Visualización de publicaciones científicas empleando D3

José Federico Medrano[1], José Luis Alonso Berrocal[2], Carlos G. Figuerola[2]

jfmedrano@fi.unju.edu.ar, berrocal@usal.es, figue@usal.es

[1]VRAIn / Visualización y Recuperación Avanzada de Información / Facultad de Ingeniería

Universidad Nacional de Jujuy - Ítalo Palanca 10, +54 (388) 4221587

[2] REINA / Recuperación de Información Avanzada / Facultad de Traducción y Documentación

Universidad de Salamanca – España - C/ Francisco de Vitoria, 6-16

RESUMEN

Las bases de datos bibliográficas poseen una enorme cantidad de registros en forma de publicaciones científico-académicas, al realizar una búsqueda por autor, por ejemplo, ofrecen un conjunto de resultados, el cálculo de indicadores y métricas, y en algunos casos opciones de visualización de resultados. Este último aspecto, la visualización, ha sido un poco descuidado. Estas herramientas se limitan a presentar simples estadísticas o datos tabulados, o los conocidos grafos de co-autor o co-citas, dejando de lado algún otro tipo de visualización más compleja como la evolución temporal involucrando varias dimensiones más que la cantidad de citas recibidas por año. Una adecuada visualización de datos permitirá entender la realidad desde distintas ópticas aportando un mayor entendimiento y conocimiento a veces oculto en representaciones básicas y estáticas.

El desarrollo de este proyecto plantea encontrar la mejor forma, la más adecuada o proponer una nueva visualización para representar las relaciones existentes entre autores, publicaciones, citas recibidas, y cualquier otro tipo de información o relación que se considere relevante durante el estudio.

Palabras clave: *Data Visualization; Publicaciones científicas; DataVis; Visualización temporal*

CONTEXTO

La línea de investigación aquí presentada se encuentra enmarcada dentro del Proyecto Consolidado D/B029 denominado "*Aplicación de técnicas de Inteligencia Artificial para evaluar la producción científico-académica de investigadores de Universidades públicas del Noroeste Argentino*", aprobado y financiado por la Secretaría de Ciencia y Técnica y Estudios Regionales de la Universidad Nacional de Jujuy. Esta etapa del proyecto es llevado a cabo en conjunto por dos grupos de investigación. En primer lugar liderado por el grupo de investigación VRAIn de la Facultad de Ingeniería de la Universidad Nacional de Jujuy, y en segundo lugar como colaboradores, el grupo REINA de la Facultad de Traducción y Documentación de la Universidad de Salamanca

1. INTRODUCCIÓN

En las décadas anteriores era muy común que los datos se presentaran textualmente o mediante gráficos estáticos, en estos casos la información representada estaba limitada a cantidades

pequeñas, pero en los últimos años estos tipos de representaciones se han tornado poco útiles cuando se trata de conjuntos de datos que contienen millones de elementos de datos (Keim, 2002). Los sistemas actuales almacenan grandes cantidades de datos y al no tener la posibilidad de explorarlos adecuadamente, los datos se vuelven inútiles y las bases de datos se convierten en meros depósitos.

La aparición de interfaces gráficas ha permitido una interacción directa con la información visualizada, dando lugar a más de una década de investigación en Visualización de Información (InfoVis) (Heer, Card, & Landay, 2005). InfoVis busca aumentar el conocimiento humano mediante el aprovechamiento de las capacidades visuales humanas para dar sentido a la información abstracta (Card, Mackinlay, & Shneiderman, 1999), proporcionando los medios por los cuales los seres humanos mediante sus capacidades perceptivas, pueden lidiar con el constante aumento de la cantidad de datos disponibles.

El objetivo de InfoVis es profundizar en los datos o conceptos ocultos. A menudo la información se oculta simplemente por la enorme cantidad de datos disponibles. De este modo, la Visualización de Información también puede ser vista como un convertidor entre los datos subyacentes y la percepción humana de la misma (Prinz, 2006). El representar grandes cantidades de información mediante abstracciones no es una tarea fácil ya que el usuario no tiene ninguna idea preconcebida de cómo estos datos pueden ser representados.

