Trabajo de fin de Máster
Julio, 2017

Analítica visual de datos para representación de la interacción en una red social privada y con restricciones de privacidad

Jorge Durán Escudero

Departamento de Informática y Automática
Universidad de Salamanca

Declaro que he redactado el Trabajo de Fin de Máster (TFM) titulado: Analítica visual de datos para representación de la interacción en una red social privada y con restricciones de privacidad, del Máster Universitario en Sistemas Inteligentes de la Universidad de Salamanca en el segundo semestre del curso académico 2016-2017 de forma autónoma, con la ayuda de las fuentes y la literatura citadas en la bibliografía, y que he identificado como tales todas las partes tomadas de las fuentes y de la literatura indicada, textualmente o conforme a su sentido.

Además, soy conocedor de que el citado TFM forma parte de los trabajos de investigación que llevan a cabo mis directores Francisco José García Peñalvo y Roberto Therón Sánchez dentro del grupo de investigación GRIAL de la Universidad de Salamanca y, en consecuencia, comparto con ellos la propiedad intelectual de los resultados alcanzados.

En Salamanca, a 10 de julio de 2017

Fdo.: Jorge Durán Escudero

Dr. D. Francisco José García Peñalvo y Dr. D. Roberto Therón Sánchez profesores titulares del Departamento de Informática y Automática de la Universidad de Salamanca

CERTIFICAN:

Que el trabajo de Fin de Máster que se recoge en la presente memoria, titulado "Analítica visual de datos para representación de la interacción en una red social privada y con restricciones de privacidad", ha sido realizado por D. Jorge Durán Escudero, con DNI 70914526Y.

Y para que así conste a todos los efectos oportunos.

En Salamanca, a 10 de julio de 2017

Dr. D. Francisco José García Peñalvo
Dpto. Informática y Automática
Universidad de Salamanca

Dr. D. Roberto Therón Sánchez
Dpto. Informática y Automática
Universidad de Salamanca

Agradecimientos

Este trabajo no podría haberlo llevado a cabo de no ser por las siguientes personas, a las que quiero transmitir mis agradecimientos:

- A mis tutores, por sus indicaciones que han guiado este trabajo.
- A Alejandro Benito, mi compañero de trabajo y mi mentor en el uso de D3, por todas sus indicaciones para manejar esta compleja librería.
- A los distintos expertos que han dedicado parte de su tiempo a proporcionar información para este trabajo.
- A todos los miembros del proyecto WYRED que con su trabajo, lo han hecho posible y a la Unión Europea por financiarlo.
- A todos los profesores que se han implicado buscando transmitir el mayor número de conocimientos de la mejor manera posible. Gracias, sin vuestros conocimientos, consejos y enseñanzas esto no habría sido posible.
- A mis amigos y mis compañeros de ACM, por todas esas charlas y discusiones que me han proporcionado nuevas ideas y grandes momentos de relax.
- A mi familia, por su apoyo constante cada día.
- A todos aquellos desarrolladores que han publicado sus conocimientos y desarrollos, para que otros podamos reutilizarlos, en especial a Mike Bostock (creador de D3) y a Joel Spolsky y Jeff Atwood (creadores de StackOverflow).

"*El futuro mostrará los resultados y juzgará a cada uno de acuerdo a sus logros*"

Nikola Tesla

Resumen

En este trabajo se realiza una propuesta para estudiar los datos que se van a generar en la red social privada y anónima del proyecto WYRED, con el fin de extraer conocimiento sobre cómo interaccionan sus usuarios, tanto entre ellos, como con la propia plataforma. Para ello se parte de la creación de un sistema que generará un conjunto de datos de prueba, lo más parecido posible al original, y de una revisión sistemática de la literatura que ha permitido conocer las principales visualizaciones y el contexto en el que se aplican. Con esta información y teniendo en cuenta el impacto de la privacidad a la hora de tratar los datos del proyecto, se ha propuesto una arquitectura flexible y completa para el desarrollo de las visualizaciones interactivas que van a permitir visualizar los datos anteriormente generados. Finalmente, se presentan varios casos de uso donde se demuestra la idoneidad de la analítica visual para realizar análisis de los datos del proyecto y extraer conocimiento, de manera sencilla.

Abstract

In this document a proposal is made to study the data that will be generated in the private and anonymous social network of the WYRED project, in order to extract knowledge about how their users interact, both between them, and with the platform. To do this, it is started with the creation of a system that will generate a set of test data, as close as possible to the original, and a systematic literature review that has allowed to know the main visualizations and the context in which they are applied. With this information and considering the impact of privacy when dealing with the data of the project, a flexible and complete architecture has been proposed for the development of interactive visualizations that will allow to visualize the previously generated data. Finally, several use cases are presented where the suitability of the visual analytic is demonstrated to perform analysis of the data of the project and to extract knowledge, in a simple way.

Índice

Índice de figuras

Indice de tablas

1. Introducción

Hoy en día, las redes sociales son uno de los tipos de comunidades que mayor crecimiento están teniendo, gracias a la amplia difusión de las tecnologías de la información y la comunicación [3]. Sin embargo, las mismas siguen presentando algunos problemas, como la gestión de la privacidad o el análisis de los datos, para incrementar el conocimiento que se tiene de lo que está sucediendo dentro de ellas. Además, los expertos se encuentran con que, debido al volumen de información que generan, actualmente no es posible realizar análisis de manera manual de lo que ocurre en las mismas. Esto lleva a centrar este trabajo en la problemática de la gestión automática de estos datos y el planteamiento de un sistema que permita comunicarlos de manera efectiva.

1.1. Interés en el tema

En este ámbito ya se han realizado multitud de trabajos, tanto de investigadores independientes, como de algunos ligados a las principales empresas del sector (Facebook, Twitter, Google, etc.). Pero estos suelen estar relacionados muy estrechamente con el tipo de comunidad con la que ellos están trabajado. Es por ello, que muchas de sus propuestas no son aplicables de manera directa a otro tipo de comunidades.

En el caso de esta propuesta, el proyecto busca sacar partido de los datos que van a ser generados por la red social del proyecto WYRED [4]. El cual tiene algunas peculiaridades que serán detalladas en la Sección 1.2. Para ello se plantea una revisión del contexto del trabajo, de los datos del proyecto y de las preguntas de investigación más importantes, y cómo se puede ayudar a un investigador a resolverlas, realizando una propuesta de arquitectura para la construcción de una herramienta de visualización interactiva de datos.

Este enfoque de apoyar el análisis de los datos en herramientas que permitan visualizarlos de manera interactiva, ya ha sido usado con buenos resultados por multitud de empresas en todo el mundo, utilizando sistemas como Tableu[1]. Pero, como se ha comentado con anterioridad, una herramienta genérica muchas veces no se adapta de forma precisa a la tipología de cada uno de los proyectos individuales. Por esta razón, queda justificado el interés de realizar un sistema ad hoc, para ayudar a los investigadores a entender mejor los datos del proyecto.

1.2. El proyecto WYRED

El proyecto WYRED [5] es el marco de trabajo bajo el cual se desarrolla este proyecto; a continuación se presenta un breve resumen de los principales aspectos

[1] Tableau es una empresa que desarrolla *software* capaz de transformar los datos en visualizaciones interactivas, para aprovechar y conocer mejor los datos de negocio. https://www.tableau.com/es-es

que influyen en este trabajo. Para conocer este proyecto en mayor profundidad se puede consultar el Apéndice A.

Este proyecto nace con el objetivo de dar voz a los jóvenes, en un contexto plurinacional europeo, para que puedan plantear cuáles son los problemas que más les preocupan, sus opiniones sobre diversos asuntos, algunas posibles soluciones, ideas innovadoras para afrontar algunos desafíos, etc. Para la gestión de la comunicación entre ellos, se ha desarrollado una plataforma que actúa de manera similar a un foro de discusión, donde los usuarios organizan los debates por medio de comunidades, temas y comentarios en los mismos. Hasta aquí podría tratarse de un foro similar a muchos que hay en la red, sin embargo, el proyecto también tiene una serie de características propias que le distinguen del resto [6]. A saber:

- Los usuarios provienen de distintos países, lo que dota al proyecto de un contexto internacional. Esto implica multitud de aspectos, como el uso de varias lenguas, distintas características socioculturales y muy variados puntos de vista. Estos aspectos serán muy relevantes para los investigadores a la hora de extraer conclusiones sobre los diversos temas que tomen como objeto de su investigación.
- La necesidad de salvaguardar la privacidad de los usuarios, debido, en primer lugar, a que estos pueden ser menores y en segundo lugar, a que se busca hacer de la plataforma un lugar donde los mismos puedan interactuar libremente, para lo cual se requiere un alto grado de anonimidad.

Además de las restricciones ya mencionadas, el trabajo a desarrollar consiste en una propuesta de una arquitectura que sirva para visualizar y analizar los datos generados por el proyecto WYRED. Así, algunas cuestiones como la arquitectura o las visualizaciones propuestas, vendrán determinadas por las propias características del proyecto.

1.3. Objetivos

El objetivo principal de este trabajo es plantear, en estas primeras etapas del proyecto WYRED, una propuesta de arquitectura de un sistema que permita dar soporte a la construcción de visualizaciones interactivas que ayuden a comprender mejor los datos, para anticiparse a las necesidades futuras del proyecto.

Esta arquitectura tiene que ser lo suficientemente flexible para poder adaptarse a las diversas características del proyecto, permitiendo además, construir sobre ella cualquier tipo de visualización que sea requerida, en esta etapa o en un futuro. Para ello debe apoyar a los investigadores en dos tareas principales:

- Conocer cómo evoluciona la comunidad y el contenido que se está generando.
- Ayudar en el proceso de la toma de decisiones.

Es importante destacar que, aunque el principal objeto de estudio del proyecto sean los jóvenes, la arquitectura va a tener como usuarios finales, a los propios investigadores del proyecto.

