representation that we have made for the following task: **summarize** the **distribution**, of the new classification: a tree map. This tree map uses the derived dataset that contain node, each one being a *Statistical Operation* or *Administrative Register*. We used clustering here for grouping them by new theme. For separating them we used an algorithm called force-in-a-box by J.Guerra (https://github.com/john-guerra/forceInABox). This representation allows to have a global vision of the new themes (transporte, investigación, derecho, agua etc.) It also to allow to detect that "servicio" and "credito" are the themes that most appear.

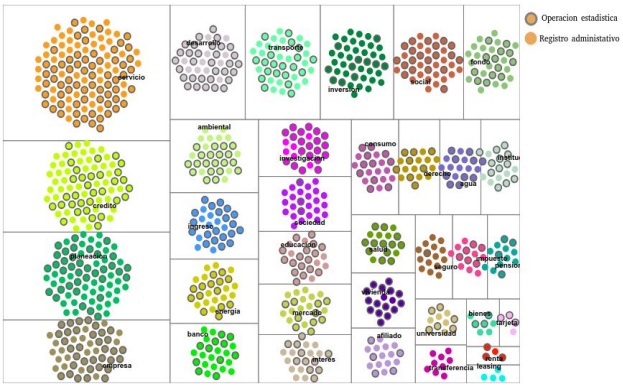

Figure 5: Tree map chart visualization on new themes

The following visualization is a table coupled to the previous treemap. It allows to give information about one item, after clicking on it in the treemap. As a result, we can get information on a specific node.

Tabla de detalle

Nombre: ESTADISTICAS SOBRE TERMINOS DE INTERCAMBIO
Tipo: OOEE
Operaciones estatisticas:
Objetivo: Calcular los terminos de intercambio del pais a partir de indices d e precios de importacion y exportacion del comercio exterior Col ombiano
Unidad de observación: Transacciones en el comercio exterior
Variables: Indices de precios de bienes importados - indices de precios de bienes exportados - terminos de intercambio
Resultados estad. (indicadores-agregados): Indice de precios de exportacion total - indice de precios de impo rtacion total - terminos de intercambio - terminos de intercambio segun comercio exterior - terminos de intercambio segun IPP
Area temática: ECONOMICA
Témática 2: MONEDA - BANCA Y FINANZAS
Nuevo tema principal descubierto: bienes
Otros nuevos temas(nombre, # occurencias): banco,1,bienes,2
Desagregación geografica: Nacional
Cobertura geografica: 0
Periodicidad (captura): Mensual
Entidad productora: BANCO DE LA REPUBLICA DE COLOMBIA
Dependencia DANE: DEPARTAMENTO TECNICO Y DE INFORMACION ECONOMIC A - SECTOR EXTERNO
Temas compartidos:
Metodología estadística OOEE: APROVECHAMIENTO DE REGISTRO ADMINISTRATIVO
Entidades que lo consumen:

Figure 6: Auxiliary view of .figure 5 visualization

3.3 Radial Force Visualization

To navigate the new generated topics, we created a visualization where the user can type a word and then explore the *Statistical Operations* and *Administrative Registers* that feature the keyword in their metadata. The main task of the visualization is to **identify features**.

In the visualization, after choosing the keyword, the user can see a radial force visualization where the *Statistical Operations* and *Administrative Registers* that contain the keyword are attracted to the center depending on the number of coincidences. Thus, leaving in the nucleus the most related items, and leaving the less related towards the outskirts of the radial visualization.

Figure 7 shows the Radial Force visualization implemented, where *Statistical Operations* are orange and *Administrative Registers* are blue. If the user puts the mouse over any item, he/she will see the name of the item and its type.

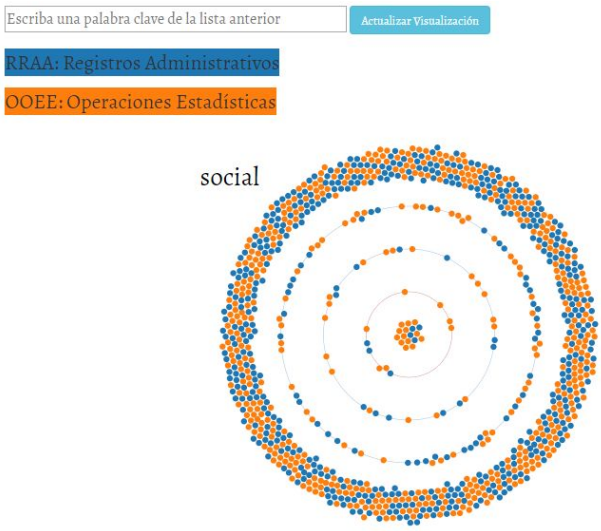

Figure 7: Radial force visualization.

To help **Identifying** the **Extremes**, we build an auxiliary view (Table - figure 8) to provide the user with a visualization to find the *Statistical Operation* or the *Administrative Register* that features the major occurrences of the keyword in all of its attributes.