La idea básica de la exploración visual de los datos es la de presentar los datos en alguna forma visual, permitiendo que los humanos puedan obtener conocimiento, sacar conclusiones, e interactuar directamente con los mismos. Con este tipo de representaciones basadas en grandes cantidades de datos, los usuarios pueden detectar patrones o comportamientos que se deseaban evaluar, como así también descubrir comportamientos y relaciones entre los datos desconocidos hasta el momento

La interacción es especialmente importante en InfoVis, ya sea para la exploración, análisis y/o presentación de los datos (Kosara, Hauser, & Gresh, 2003). La interacción permite al usuario implícitamente formar modelos mentales de las correlaciones y las relaciones entre los datos, a través del reconocimiento de patrones. El uso de los ordenadores permite ir un paso más allá de simples representaciones del mundo real, permitiendo hacer agrupaciones o asociaciones impensables, o distorsiones sobre dichas representaciones proporcionando un mayor nivel de abstracción, aspecto fundamental de las técnicas de InfoVis.

La Visualización de Información cumple un papel relevante al realizar cualquier estudio del estado de la producción científico-académica de un investigador, y más aún al considerar el recuento de citas de dichas publicaciones. Poder plasmar mediante una representación, la información recolectada se vuelve una tarea compleja no solo según aumente la cantidad de información, sino también según se incremente el número de dimensiones objeto de estudio. Las bases de datos bibliográficas hoy en día ofrecen un pequeño número de visualizaciones relacionadas con esta temática, se limitan a un conjunto de gráficos estáticos como la evolución de algunos indicadores calculados o el recuento o clasificación de las publicaciones de un autor en particular (Medrano, 2017).

El diseño de visualizaciones para la exploración de datos temporales requiere varias opciones basadas en aspectos de tiempo y representación visual. La elección del diseño o técnica debe adecuarse al problema en cuestión y a los valores que se desean informar, la misma magnitud puede representarse de múltiples formas y el

conocimiento que se pueda obtener de ella no tendrá el mismo impacto visual aplicando un diseño u otro. Como indica (Henkin, Dykes, & Slingsby, 2016) los aspectos a tener en cuenta deberían basarse en el diseño, la forma y el tamaño de las marcas visuales. En (Bach, Dragicevic, Archambault, Hurter, & Carpendale, 2016; 2014) se presenta una taxonomía completa de las distintas técnicas y modelos utilizados para representar datos temporales, tanto en 2 dimensiones (2D) como en 3 dimensiones (3D).

El *Scatter plot* también llamado *scatter graph*, *scatter chart*, *scattergram*, *scatter diagram* o diagrama de dispersión, es un tipo de diagrama matemático que utiliza coordenadas cartesianas para graficar puntos que muestran la relación entre dos variables de un conjunto de datos. Los puntos pueden ser coloreados para indicar los valores de una variable adicional. Al ser puntos dentro de un eje de coordenadas, el valor de cada punto está dado por la posición que ocupa, es decir, el valor de una variable según la posición en el eje horizontal y el valor de la otra variable según la posición en el eje vertical. A menudo se utiliza este tipo de diagramas para identificar asociaciones potenciales entre dos variables, en las que se puede considerar una variable explicativa (como años de educación) y otra puede considerarse una variable de respuesta (como el ingreso anual) (Lacey, 2017).

Los conjuntos de datos que involucran más de dos magnitudes resultan difíciles de visualizar en un espacio de 2D, en algunos casos visualizaciones en 3D pueden aportar ciertas mejoras o nuevas posibilidades, sin embargo al aumentar la cantidad de magnitudes o dimensiones del problema resulta necesario recurrir a visualizaciones un tanto más complejas que una representación lineal en ejes de coordenadas. Para este trabajo se diseñó una visualización en el espacio 2D recurriendo a ciertos elementos para poder incluir todas las magnitudes que debían ser representadas para capturar, en una sola representación, el entendimiento global del estado actual de los resultados de la investigación de un científico.