En línea con el objetivo final del proyecto, lo que se busca en última instancia, es ofrecer un soporte a la toma de decisiones de los representantes públicos, para que tomen medidas que ayuden a mejorar la vida de los jóvenes y, en definitiva, que aprovechen sus aportaciones.

Por otro lado, uno de los objetivos secundarios que se busca alcanzar con este trabajo, es validar esta propuesta, para decidir si la misma debe formar parte de la plataforma o no, en un futuro próximo. Otro de los mismos es estudiar los distintos mecanismos para representar la gran cantidad de información que generan este tipo de plataformas.

1.4. Organización del documento

Bajo este apartado queda recogido cómo se ha estructurado este trabajo. En el apartado de metodología (Sección 2) se abordan los distintos pasos dados para la realización de este trabajo. En el estado del arte (Sección 3) se analizan las distintas propuestas que han realizado otros autores para resolver los problemas de este estudio. La generación del conjunto de datos se trata en la Sección 4, donde se aborda cómo crear un conjunto de datos que simule los que ofrecerá la plataforma una vez esté en funcionamiento. En la propuesta de arquitectura (Sección 5) tiene lugar la explicación detallada de la arquitectura para construir las visualizaciones interactivas, así como ofrece una descripción de los módulos de los que consta. En la Sección 6 se muestra cómo se ha implementado la arquitectura y algunos casos de uso donde se resuelven preguntas de investigación. En las conclusiones y líneas de trabajo futuras (Sección 7) se recogen los objetivos logrados y algunos caminos para ampliar este trabajo. Además de lo anterior, el trabajo cuenta con tres apéndices, siendo el primero el Apéndice A, donde se expone el proyecto WYRED, seguido del Apéndice B en el que se recoge la revisión sistemática de la literatura realizada y, finalmente, en el Apéndice C se aborda una introducción a las principales tecnologías web utilizadas.

2. Metodología utilizada

En este apartado se recogen los pasos y técnicas utilizados para la realización de este trabajo, haciendo hincapié en describir con mayor profundidad aquellos que han supuesto una mayor aportación para la realización del mismo.

2.1. Búsqueda de necesidades

En primer lugar, se ha preparado un pequeño cuestionario con el que captar aquellas tareas más relevantes para los usuarios potenciales del sistema. Esto se convierte en un aspecto básico para luego poder proponer las visualizaciones que mejor se adapten a ellas. Además, en el mismo también se indaga sobre cuáles son las preguntas de investigación que les parecen más importantes a los investigadores que podrían usar el sistema en el futuro.

El formulario está estructurado en dos secciones principales: la primera contiene un pequeño resumen de la información del proyecto, para poner al investigador en contexto y la segunda está formada por las siguientes preguntas formuladas:

1. ¿De las siguientes preguntas y tareas, cuáles crees que son más relevantes? Siendo el valor 5, relevancia máxima y 1, relevancia mínima.

   a) Conocer la evolución de un tema a lo largo del tiempo.

   b) Conocer cómo se relacionan los usuarios dentro de una comunidad.

   c) Conocer cuáles son los países más activos.

   d) Conocer cuáles son los lenguajes más usados.

   e) Conocer cuáles son los temas más frecuentes.

   f) Poder filtrar los datos por fechas.

   g) Poder realizar comparativas por género.

2. Si has identificado otras preguntas o tareas que pueden ser relevantes en este contexto. Por favor inclúyelas a continuación, junto con el valor de relevancia correspondiente.

Como se puede apreciar, la primera pregunta recoge algunas de las tareas y preguntas de investigación que se han considerado que podrían ser relevantes para los investigadores. Con respecto al tipo de pregunta, se ha decidido que en lugar de aceptar una respuesta binaria a las mismas, el usuario pueda emitir una valoración indicando cómo de relevante considera cada una de ellas, para extraer así más información. La segunda pregunta está realizada con el fin de permitir a cada uno de los expertos incluir sus propias aportaciones, siendo opcional la contestación a esta cuestión.

Para seleccionar a los expertos que han realizado la encuesta, se ha optado por sólo incluir a aquellos que tengan experiencia en la gestión o el análisis de comunidades en línea. Los mismos pertenecen principalmente a la Universidad de Salamanca y a la Universidad de Tel Aviv, siendo todos ellos investigadores en el proyecto WYRED. La lista completa de los mismos y su filiación se puede consultar en la Tabla 2.

Investigador	**Universidad**
Francisco José García Peñalvo	Universidad de Salamanca
Roberto Therón Sánchez	Universidad de Salamanca
Juan Cruz Benito	Universidad de Salamanca
Aharon Hauptman	Universidad de Tel Aviv

Tabla 2: Investigadores encuestados

Analizando las respuestas aportadas por los mismos, se desprende que todas las preguntas de investigación se consideran relevantes, menos la de conocer los lenguajes más usados, por su bajo respaldo, siendo las más valoradas el conocer cómo se relacionan los usuarios dentro de una comunidad, la exploración de los temas más frecuentes y la capacidad de realizar comparativas, como se puede comprobar en la Tabla 3.

Pregunta	**Valoración media**
Conocer cómo se relacionan los usuarios dentro de una comunidad	4.50
Conocer cuáles son los temas más frecuentes	4.25
Poder realizar comparativas por género	4.25
Poder filtrar los datos por fechas	4.00
Conocer la evolución de un tema a lo largo del tiempo	3.75
Conocer cuáles son los países más activos	3.25
Conocer cuáles son los lenguajes más usados	2.75

Tabla 3: Resultados del formulario de búsqueda de necesidades con expertos

Además de las respuestas anteriores, es de vital importancia analizar aquellas sugerencias que han ofrecido los expertos para este trabajo, siendo las más destacables las siguientes:

- La capacidad de comparar el interés en un tema por género, país y rango de edad.
- La posibilidad de conocer los términos más usados dentro de un tema.
- El soporte para la detección de valores anómalos (usuarios que no se relacionan, temas que en un instante se dispara su uso, etc.).
- La posibilidad de conocer si comunidades distintas se comportan de diferente manera ante determinados tópicos.

2.2. Revisión sistemática de la literatura

Una revisión sistemática de la literatura, conocida por sus siglas en inglés como SLR (*Systematic Litarature Review*), es un proceso sistemático que permite recopilar, analizar y conocer los documentos de mayor calidad en un campo específico [7] [8].

El nacimiento de esta técnica se puede enclavar en el campo de la salud, donde es común que los autores realicen una revisión exhaustiva de un campo, con el fin de convertirse en expertos en ese ámbito [9] [10]. Además, este tipo de revisiones son muy apreciadas por el resto de investigadores del sector, ya que estos trabajos se convierten en manuales de referencia.

La importancia del uso de esta técnica radica en la posibilidad de poder replicar el estudio, para verificarlo. Esto es posible debido a que el autor o autores de la revisión lo realizan de manera sistemática, es decir, anotando cada paso que dan en el mismo, así como informando de las decisiones que toman en cada apartado. Para ello es básico que el proceso cumpla las siguientes 3 características [11]:

- **Sistematicidad:** lo que significa que no es arbitraria, ni subjetiva, sino que se ha realizado buscando evitar cualquier tipo de sesgo, utilizando las mejores herramientas y fuentes de información.
- **Completitud:** lo que implica que se han examinado todos los documentos al alcance del autor sobre la temática concreta.
- **Explicitud:** lo que denota que la revisión documenta todos los pasos dados y los métodos aplicados en cada uno de ellos.

Una revisión sistemática de la literatura parte de una pregunta o preguntas de investigación, para las cuales se quiere encontrar una respuesta. Además, en este momento se tiene que decidir qué tipo de documentos van a ser válidos para la revisión. El siguiente paso consiste en establecer las palabras clave y las búsquedas a ejecutar en las bases de datos bibliográficas, para obtener un primer conjunto de documentos a analizar. Después, se realizarán varias etapas de filtrado; normalmente, una primera basada en el título y el resumen del documento, teniendo también presente los criterios de inclusión y exclusión, y otra basada en el propio contenido de los documentos. Una vez que ya se ha realizado el proceso de filtrado, hay que realizar una etapa de aseguramiento de la calidad. Para ello se definen una serie de preguntas con las que valorar todos los documentos recuperados, así como las posibles respuestas y su puntuación. Después, se realiza la valoración de los documentos y, finalmente, se escribe un documento donde se resume la información que proporcionan, respecto a las preguntas de investigación planteadas en un principio [12].

Además de la revisión sistemática de la literatura, este tipo de trabajos suelen ir acompañados de un *mapping* de la literatura. Este último es un trabajo donde los autores resumen, en líneas generales, el estado del arte. Algunos de los temas de interés para un *mapping* de la literatura son los siguientes:

- La evolución temporal de un campo.
- Los principales autores del sector, tanto por número de artículos, como por relevancia.
- Los medios que recogen con mayor asiduidad los trabajos en un campo de investigación.
- Las principales técnicas y su grado de utilización.

En definitiva, el *mapping* sistemático de la literatura sirve para conocer el contexto de un campo de investigación, siendo además, un buen punto de partida para realizar una revisión sistemática.

3. Estado del arte

En este apartado se van a analizar las principales aportaciones que han realizado los autores en los diversos campos que se abordan en este trabajo.

3.1. Redes sociales

Una red social es un lugar que pone en contacto a un conjunto de organizaciones y/o personas con algún interés común. La redes sociales han sido objeto de estudio por parte de los investigadores desde hace varias décadas, pero la importancia de este tema se ha multiplicado con el auge de las redes sociales en línea. Este tipo de redes, al almacenar toda la información generada en línea, han permitido estudiarlas con mayor profundidad, ya que los investigadores pueden acceder a todo, o a la mayor parte de su contenido [13].

La evolución de las redes sociales en línea ha venido de la mano de la propia evolución tecnológica [14]. Esta ha supuesto multitud de cambios respecto a las redes sociales o comunidades que se habían constituido anteriormente. Algunos de estos son: la creación de nuevos roles en la comunicación, el surgimiento de jerarquías menos estrictas, donde hay espacio para discusiones más plurales o la capacidad de contactar con un mayor número de individuos.