Tipo	Nombre del Nodo	Puntaje	Keyword
RRAA	INFRAESTRUCTURA PECUARIA	24	social
OOEE	ACTIVIDAD JUDICIAL DE ENTIDADES PUBLICAS DE ORDEN NACIONAL	12	social
OOEE	ESTADISTICAS DE AFILIACIONES A LOS DIFERENTES SISTEMAS(O COMPONENTES) DE LA PROTECCION SOCIAL A PARTIR DEL REGISTRO UNICO DE AFILIACIONES - RUAF	8	social
RRAA	EVALUACION AL SISTEMA DE CONTROL INTERNO	7	social
OOEE	ESTADO DE MADUREZ DEL SISTEMA DE CONTROL INTERNO	7	social
OOEE	ESTADISTICAS SOBRE APORTES Y CONTRIBUCIONES A LAS OBLIGACIONES DE LA SEGURIDAD SOCIAL EN COLOMBIA	5	social
OOEE	ENCUESTA NACIONAL DE SALUD (ENS)	5	social
RRAA	OPERACIONES RECIPROCAS TIPO DE INFORME CERO (0)	4	social
OOEE	ESTADISTICAS SOBRE VIVIENDA DE INTERES PRIORITARIO Y SOCIAL INICIADAS CON APOYO DE FONVIVIENDA	4	social
RRAA	REGISTRO DE GENERADORES DE RESIDUOS O DESECHOS PELIGROSOS	4	social
OOEE	ENCUESTA LONGITUDINAL DE PROTECCION SOCIAL (ELPS)	4	social
OOEE	ENCUESTA NACIONAL DE SALUD - BIENESTAR Y ENVEJECIMIENTO (SABE)	4	social
OOEE	ESTADISTICA DE COBERTURA DE LA POBLACION AFILIADA AL SISTEMA DE SALUD	4	social
OOEE	ESTADISTICAS DE APORTES REALIZADOS A LOS DIFERENTES SISTEMAS DE LA PROTECCION SOCIAL A PARTIR DE LA PLANILLA INTEGRADA DE LIQUIDACION DE APORTES - PILA	4	social

Figure 8: Auxiliary view of .figure 7 visualization.

4 Experiment and results

4.1 Experiment

To validate our work, we organized an experiment where the experts from the DANE were invited to try our visualization according to the following story board that we provided to them (short summary here): first try to get new themes about registers and operations (with the treemap) ; then get some information about one item (with the coupled table); after that, write a word about one theme of your interest and discover which are the registers and operations with more relation with this keywords (with the radial visualization) ; finally read the most important register or operation in the coupled table.

4.2 Results

Note that the following results come from a questionnaire that the users filled after the experiment. The aim was here to evaluate the quality of our visualization. As a result we asked close questions using Likert scale, one for each visualization and task. And as the experiment was with experts from the DANE, we also asked open questions to get some feedback about the visualization and to make corrections. According to these results, we created the final visualizations that are presented here in the paper - before the users' experiment, we had four visualizations: the tree map (where the forces were not separating so good the clusters) and its coupled table, the radial force visualization and its coupled table.

Q1 - General impression ?

Q2- Explore a new classification of the items ?

Q3 - Obtain the detail for one item ?

Q4 - Discover the items (ordered) in relation with a theme?

Q5 - Identify the item with more relation with a theme ?

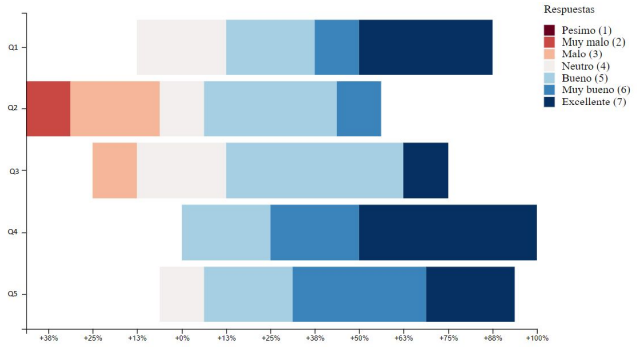

Figure 9: Results of the user tests.

5 Discussion

According to the results, both the quality results shown on the Likert scale figure and the feedback given by the experts, before our corrections, the task that costed a lot to the users was to explore the new classification (Q2) whereas that on of the easiest task for them was to discover the items in relation with a theme in particular (Q4). As a result, we did more corrections on the visualization used in Q1.

These visualizations allow to discover some insights about the statistical data and their classification in new themes. Here we provide some of them: with the initial data, we can discover that in the category "Economic", more information could be generated because there are more registers than operations (the registers produce and create the operations) ; thanks to our new derived data, we can discover more precisely that in the category "Economic", the focus should be done on "Leasing" and "Transactions" because there are no operations about these subjects. And about the public policy, thanks to the visualizations we can identify which registers and operations are more related to a theme: for example about "Housing", the more important register is "Financiation of housing VIS".

6 Conclusion

Finally, the DANE owns highly relevant information for the country and it is important that it continue with its efforts to develop more data analysis tools to provide its different stakeholders tools to maximize the usage of the data. Visual analytics is a tool that allows both policymakers and citizens to locate where the information is, and allows its understanding, ultimately enhancing policies and fostering data driven businesses.

References

[1] S.Elzer, S.Carberry, I.Zukerman, The automated understanding of simple bar charts, Artificial Intelligence, Volume 175, Issue 2, Pages 526-555, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2010.10.003. 2011

[2] M.Hilbert, P.Oh, P.Monge, Evolution of what? A network approach for the detection of evolutionary forces, Social Networks, Volume 47, Pages 38-46, ISSN 0378-8733, https://doi.org/10.1016/j.socnet.2016.04.003. 2016

[3] Indratmo, L.Howorko, J.Boedianto, B.Daniel, The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons, Visual Informatics, Volume 2, Issue 3, Pages 155-165, ISSN 2468-502X, https://doi.org/10.1016/j.visinf.2018.09.002.2018

[4] T.Munzner, Visualization Analysis and Design. A K Peters Visualization Series, CRC Press, https://books.google.de/books?id=NfkYCwAAQBAJ 2014.

[5] C.Ochs, J.Geller, Y.Perl, M.A. Musen, A unified software framework for deriving, visualizing, and exploring abstraction networks for ontologies, Journal of Biomedical Informatics, Volume 62, Pages 90-105, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2016.06.008. 2016