Para esto se desarrolló un primer prototipo de visualización utilizando la librería D3.js (Data Driven Documents) (Teller, 2013; Zhu, 2013).

D3 (creada por Mike Bostock[1]) es una elegante pieza de software que facilita la generación y manipulación de documentos web con datos. Lo realiza mediante (Murray, 2013):

- Carga de datos en la memoria del navegador.
- Vincular datos a elementos dentro del documento, creando nuevos elementos según sea necesario.
- Transformar esos elementos mediante la interpretación de los datos enlazados de cada elemento y establecer sus propiedades visuales en consecuencia.
- Transición de elementos entre estados en respuesta a la entrada del usuario.

D3.js fue creado para llenar una necesidad apremiante de una sofisticada visualización de datos accesible desde la web. La visualización de datos ya no se refiere a los gráficos de torta y gráficos de líneas. Ahora significa mapas y diagramas interactivos y otras herramientas y contenidos integrados en noticias, cuadros de mando de datos, informes y todo lo que se ve en la web.

La visualización desarrollada se basó en un ejemplo sencillo y estático de *Scatter plot* disponible en el blog de su creador (Bostock, 2019), ver Figura 1. Por su parte, en la Figura 2

[1] https://bost.ocks.org/mike/

se observa el primer prototipo de la visualización desarrollada.