Los principales métodos para analizar redes sociales no son muy recientes, sin embargo, suponen el punto de partida para el estudio de una red social en línea. El mecanismo más estudiado y aplicado en este ámbito, son los grafos [15]. Mediante los mismos, se pueden representar a los usuarios y sus interacciones por medio de un método matemático ampliamente estudiado. Esto permite la utilización de un vocabulario común y la aplicación de los conceptos ya conocidos de la teoría de grafos: teoremas, derivaciones, deducibilidad de un concepto, etc. Además, la estructura de los grafos es muy flexible, permitiendo representar tanto las interacciones o comunicaciones unidireccionales, como las bidireccionales. Este mecanismo también posee la ventaja de que es fácilmente transformable en forma matricial, lo cual facilita mucho el tratamiento de los datos y la operación con los mismos.

La representación más común usa el enfoque de nodos y arcos, siendo los primeros la representación de los usuarios y los segundos, la de las relaciones entre ellos. Una vez realizado este paso, se pueden calcular multitud de métricas.

Una de las métricas más importantes en el análisis de las redes sociales es la centralidad, la cual permite conocer los individuos que tienen un mayor grado de conexión con el resto. El cálculo de la misma es sencillo:

Sea G:=(V,E), dónde G es un grafo, V un conjunto de vértices y E un conjunto de aristas. Para todo $v \in V$, su centralidad de grado $C_{DEG}(v)$ se define como:

$C_{DEG}(v) = grado(v)$

Siendo el $grado(v)$ el número de enlaces que posee v con el resto de vértices

$(V - \{v\})$.

En relación a la métrica anterior, otros autores como Freeman [16] proponen calcular la cercanía ($C_{CLO}(i)$) para conocer mejor cuáles son los nodos centrales del grafo. Esta métrica se basa en calcular la suma de las distancias más cortas de un vértice a todos los demás:

$C_{CLO}(i) = \sum_{j=1}^{n} S_{i,j}$

Siendo S la matriz que contiene el valor de la distancia mínima para cualquiera de dos vértices dados.

Siguiendo con la idea de transformar una red social *online* en un grafo, otros autores [17] centran su estudio en describir cómo evolucionan este tipo de comunidades. Para ello analizan distintas métricas que calculan la densidad de una red, es decir, el número de interconexiones por persona. Esto les permite afirmar que hay un patrón subyacente en la evolución de este tipo de redes, el cual está formado por tres fases: la primera de ellas presenta un crecimiento muy rápido de la comunidad, seguido de una contracción para, finalmente, mantener un crecimiento lento.

Además, los investigadores anteriores descubrieron que los usuarios de las redes sociales se pueden clasificar en 3 grupos:

- *Singletons*: pertenecen a este grupo los usuarios caracterizados por nodos de grado 0 y representan a las personas que se unieron a la red, pero que nunca se han conectado con otros usuarios. Estos quedan descritos como seres solitarios que no participan activamente en la comunidad.

- *Giant component*: forman parte de este grupo los usuarios que están conectados a otros por varios caminos. Este conjunto contiene a los más activos de la red, cuya característica principal es que suelen estar en contacto con un amplio número de usuarios de la red.

- *Middle region*: este conjunto agrupa al resto de usuarios no categorizados. Normalmente está compuesto por varias comunidades aisladas o pequeños grupos que interaccionan entre ellos, sin contacto con el resto de la red.

Otros autores al analizar redes sociales ya creadas [18], han abordado ambos aspectos en sus estudios: la relación entre usuarios y el análisis de los grupos que forman estos. Con estos estudios, han llegado a la conclusión de que las redes sociales están formadas por usuarios muy activos y con muchos enlaces, que actúan de supernodos. Estos usuarios tienen la capacidad de lograr difundir sus publicaciones a lo largo de toda la red social. Esto es posible debido a que, cuanto mayor es el grado de socialización de un usuario, más aumenta la probabilidad de que otros que participan en la red confíen en él.

Esta investigación detecta uno de los mayores problemas que surgen al analizar las redes sociales, el acceso a la información. Para solucionarlo, proponen usar una

serie de *crawlers*[2], teniendo en cuenta las limitaciones de cada plataforma.

Algunos estudios más recientes se centran en el análisis de una red social en concreto. Por ejemplo, en el estudio de la red social Twitter [19], se puede ver cómo se comportan los usuarios. De este trabajo, se extraen dos conclusiones importantes:

- A mayor número de contactos, mayor es el número de publicaciones. Lo que indica que los usuarios con muchos contactos, son muy activos en la red.
- Al principio, debido al pequeño alcance de las publicaciones, los usuarios son muy activos para ganar seguidores. Pero una vez se consigue cierta repercusión, el número de publicaciones se estabiliza.

Otros trabajos importantes realizados en este ámbito se encargan de analizar la estructura de una comunidad en línea y de proponer algoritmos para la detección de subcomunidades [20], siendo este tema uno de los más estudiados.

En un primer momento, para la detección de subcomunidades se utilizó el *clustering* jerárquico [21]. Esta técnica permite ir agrupando los nodos más próximos, de tal manera que cuando la diferencia entre un nodo y el grupo al que se iba a unir sobrepasa un umbral, se detecta ese grupo como una subcomunidad.

Más adelante las técnicas utilizadas se basaron en detectar las subcomunidades, centrándose en aquellos nodos y arcos más centrales del grafo, como se ha comentado anteriormente [21]. Sin embargo, la técnica propuesta por los autores ([20]) tiene un enfoque totalmente distinto para resolver el problema. Esto es debido a que parte del grafo completo y lo que va haciendo es ir eliminando los enlaces menos significativos de manera progresiva.

3.2. Foros de discusión

Los foros de discusión en línea, a los que me referiré simplemente como foros en este trabajo, han tenido una gran influencia en el surgimiento de las comunidades en línea. Estos sistemas nacieron como una evolución de los grupos de noticias albergados dentro de *Usenet* en los ochenta, volviéndose populares entre los noventa y los primeros años del nuevo milenio. Hoy en día, con el auge de las redes sociales como Facebook o Twitter, los foros han ido perdiendo usuarios y su popularidad también se ha visto reducida, aunque todavía, algunos siguen teniendo una gran importancia como Forocoches, la 46º página web más visitada en España y una de las mayores comunidades de habla hispana [22].

En cuanto a las características, los foros cuentan con algunas que los distinguen de otra comunidades y de las actuales redes sociales:

- Las discusiones en los foros están organizadas dentro de comunidades, subcomunidades y temas.

[2]Un *crawler* es un programa que descubre, inspecciona y, en algunos casos descarga, el contenido de una web. Para ello utiliza la estructura de enlaces de la propia página.

- No todos los usuarios tienen los mismos roles, siendo frecuente encontrar una jerarquía de tipos de usuarios.
- La mayoría cuentan con una moderación, más o menos estricta, que filtra el contenido.
- La mayor parte de ellos son temáticos, al contrario que las redes sociales donde, normalmente, cualquier tema tiene cabida.
- El acceso a algunas partes y/o herramientas suele estar limitado a los usuarios de mayor jerarquía.
- La comunicación no es de usuario a usuario, sino que, principalmente, es de un usuario hacia un tema.

Debido a las características intrínsecas de este tipo de comunidades, algunos autores han centrado sus estudios en ellos [23]. Estos llegan a la conclusión de que los principales trabajos realizados para estudiar los foros, se basan en análisis cuantitativos (número de usuarios, frecuencia de publicación, crecimiento, número de interacciones, etc.). Sin embargo, este no es el único enfoque posible para estudiar un foro, ya que existen distintos protocolos para analizar el contenido generado.

Una de las formas de sacar partido al contenido escrito en un foro es la aplicación de las técnicas de minería de textos. Estas técnicas sumada a las del análisis del sentimiento, proporcionan multitud de información de gran valor [24]. En ese trabajo, los autores proponen utilizar los algoritmos de las K-medias y las máquinas de vectores de soporte (SVM). Ambos son usados para agrupar y clasificar las comunidades en grupos más compactos, para después calcular el centroide que será la comunidad a analizar. Además de aplicar el análisis del sentimiento sobre el tema tratado, para conocer la opinión (positiva o negativa) de los usuarios sobre el mismo, los investigadores se centran en predecir la evolución del tema y de la opinión de los usuarios sobre el mismo. Esto último, se considera muy relevante, debido a que este tipo de comunidades evolucionan muy rápido al generar de manera dinámica multitud de información.

Además de analizar cuantitativamente y cualitativamente la actividad de un foro, algunos autores prefieren indagar sobre cómo afectan los distintos roles en el desarrollo de este tipo de comunidades [25]. La figura que analizan estos autores es la de los instructores, es decir, aquellos encargados de guiar las discusiones y moderarlas, siendo este un rol clave, por ejemplo, en los foros de discusión escolares.

Una de las conclusiones más importantes del estudio anterior es que cuanto mayor es la participación de los instructores, más se reduce la de los alumnos. Por lo tanto, la estrategia de iniciar muchos debates, por parte de los instructores, para incrementar la participación, no es fructífera. Sin embargo, los investigadores llegaron a la conclusión de que, para aumentar la valoración de los instructores, estos deberían encargarse de iniciar algunas preguntas genéricas y de responder aquellas que han quedado sin respuesta, después de un tiempo.

3.3. Privacidad

La gestión de la privacidad siempre es un tema complejo cuando se tienen que manejar grandes volúmenes de datos, de los cuales se puede extraer gran cantidad de información personal. Además, a esto se une que los usuarios cada vez están más concienciados con mantener su privacidad *online* [26]. Fruto de ello, es la negativa que presentan muchos de ellos ha participar en comunidades o lugares que van a ser controlados con el objetivo de recolectar datos para realizar investigaciones. Lo cual se convierte en uno de los mayores impedimentos para llevar a cabo, por ejemplo, estudios sociológicos o de comportamiento.

Para solventar este problema, en la medida de lo posible, es necesario definir cómo va a ser la comunidad. Si esta es abierta, los usuarios fácilmente van a entender que los datos pueden ser tratados, sin embargo, si es cerrada, por defecto los usuarios pensarán que sus datos van a permanecer privados. Esto va a hacer necesario que el usuario sea informado y acepte el consentimiento por el cual permite el acceso a todos o parte de sus datos. En [26], los autores plantean que el consentimiento pueda ser dado de manera pasiva, siempre y cuando, todos los datos que se proporcionen estén anonimizados.

Anonimizar los datos de los usuarios tampoco es una tarea sencilla, ya que según diversos autores [27], se puede identificar a un usuario debido al uso en diversos sitios de sus mismas fotos de perfil. Este proceso también puede ser aplicado si se tienen los suficientes datos sociológicos y la muestra es pequeña.

El estudio anterior también analiza un comportamiento curioso por parte de los usuarios, mientras evitan participar en comunidades donde sus datos van a ser tratados, no ponen el mismo empeño en proteger su información personal en las redes sociales como Facebook:

- El 89 % usa su nombre real completo y el 3 %, parte del mismo.
- La mayoría de los usuarios ha subido alguna imagen (90.8 %), pudiendo ser identificados completamente por su imagen de perfil en el 61 % de los casos.
- Más del 80 % de los usuarios muestran de manera pública su fecha de nacimiento y el lugar donde estudian.
- Más del 60 % de los mismos, comparten su música, libros y películas favoritas, su estado civil y su lugar de nacimiento.
- Solo el 1.2 % de los usuarios tienen activada la opción para ocultarlos de los resultados de las búsquedas.

La colección de los datos anteriores, como bien afirman en el estudio, permite construir de manera sencilla una base de datos con mucha información privada de cada uno de los usuarios. Lo cual puede servir para espiar a un amplio conjunto de usuarios o para realizar ataques basados en ingeniería social.

Análisis con mayor profundidad sobre afecta la privacidad a los jóvenes, se encuentran en algunos estudios [28][29], donde se detallan aspectos como el uso cada vez mayor de este tipo de redes por estos usuarios, siendo utilizadas principalmente para mantener el contacto con sus amigos. Sin embargo, en una proporción nada desdeñable (21 %), estos han sido contactados por algún extraño por medio de las mismas, lo cual les ha resultado incómodo.

En referencia a su comportamiento y los datos compartidos públicamente, los autores de los estudios anteriores, concluyeron que tanto la edad como el género, son factores muy influyentes:

- Las chicas suelen compartir un mayor número de imágenes, tanto suyas como de sus amigos.
- Los chicos utilizan con mayor asiduidad información falsa en sus perfiles.
- Los adolescentes comparten más datos verídicos que los niños.

Debido a todo lo anterior, se puede concluir que las personas no cuidan lo suficientemente su privacidad y por ello, si se quiere analizar los datos de una comunidad, es necesario encargarse de proteger la privacidad de los mismos, más si cabe, en el caso de los jóvenes.

3.4. Analítica visual

Hoy en día, la cantidad de datos almacenada cada vez crece con mayor velocidad, gracias al abaratamiento de los costes de almacenamiento. Esto ha provocado que se guarden multitud de datos en bruto, sin ser procesados ni tratados, por lo que no se aprovechan para extraer nueva información [30].

El manejar este amplio conjunto de datos es muy costoso, tanto en tiempo como en dinero. Por esta razón surge la necesidad de plantear nuevos mecanismos que permitan tratar estos datos, ya que en ellos se encuentra información muy valiosa. Estas metodologías necesitan resolver algunas cuestiones como el cálculo de la relevancia de un dato o la identificación de patrones de comportamiento.

Aunque no existe una única manera para abordar esta problemática, una de las más utilizadas es la Analítica Visual. Esta ciencia provee un conjunto de tecnologías para sacar partido de las fortalezas de los humanos y del procesamiento computacional de la información, permitiendo que ambos colaboren para procesar y analizar los datos de una manera más transparente [31]. Este punto de encuentro donde ambos colaboran son las visualizaciones.

Históricamente, la visualización de los datos ha tenido un rol menor, dentro de las investigaciones. Por lo que, en muchos casos, esta se ha visto relegada a un mero sistema para representar los resultados de un estudio. Para esta tarea, los autores han utilizado de manera provechosa los gráficos de barras, de sectores, de líneas, etc.

Sin embargo, la analítica visual alcanza su verdadera potencia, al pasar de ser una herramienta para confirmar hipótesis sobre los datos, a un sistema para explorarlos.

Además, este ámbito se ha visto muy beneficiado por el desarrollo de la computación, ya que la informática le ha dado las herramientas para poder construir nuevas visualizaciones y adaptar las existentes. Para ello, en primer lugar, se popularizó Java [32] al soportar diversos sistemas operativos y ser fácilmente exportable a la web, y hoy en día, con la madurez de las tecnologías web, el enfoque más popular es la combinación HTML+CSS+JavaScript, con la biblioteca para visualización de datos D3.js [33] como su mayor exponente. Fruto de esta evolución, se ha conseguido dotar de mayor interactividad a las distintas visualizaciones, lo que permite mejorar el desempeño de la analítica visual.

Una vez descritos los medios necesarios para aplicar esta ciencia, se considera importante analizar las principales tareas que se abordan mediante la analítica visual [31]:

- Sintetizar la información para extraer conocimiento.
- Trabajar con datos masivos, dinámicos y ambiguos.
- Detectar las características esperadas y descubrir otras nuevas.
- Proporcionar conclusiones justificadas.
- Comunicar las conclusiones de manera efectiva e inteligible.

Para llevar a cabo estas tareas es necesario seguir un proceso claro y justificado, que parta de los datos en bruto iniciales hasta conseguir distintas propuestsa de visualización, que sirvan a los usuarios para analizar los datos y extraer conclusiones. Algunos de los pasos más importantes en el proceso son los siguientes:

- El preprocesado de los datos es fundamental para alcanzar visualizaciones óptimas. En esta fase se eliminan registros erróneos, se normalizan los valores, se agrupan y/o se integran los datos desde distintas fuentes [34].
- Una vez se tienen los datos ya listos para ser estudiados, hay dos maneras, principalmente, de proceder:
    - Aplicando una técnica de análisis automático, como la minería de datos, para descubrir modelos en los datos, y luego plantear visualizaciones basadas en esos modelos.
    - Planteando en un primer momento un conjunto de visualizaciones que permitan obtener los modelos intrínsecos de los datos, y luego confirmar estos datos mediante cálculos.

La analítica visual es una disciplina que agrupa multitud de campos de estudio, como bien recoge [35]. Estos son, principalmente, la minería de datos, la estadística,

la psicología cognitiva y el diseño. La combinación de estas áreas de investigación tan diversas, y a la vez tan distantes las unas de las otras, proporciona soluciones al problema de la representación de conjuntos masivos de datos.

La analítica visual, se ha aplicado con éxito en multitud de áreas como:

- Las ciencias de la tierra, por ejemplo en la astronomía, al tener que manejar la ingente cantidad de datos generados por los distintos aparatos de exploración espacial y las simulaciones. Este área del conocimiento trabaja con información que presenta ruido, lo cual es un desafío a la hora de ser analizada [36][37]. También se aplican para representar modelos climáticos o las dinámicas oceánicas debido a que permiten conocer mejor sus efectos y variaciones en distintos puntos de la tierra [38].
- La gestión de las incidencias, que hace uso de la analítica visual para detectar rápidamente valores y/o patrones de comportamiento inusuales [39][40].
- La sanidad, en campos como la medicina, para comunicar sus estudios clínicos [41], en la biología para explorar los datos resultantes de una secuenciación [42] o del uso de *microarrays* [43][44], o en el de la bioinformática, al tener que tratar con grandes volúmenes de datos como los genes presentes en un organismo o sus proteínas [45][46].
- La educación, para representar la evolución en la adquisición de conocimientos de los alumnos [47], y explorar sus patrones de comportamiento [48] o con el fin de medir los progresos en plataformas de *e-learning* [49][50][51].
- La inteligencia de negocio, donde es muy importante poder detectar cambios en el comportamiento de los usuarios o las ventajas competitivas de una empresa frente a otra [52][53].
- La informática, al tratar con sistemas muy complejos, como los ecosistemas tecnológicos [54][55] o la gestión de proyectos *software* con multitud de componentes y líneas de código, los cuales tienen que ser analizados [56][57] y configurados [58].
- El deporte, especialmente en los de grupo, como el rugby [59], el béisbol [60] o el baloncesto [61][62], donde se utiliza para representar el movimiento de los jugadores y sus acciones a lo largo de los partidos.
- Las humanidades digitales, en las que se trabaja, habitualmente, con corpora multimedia [63] o textuales [64][65], de vastas dimensiones para generar nuevo conocimiento.

Como se ha podido apreciar, la analítica visual se ha aplicado en multitud de áreas. Sin embargo, las distintas propuestas son muy dependientes del conjunto de datos con el que se trata. Por esta razón se ha decidido realizar un estudio pormenorizado de la aplicación de la analítica visual en el estudio de las comunidades.

Este estudio, el cual se puede consultar íntegramente en el Apéndice B, es una revisión sistemática de la literatura. Esta técnica, como ya se comentó en la Sección 2.2, recopila, filtra y analiza los principales trabajos de desarrollados en un campo. Las conclusiones más importantes de este estudio, se detallan a continuación.

Respecto a la evolución de este campo, se puede afirmar que es un concepto bastante moderno, ya que los artículos analizados son posteriores al año 2000 y que además, sigue en desarrollo, al tener varias propuestas muy recientes.