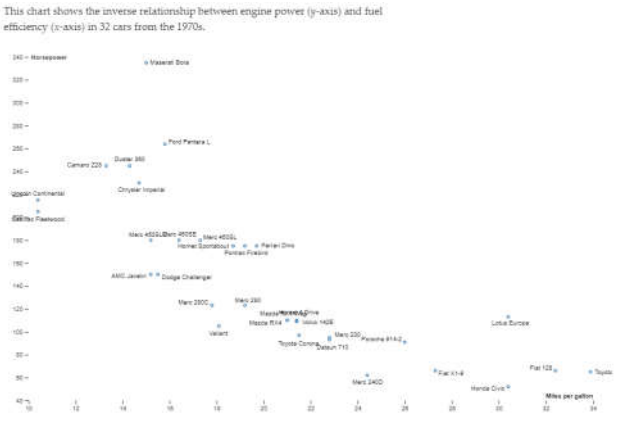

Figura 1: Visualización de Scatterplot provista por Mike Bostock

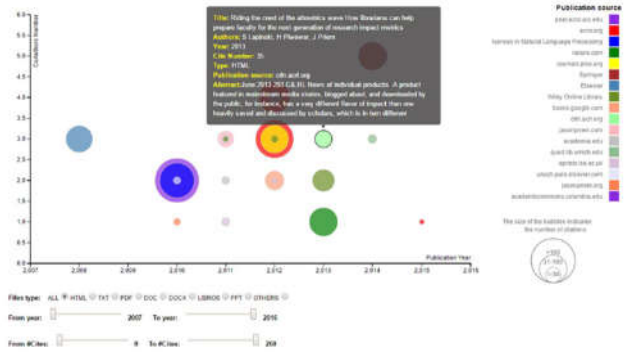

Figura 2: Visualización de Scatterplot adaptado

En el eje vertical se encuentra la cantidad de coautores de las publicaciones, en el eje horizontal el año de publicación, cada publicación se representa por una burbuja donde el tamaño de las mismas indica la cantidad de citas recibidas, y por último el color de las burbujas indica el nombre de la revista donde fue publicado el registro bibliográfico. Este prototipo ofrece un conjunto de opciones de interacción entre las que se incluye la posibilidad de filtrar por el origen de la publicación (mostrando u ocultando las publicaciones de dicho origen), filtrar por el tipo de archivo de la publicación eligiendo de la lista disponible, filtrar los registros por año de publicación, filtrar las burbujas de acuerdo a la cantidad de citas recibidas. Además permite visualizar el detalle de la publicación al pasar el mouse por encima de una burbuja y acceder al recurso publicado haciendo doble-clic sobre la burbuja.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de trabajo principal de este proyecto de investigación es el estudio, diseño e implementación de técnicas de Visualización de Información aplicables a representar el conjunto de publicaciones científico-académicas de un investigador a partir de la recolección de registros bibliográficos almacenados en bases de datos académicas tanto comerciales (Scopus, Web of Science) como de libre acceso (Google Scholar, Microsoft Academic), con el objeto de mejorar los procesos de comunicación de resultados de los análisis bibliométricos que se realizan sobre estas bases de datos. A partir de la generación de visualizaciones de dichos resultados, se logrará una mejor comprensión de los datos.

3. RESULTADOS OBTENIDOS/ESPERADOS

Con este proyecto se espera encontrar un modo novedoso que satisfaga la enorme necesidad de contar con visualizaciones de información de publicaciones científicas que permitan brindar una óptica diferente a resultados tabulares o gráficos sencillos y estáticos.

Particularmente se espera lograr:

- Mejorar los procesos de comunicación de resultados de los análisis bibliométricos que se realizan sobre bases de datos bibliográficas.
- Mejorar la comprensión de los datos a partir de la generación de visualizaciones de dichos resultados.
- Diseñar una herramienta web de libre acceso que permita generar visualizaciones a partir de un dataset específico.
- Diseñar una herramienta interactiva, agradable e intuitiva que ofrezca un

conjunto de opciones para aumentar el entendimiento sobre la información presentada.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto brinda un marco para que docentes y estudiantes lleven a cabo tareas de investigación y se desarrollen en el ámbito académico.

El área de visualización de información es incluida y desarrollada en cada proyecto que los equipos de investigación llevan a cabo, además de presentar un campo atractivo y novedoso a los alumnos que forman y formarán parte de los distintos proyectos.

5. BIBLIOGRAFÍA

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., & Carpendale, S. (2014). A review of temporal data visualizations based on spacetime cube operations. *Eurographics Conference on Visualization.*

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., & Carpendale, S. (2016). A descriptive framework for temporal data visualizations based on generalized space-time cubes. *Computer Graphics Forum.*

Bostock, M. (2019). *Observable*. Obtenido de https://observablehq.com/@d3/scatterplot

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think.* San Francisco: Morgan-Kaufmann.

Heer, J., Card, S. K., & Landay, J. (2005). Prefuse: A toolkit for interactive information visualization. *ACM Human Factors in Computing Systems (CHI)*, 421-430.

Henkin, R., Dykes, J., & Slingsby, A. (2016). Characterizing representation of temporal data visualization. *Poster presented at the VIS 2016*.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 7*(1).

Kosara, R., Hauser, H., & Gresh, D. (2003). An interaction view on information visualization. *EUROGRAPHICS 2003 State-of-the-Art Re-ports*, (págs. 123-137).

Lacey, M. (2017). *Statistical topics*. Obtenido de http://www.stat.yale.edu/Courses/1997-98/101/scatter.htm

Medrano, J. F. (2017). *Evaluación de la producción científica mediante motores de búsqueda académicos y de acceso libre.* Tesis doctoral, Universidad de Salamanca, Informática y Automática, Salamanca.

Murray, S. (2013). *Interactive Data Visualization for the Web.* O'Reilly.

Prinz, W. (2006). The Graph Visualization System (GVS): A Flexible Java Framework for Graph Drawing. *Master's thesis, Graz University of Technology*.

Teller, S. (2013). *Data Visualization with D3.js.* Packt Publishing.

Ware, C. (2004). *Information Visualization - Perception for Design.* Morgan-Kaufmann.

Ware, C. (2008). *Visual Thinking for Design.* Morgan Kaufman/Elsevier.

Zhu, N. Q. (2013). *Data Visualization with D3.js Cookbook.* Packt Publishing.