Al analizar el tipo de datos con el que trabajan los investigadores, se llega a la conclusión de que la mayoría de los artículos utiliza, principalmente, estadísticas de uso o el propio contenido generado (corpora documentales), sin embargo, algún autor propone un conjunto integral de visualizaciones donde trabaja con ambos al mismo tiempo.

Las propuestas más utilizadas para explorar estos datos son: los grafos, cuando se quieren analizar las estadísticas de uso con el fin de mostrar o agrupar los usuarios en comunidades [66][67], y los gráficos de área, para mostrar la evolución temporal de los temas más frecuentes en el contenido [68][69]. Además de estas dos representaciones, también son utilizadas de forma recurrente las coordenadas paralelas [70], a fin de representar las múltiples características de un individuo o tema [71].

Llegados a este punto, es importante destacar que la analítica visual requiere de visualizaciones interactivas para representar amplios conjuntos de datos [72]. Además, la interactividad es el componente que permite realizar representaciones bajo demanda, ajustándose al criterio del usuario y modificando su representación para permitir descubrir el conocimiento implícito en los datos, mediante técnicas como el filtrado múltiple o el *zoom* semántico. Sin embargo, debido a la complejidad de generar este tipo de visualizaciones, este aspecto sólo ha sido abordado de manera satisfactoria recientemente [73] [74]. Finalmente, hay que recalcar que el componente interactivo dota de mayor eficacia a las visualizaciones realizadas [75], como han demostrado los estudios realizados con usuarios, presentes en algunos trabajos analizados [71] [76].

3.5. Creación de un conjunto de datos

La creación de un conjunto de datos que sirva para probar las distintas propuestas o hipótesis del presente trabajo, no es una tarea sencilla, pero en muchos casos se vuelve imprescindible por la falta de acceso a los datos originales. Esto se puede dar por multitud de motivos, como la falta de permisos, las restricciones de privacidad o por no estar disponibles los mismos en el momento actual. Esto ha llevado a un gran número de autores a intentar generar u obtener un conjunto de datos que sea similar al conjunto objetivo, lo cual, en el caso de la simulación de una comunidad, implica simular el comportamiento de los usuarios. Entender la relación entre los distintos atributos de un usuario y la influencia en su comportamiento y en cómo interaccionan es algo específico de la sociología, y por tanto, queda fuera del alcance muchos investigadores.

Uno de los procesos más utilizados en este campo, consiste en extraer los datos de alguna comunidad próxima a la que es objeto de estudio. En este caso, se puede afirmar que el comportamiento en ambas, por parte de los usuarios, será similar y, por tanto, no es necesario realizar una generación artificial de un conjunto de datos. Las principales fuentes para la obtención de conjuntos de datos son las mayores redes sociales (Twitter, Facebook, Flickr, etc.), las cuales han sido analizadas en profundidad por multitud de autores que, en la mayoría de los casos, han puesto a disposición de otros investigadores sus datos [77] [78]. El principal problema de estos conjuntos es que suelen presentar los datos ya anonimizados y esto limita la investigación y las conclusiones que pueden extraerse con los mismos. Además, algunos de ellos son conjuntos demasiado genéricos, lo que no permite realizar estudios centrados en campos concretos.

Otros autores [79] han propuesto utilizar algunos datos que son de más fácil obtención, como las entradas en los ficheros de registro, para generar el conjunto de datos. De tal manera que aquellas características que estén presentes en los registros y en el conjunto de datos objetivo, se mantengan y las que no aparezcan, sean generadas a partir de la combinación de otras que sí formen parte de los mismos. Un ejemplo de esto último, podría ser asociar al género el valor masculino, cuando las visitas de los usuarios se produzcan en minutos con valor par y femenino en el caso contrario. Este enfoque posee la ventaja de que parte de los datos se corresponde con información real y, por tanto, es posible estudiarla para encontrar patrones y verificar hipótesis, mientras que el resto de los datos pueden servir para añadir contexto a los mismos, formando un conjunto de datos más completo que pueda servir para presentar una herramienta o un caso de uso de una metodología.

Otros investigadores centran sus estudios en generar el conjunto de datos por completo, de manera artificial. Dentro de este campo hay que destacar a los que se centran en simular la interacción y los que además de lo anterior, intentan generar el contenido que se produciría. En el primer caso, han trabajado en modelar de forma matemática el crecimiento y la evolución de las interacciones en una red [80], lo que les permite alcanzar un conjunto de datos cuyo comportamiento es representativo. En el segundo caso, los autores se enfrentan a la alta complejidad que implica la generación de contenido, por ejemplo, de tipo textual, junto con la asignación de atributos representativos a cada individuo y sus interacciones. Aquí el principal exponente es LDBC-SNB Data Generator [81] el cual es un programa desarrollado para generar conjuntos de datos de comunidades para LDBC (*Linked Data Benchmark Council*) [82], una institución que se encarga de realizar análisis del rendimiento de distintas las tecnologías aplicadas al procesamiento de información, representada, en la mayoría de los casos, en forma de grafo. Para asignar a los atributos valores de manera lógica, los autores se basan en S3G2 [83], un *framework* que define la correlación que existe entre determinados atributos. Para la elección de los valores, el *software* cuenta con un conjunto de diccionarios donde están presentes los distintos valores que pueden tomar los atributos, seleccionando el valor final mediante diversas funciones que modelan la probabilidad de un suceso.

4. Generación del conjunto de datos de prueba

Uno de los problemas que se ha tenido al realizar este trabajo, ha sido la no disposición de un conjunto de datos con el que realizar el análisis de la arquitectura propuesta. Es por ello que se ha tomado la decisión de intentar generar un conjunto de datos de prueba lo más similar a los conjuntos reales con los que debería operar esta propuesta.

En este caso, de las tres opciones para generarlos analizadas en la Sección 3.5, sólo la opción de obtenerlos de manera artificial es viable, ya que no hay suficientes datos de comunidades similares y tampoco se tiene acceso a ficheros de registro de sistemas utilizados por el tipo de público objetivo.

En un primer momento, se intentó utilizar LDBC-SNB Data Generator para recrear un conjunto de datos formado por los usuarios del sistema. Sin embargo, esto no fue factible, al generar un conjunto de datos que no es personalizable y que no contiene algunos de los atributos necesarios. En la Tabla 4, se puede consultar algunos usuarios generados por este *software*. Una de las mayores ventajas de este sistema, como se puede apreciar, es que es capaz de generar nombres y apellidos acordes para cada usuario, sin embargo, en este caso al tratarse de datos anónimos, es una característica que no aporta valor.

Id	First name	Last name	Gender	Birthday	IP
933	Mahinda	Perera	Male	1989-12-03	192.248.2.123
1129	Carmen	Lepland	Female	1984-02-18	81.25.252.111
8333	Chen	Wang	Female	1980-02-02	1.4.16.148
8698	Chen	Liu	Female	1982-05-29	14.103.81.196
8853	Albin	Monteno	Male	1986-04-09	178.209.14.40
10 027	Ning	Chen	Female	1982-12-08	1.2.9.86

Tabla 4: Usuarios generados por LDBC-SNB

Otra de las ventajas de este *software* es que es capaz de generar los mensajes (y el propio contenido de los mismos), que podrían publicar los usuarios de prueba. La generación automática de textos es un problema que se viene abordando desde hace tiempo con diversas implementaciones, basadas en la selección de palabras y su disposición adecuada mediante algoritmos que buscan el cumplimiento de las reglas gramaticales [84] [85] y, más recientemente, mediante el uso de redes neuronales profundas [86]. Sin embargo, por los resultados obtenidos, este *software* genera el contenido de los mensajes mediante la combinación de varios extractos de textos, como se puede apreciar en la Tabla 5. Este tipo de mensajes, cuyo contenido es incongruente, no es válido para la aplicación de técnicas de análisis automático del lenguaje.

Id	Content
1 236 950 581 248	About Augustine of Hippo, ustinian religious order; his memoriAbout Nicolas Sarkozy, y was also president of the General About Robert
2 061 584 302 084	About Augustine of Hippo, ears he was heavily iAbout Nicolas Sarkozy, nuary 1955) is a FrenAbout Mary, Queen of Scots, land
1 786 706 395 168	About Augustine of Hippo, 30), also known as AugustAbout Plato, o not mean the systematicAbout E
412 316 860 456	About Rory Gallagher, her recorded solo albums throughout the 197About Call the Man
962 072 674 360	About Jawaharlal Nehru, dhi and the mateAbout Gil Kane, ated Iron Fist wAbout Dancing on My

Tabla 5: Mensajes generados por LDBC-SNB

Llegados al punto en que las soluciones ya creadas no proporcionan un conjunto de datos que permita probar la solución propuesta, se decidió intentar generar el conjunto de datos, de manera automática y escalable.

El primer paso consistió en definir qué atributos van a tener los usuarios, para ello se uso como base el documento donde se describen las especificaciones principales del proyecto [87]:

- Id: un identificador único de cada usuario que sólo permita que sea identificado por el administrador del proyecto.
- La edad calculada a partir del año de nacimiento del usuario.
- El grupo de edad al que pertenece.
- El género.
- El país desde donde participa.
- La provincia desde donde interviene.
- El nivel educativo.
- El rol con el que participa en el proyecto.

Tomando como referencia el artículo original en el que se basa LDBC-SNB Data Generator [83], se separaron los atributos anteriores entre los que son independientes y los que dependen de los anteriores, como puede apreciarse en la Fig. 1.

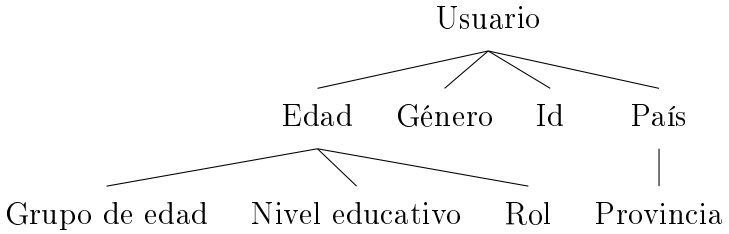


Figura 1: Dependencia entre los atributos de un usuario

Asignar valores a cada uno de los atributos no es una tarea sencilla, ya que es complejo recoger todos los patrones que se encuentran implícitos. El caso más sencillo es el del Id, el cual es simplemente un número consecutivo. Para el género, se decidió extrapolar los datos del estudio *Teens, Social Media & Technology Overview 2015* [88], donde se afirma que el 72 % de los chicos y 70 % de las chicas usan la principal red social (Facebook), asignando de manera equitativa ambos valores.

Respecto a la edad, se estableció un intervalo entre 14 años, la edad mínima para formar parte de una red social en España [89], y 29 años, al estar la plataforma destinada a un público joven. Sin embargo, la probabilidad de encontrar un usuario con una edad específica no es igual para todos los valores. Según el estudio sociológico anteriormente citado, entre los 15 y los 17 años la cuota de uso de Facebook es del 80 %, mientras que entre los 18 y los 29 años, esta sube hasta el 88 % [90]. Es por ello que se modeló la probabilidad de que un sujeto tenga una edad x ($P(x)$) como:

$$P(x) = \frac{x}{(x_1 + x_n) * \frac{n}{2}} = \frac{x}{43 * 8} = \frac{x}{344} \tag{1}$$

Una vez que se ha calculado la edad del individuo, la asignación del grupo de edad es sencilla: si es mayor de edad, *Adult*, y en caso contrario, *Teenager*.

Respecto al nivel educativo, este también se ve influido por la edad, al ser necesaria, normalmente, una cierta edad para avanzar a un nivel educativo superior. Pero, en este caso, fue complicado encontrar datos del nivel educativo por edades, ya que tanto los estudios de la OCDE [91], como los del Ministerio de Educación [92] se limitan a la población entre 25 y 64 años. Sin embargo los estudios del INE [93] si recogen esta información, la cual se puede consultar en la Tabla 6.

Nivel educativo	**Edad**		
ISCED 1997	**16-19**	**20-24**	**25-29**
0	0 %	1 %	1 %
1	7 %	4 %	6 %
2	68 %	24 %	26 %
3	22 %	37 %	12 %
4	3 %	12 %	11 %
5	0 %	22 %	44 %

Tabla 6: Nivel educativo según la edad. Fuente: INE 2017

Con estas probabilidades se simuló el nivel educativo de los usuarios potenciales del proyecto.

El siguiente atributo a modelar es el rol que tiene un usuario en la plataforma, indicando si el mismo es un usuario normal o es un moderador. En el proyecto se calcula que habrá en torno a 1 moderador por cada 50 usuarios, y estos serán siempre mayores de edad. Como la probabilidad de ser menor de edad es del 18 %, se puede despreciar esto, y suponer que 1 de cada 50 usuarios mayores de edad serán moderadores.

Respecto al país del usuario, se decidió asignar los valores en función de la población que tienen los principales países involucrados en el proyecto, como se puede ver en la Tabla 7.

País	**Población (millones de hab.)**	**Población (% respecto al total)**
Austria	8.611	4 %
España	46.620	22 %
Irlanda	4.641	2 %
Israel	8.380	4 %
Italia	60.800	29 %
Turquía	78.670	39 %

Tabla 7: Distribución de población de los países involucrados en el proyecto

Una vez conocido el país, se procedió a obtener una provincia de ese país para asignársela a cada usuario, quedando de esta manera completada la información de cada usuario.

De igual manera se generaron los mensajes, para ello se definieron los atributos y las dependencias entre ellos, según la Fig. 2.

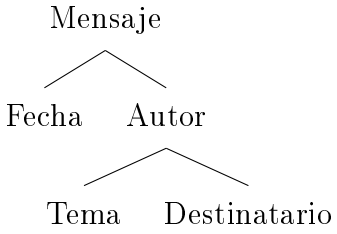


Figura 2: Dependencia entre los atributos de un mensaje

En el caso de la fecha, se tomó la decisión de simular un intervalo temporal de 30 días. Respecto al autor, se decidió que cada usuario pudiera ser autor de entre 5 y 40 mensajes, para así abarcar tanto a los que son muy poco activos, como a los que participan de forma diaria.

Para los temas, se cuenta con los 7 identificados por los expertos: educación, tolerancia, inmigración, imagen personal, empleo, acoso y privacidad [94]. Sin embargo,

elegir qué características de los usuarios influyen en los temas de los que hablan, es algo que queda fuera del alcance de este proyecto. Para simular la dependencia, según cada país, se asignaron de manera aleatoria las siguientes probabilidades a los anteriores temas: 0.05, 0.05, 0.10, 0.10, 0.20, 0.20 y 0.30. Este implica que cada país tiene unos temas principales y otros que generan un menor número de mensajes.

El siguiente paso, es establecer el destinatario de un mensaje, esto también es algo complejo de simular al depender de muchos factores; en este caso, se estableció que sean el país, el rol y la edad, los atributos de los que dependa, con la siguiente función que calcula la fuerza del enlace entre dos usuarios.

$$F(a,b) = 30m + (15 - |a_{edad} - b_{edad}|) * 60p \begin{cases} p = 1, & a_{pais} = b_{pais} \\ p = \frac{1}{60}, & a_{pais} \neq b_{pais} \\ m = 0, & a_{rol} = b_{rol} \neq Moderador \\ m = 1, & (a_{rol} \vee b_{rol}) = Moderador \\ m = 2, & (a_{rol} \wedge b_{rol}) = Moderador \end{cases} \tag{2}$$

Como se puede apreciar en la fórmula anterior, la característica más importante que influye en que un usuario conecte con otro es el país al que pertenecen. Esto está orientado a reflejar cómo los usuarios suelen buscar a otros que también hablan su mismo idioma y, en definitiva, que tengan un contexto lo más similar posible. Sin embargo, y como se puede apreciar en la Fig. 3, la fórmula anterior crea comunidades muy compactas basadas en el país del usuario. Para limitar esto, se estableció otra restricción, centrada en reducir el número de usuarios con los que entra en contacto cada uno de ellos. Esta limitación implica que la función anterior sólo se calcula si el identificador de usuario de a es múltiplo del de b o viceversa. El resultado de su aplicación queda patente en la Fig. 4.

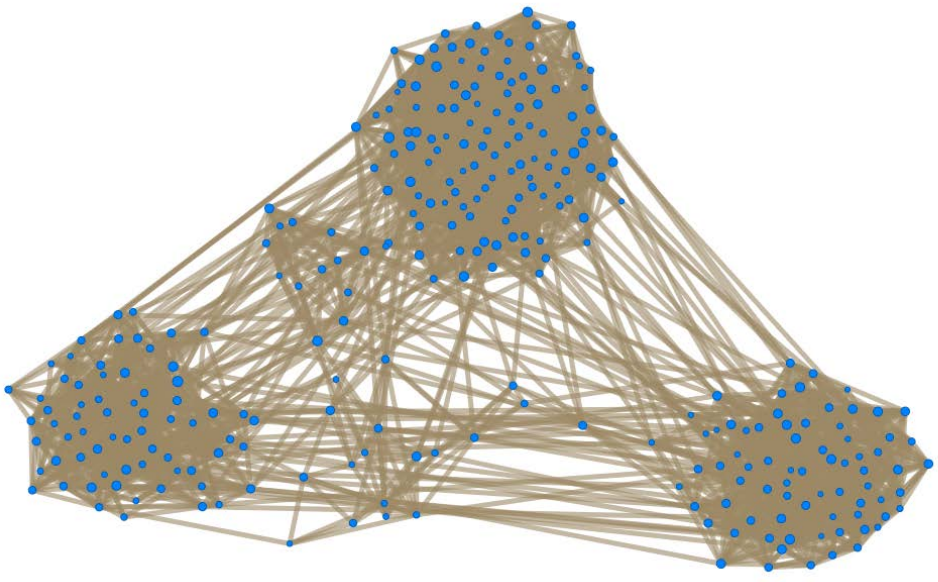


Figura 3: Comunidades compactas por países

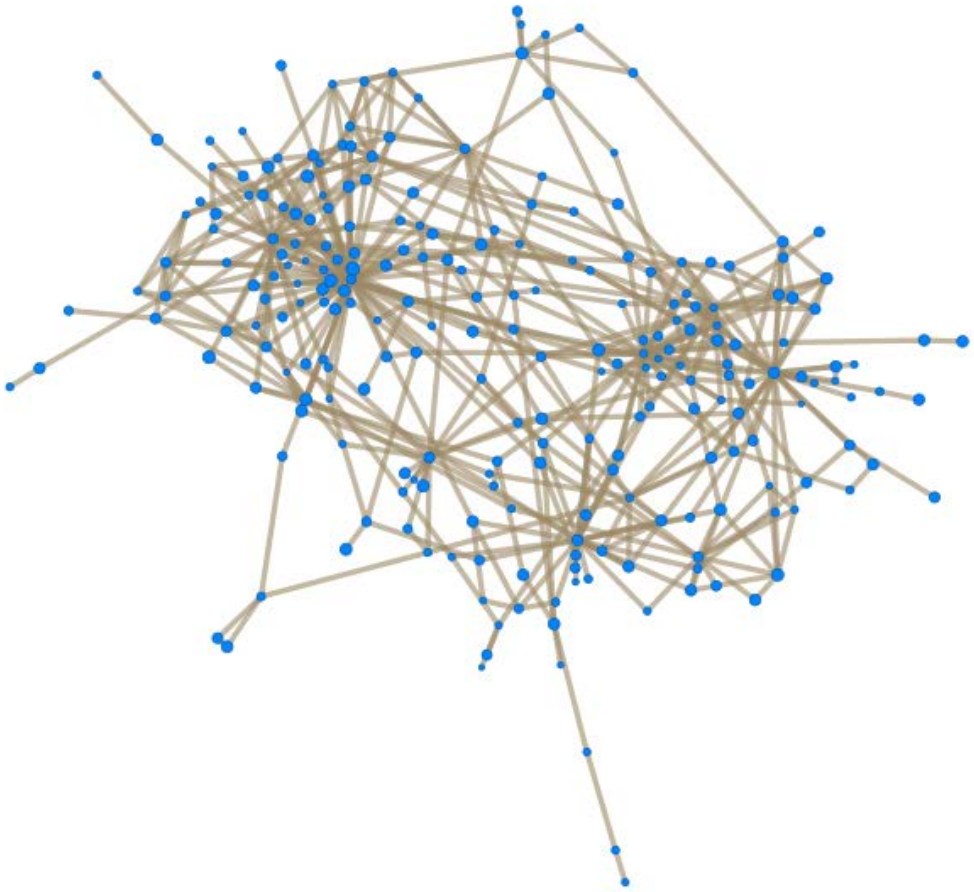


Figura 4: Dispersión de las comunidades al variar el cálculo de los enlaces

Para finalizar, sólo queda establecer el número de usuarios que van a ser generados. El valor de este dato se fijó teniendo en cuenta las previsiones del proyecto (300 usuarios).

5. Propuesta de arquitectura

Una propuesta de arquitectura *software* consiste en definir cada uno de los elementos de un sistema y cuál va a ser el modo en el que interaccionan los mismos. Este tipo de trabajo se vuelve necesario cuando se plantea la realización de un proyecto de cierto tamaño, ya que en él están presentes multitud de requisitos que se deben cumplir, para alcanzar un alto grado de satisfacción de los usuarios. En caso de no establecerla, se corre el riesgo de que el proyecto no permita alcanzar todos los objetivos propuestos y/o la calidad del resultado sea muy baja.

La calidad toma un rol muy importante en este trabajo, ya que el objetivo final del mismo sería desarrollar la arquitectura que aquí se propone. Lo cual implicaría que pasaría a ser usada por un amplio conjunto de usuarios, en un entorno en explotación, formando parte del proyecto WYRED.

La arquitectura de este proyecto tiene que soportar un gran número de requisitos, siendo los principales:

- La capacidad de trabajar con distintas fuentes de datos.
- El soporte para gestionar la privacidad de los mismos.
- El análisis automático de los datos (en la medida de lo posible).
- La capacidad de representar los datos mediante visualizaciones interactivas.

Para soportar estos requisitos, se ha decidido utilizar una arquitectura denominada de micronúcleo [95]. Esta arquitectura se basa en ofrecer una funcionalidad mínima en el núcleo, y complementar al mismo con un conjunto de componentes que son los que realizan las tareas requeridas por los usuarios. Este modelo presenta un cambio de filosofía respecto al patrón de capas, que se caracteriza por apilar las capas de manera horizontal, teniendo cada una ellas un rol específico dentro de la aplicación.

Aunque la arquitectura basada en capas se ha venido utilizando en la mayoría de los desarrollos [96], dando lugar a arquitecturas como MVC (Modelo Vista Controlador) [97] o MVVM (Modelo Vista Modelo de la Vista) [98] muy usadas en el desarrollo web y de escritorio. La arquitectura de micronúcleo ha incrementado su uso, y hoy está presente en multitud de desarrollos [99], siendo Docker [100] uno de los mayores referentes. En la Fig. 5 se puede apreciar su arquitectura.

La gran ventaja de aplicar esta arquitectura en este caso es que el núcleo solo se va a encargar de obtener los datos y anonimizarlos, siendo cada uno de los componentes, los encargados de procesar esos datos y realizar la visualización correspondiente. Esto además permite conseguir una arquitectura muy flexible, donde se pueda añadir fácilmente nuevas visualizaciones o eliminar alguna de las existentes, en el caso de que sus resultados no fueran satisfactorios [101].

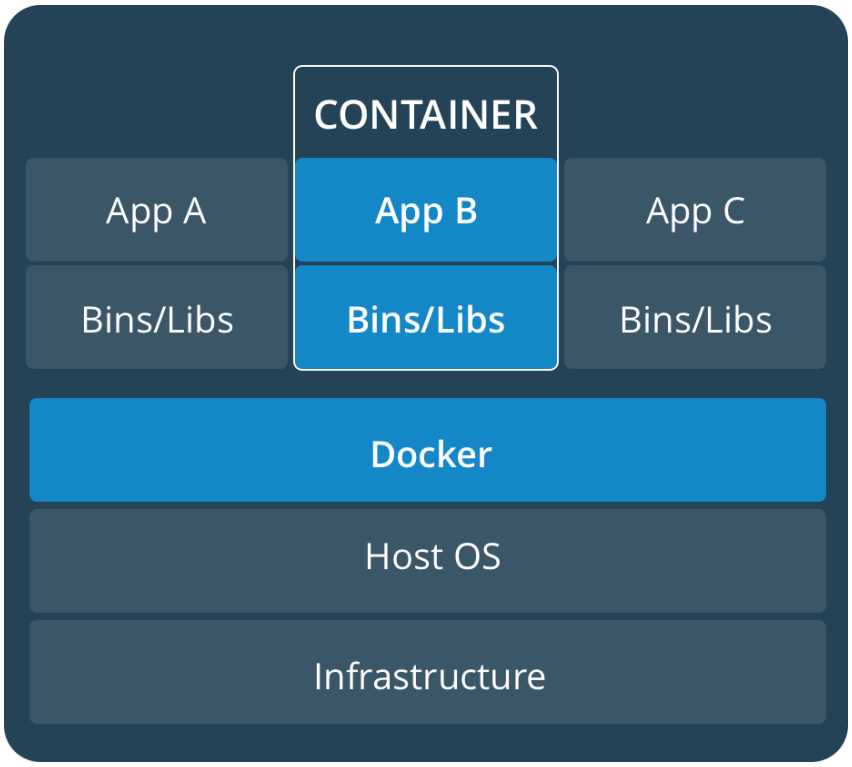


Figura 5: Arquitectura de micronúcleo de Docker

Tomando como referencia la arquitectura de Docker se ha diseñado esta propuesta, que consta de dos capas que forman el micronúcleo y dos capas principales para cada uno de los componentes, que darán lugar a la generación de las visualizaciones interactivas.

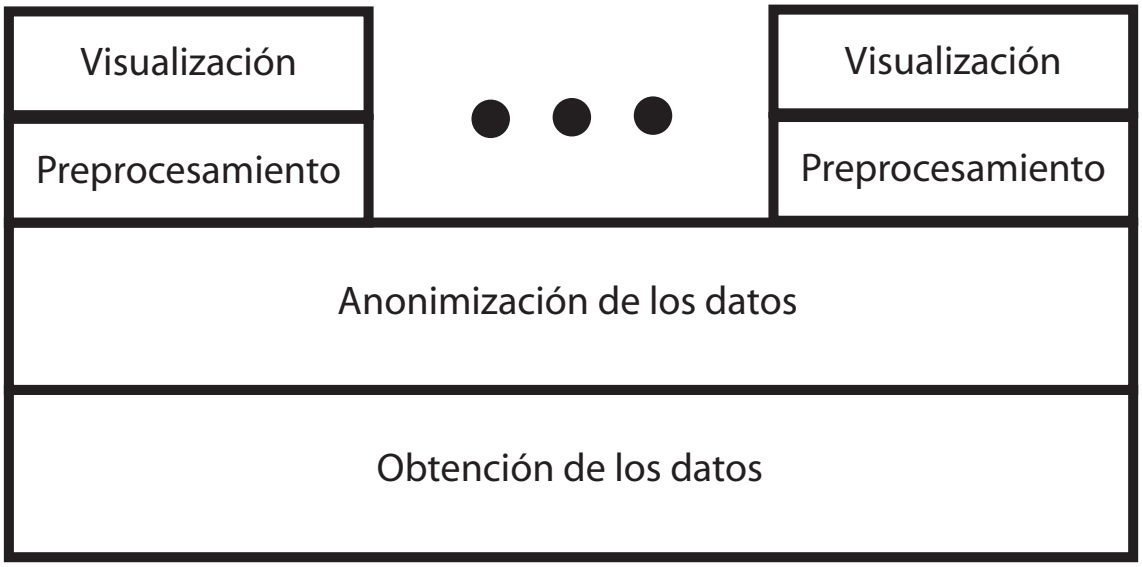


Figura 6: Esquema de la arquitectura propuesta

5.1. Obtención de los datos

La obtención de los datos en este proyecto implica algo más que unas simples consultas a una base de datos. Esto es debido a que la información del mismo se encuentra distribuida entre varios servicios, presentes en varias máquinas.

Al ser uno de los objetivos mantener la privacidad de los usuarios, se decidió guardar esta información crítica, junto con el acceso de los usuarios, en un CAS (*Central Authentication Service*) [102]. Esta decisión está respaldada por algunos estudios que se han enfrentado a este problema [103] [104]. Para aislar todavía más el acceso a esta información, se estableció que el servicio se ejecute en su propio servidor, independiente del resto del proyecto.

En el caso de la información pública de los usuarios, esta forma parte de la plataforma WYRED y está disponible en su base de datos. Finalmente, la información de la interacción de los usuarios con la plataforma, se almacena en una base de datos NoSQL, para afrontar de manera satisfactoria los problemas de escalabilidad [105] [106].

Esta capa del micronúcleo, por tanto, tendrá que encargarse de fusionar los datos desde los distintos medios, además de la recuperación de la información.

5.2. Anonimización de los datos

La capa encargada de anonimizar los datos es de vital importancia en este trabajo, al manejar datos que contienen información personal de los usuarios. Además, muchos de los usuarios son menores, por lo que este proceso es de obligado cumplimiento para acatar la legislación de protección de datos vigente.

La manera de trabajar con parte de estos datos es sencilla, ya que datos como el nombre, los apellidos o su correo electrónico, pueden ser eliminados sin perder información representativa. Sin embargo, esto no es suficiente para asegurar que los datos ya estén anonimizados, ya que mediante la combinación de los datos restantes puede ser posible identificar al usuario inicial.

En el estudio realizado por Sweeney en 2002 [107], el autor muestra cómo es posible realizar el proceso de identificación partiendo de un conjunto de datos anonimizados. Para ello, parte de la tenencia de los datos de una persona y encuentra que: sólo hay, en el conjunto de datos, 6 personas con su misma fecha de nacimiento, sólo 3 de ellas son varones y sólo una vive en su distrito. Esto implica que el usuario ha sido identificado dentro del conjunto de datos anonimizado, mediante el uso de tres datos: la fecha de nacimiento, el género y el código postal que, en un principio, no identifican únicamente a un usuario. A este tipo de datos, que no son identificadores únicos, pero que tienen valores que no se suelen repetir (o su índice de repetición es bajo) en un conjunto de datos, se les denominan quasi-identificadores.

La propuesta para la anonimización de los datos consiste en analizar y detallar cuáles son los atributos quasi-identificadores que se van a tener, e intentar reducirlos:

- En el caso de la fecha de nacimiento, se propone transformar este dato en el año de nacimiento. De esta manera el número de usuarios con un valor único para este campo será muy reducido o nulo. Este cambio si genera pérdida de información, pero se considera que es asumible en el contexto actual.
- En el caso del lugar de residencia, se planta realizar un proceso similar, reduciendo la información a la provincia desde donde se participa.

Además de estas transformaciones, se propone que los resultados que se ofrezcan sean siempre k-anónimos con un valor de k=2. Lo que significa que no pueden existir registros con valores únicos, ya que, como mínimo, de cada registro deben existir 2 usuarios con iguales valores. En la Tabla 8 se puede ver un ejemplo de cómo serían los datos una vez anonimizados.

Año de nacimiento	**Género**	**Lugar**	**Tema de interés**
1994	M	Salamanca	Música
1992	F	Madrid	Tecnología
1994	M	Salamanca	Música
1980	M	Cuenca	Deportes
1992	F	Madrid	Tecnología
1980	M	Cuenca	Deportes

Tabla 8: Ejemplo de datos k-anónimos con k=2

Al aplicar este proceso de anonimización, podría ocurrir que sólo hubiera un usuario con los 4 atributos iguales. En este caso, habría que descartarlo ya que si no se procediera de este modo, se pondría en peligro la privacidad del mismo. La eliminación de este registro implicaría una pérdida de información, sin embargo, se considera que en este contexto tiene mayor prioridad mantener los datos anónimos. Esto es debido, a que también los datos serán publicados de manera abierta, para que otros investigadores puedan utilizarlos como fuente de información en sus investigaciones, tal y como estipula la Unión Europea para los proyectos financiados bajo el paraguas del Horizonte 2020 [108].

5.3. Módulo para el análisis de los temas más frecuentes

El análisis de los temas más frecuentes es una de las cuestiones más repetidas por los distintos investigadores. Algunos únicamente se centran en la evolución temporal de los mismos para poder responder a preguntas como: ¿sigue siendo relevante este tema?, o ¿la presencia de este tema se rige por algún patrón?, sin embargo, otros investigadores consideran también muy importante la capacidad de poder explorar el uso de estos temas atendiendo a las características de los individuos (edad, género, país, etc.).

5.3.1. Extrayendo los temas más frecuentes

Gracias a la arquitectura propuesta anteriormente, este módulo es capaz de acceder a los datos de la plataforma para poder preprocesarlos. En este caso, se plantea realizar un análisis automático de los temas más frecuentes utilizando para ello LDA (*Latent Dirichlet Allocation*) [109]. LDA es un modelo probabilístico generativo capaz de extraer, a partir de un corpus documental, los principales temas presentes en esos textos. Este sistema se basa en representar cada documento como una mezcla de temas, para los cuales hay asociadas una serie de palabras con una cierta probabilidad [110]. Siguiendo esta idea, se asume que un texto habría sido escrito siguiendo los siguientes pasos:

1. Se fijaría, en primer lugar, el número de palabras que va a tener el texto.
2. Se escogería el número de temas que van a estar presentes, y la proporción de los mismos.
3. Se elegirían las palabras, siguiendo este procedimiento para cada una de ellas:
    a) Se seleccionaría el tema correspondiente, teniendo en cuenta las probabilidades asignadas en el paso 2.
    b) Se tomaría una palabra relacionada con ese tema.

Partiendo de la idea anterior, LDA realiza el proceso inverso, es decir, parte del texto ya generado, e intenta llegar a conocer los temas que lo forman y en qué proporción intervienen.

Uno de los problemas que tiene este método, es que no está pensado para trabajar en sistemas multilingües, cuestión muy importante al ser una de las características del contexto de uso, sin embargo, algunos autores [111][112] han propuesto diversos métodos para poder soportarlo. Otro de los hándicaps de este mecanismo es que es capaz de agrupar las palabras que forman parte de la misma temática, pero no de asociar un nombre representativo a cada tema. Este proceso se podría hacer de manera manual, o de forma automática utilizando un sistema que para cada palabra tenga presente sus temas principales.

5.3.2. Visualización propuesta

Para realizar la propuesta de visualización, lo primero que se ha tenido en cuenta son sus principales tareas asociadas:

- Conocer la evolución de una temática: máximos, mínimos, patrones, etc.
- Poder comparar la evolución de varios temas.
- Ser capaz de conocer cómo influyen los atributos de los usuarios en la evolución de los temas.

El siguiente paso fue la elección del tipo de gráfico a representar. Una decisión no trivial al existir multitud de maneras de representar los datos, como atestiguan algunos trabajos [2]. Por la importancia de la característica temporal, la primera decisión fue utilizar una representación que dispusiera de un eje horizontal donde poder mostrar cada uno de los instantes temporales. Pero todavía era necesario indicar cómo se iba a codificar la frecuencia de un tema, para lo cual había varias posibilidades como los gráficos de líneas, de áreas, o los histogramas. Se descartó esta última opción ya que las barras rompían con la idea de representar la evolución temporal, al mostrar los valores para instantes concretos. Por esta misma razón, se optó por un gráfico de áreas.

En un principio se pensó utilizar una visualización basada en el concepto de *Theme River* [69]. Este sistema ya se había usado de manera efectiva en otros trabajos [71] [113], ya que permite identificar de manera sencilla los cambios de tendencia más importantes. Sin embargo, se ha demostrado que no es útil para detectar cambios de tendencia menores y, además, no permite representar un número amplio de temas. En la Fig. 7 se puede ver un ejemplo de su uso.

Otra propuesta que se consideró al realizar la revisión sistemática de la literatura, era la posibilidad de representar cada tema de manera individual, sobre líneas temporales paralelas [114]. Este mecanismo permite apreciar con mayor claridad la evolución de cada uno de ellos. Por esta razón, se propone utilizar un enfoque combinado, en el que se aprovechen las ventajas de la representación *Theme River*, a la hora de realizar las comparaciones, y de las representaciones con líneas temporales paralelas, a la hora de conocer en mayor profundidad la evolución temporal del tema y permitir representar un número mayor de ellos. Respecto a esto último, se decidió utilizar sólo 5 temas debido a que, si se eligiera un número mayor, se corre el riesgo de aumentar la carga cognitiva del usuario, lo cual siempre hay que intentar evitar, para conseguir una mayor efectividad en las visualizaciones.

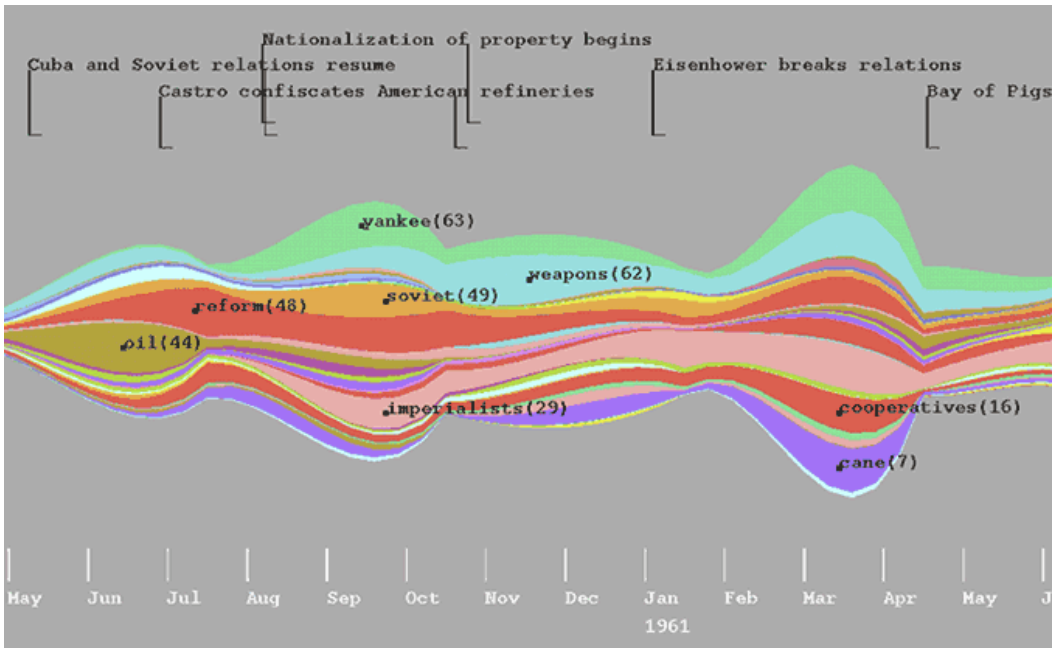


Figura 7: Ejemplo de *Theme River* [1]

El color también es una parte importante de cualquier visualización, y como se ha demostrado en diversos estudios [115] [116] es importante no usar un número elevado de colores, siendo entre 5 y 9, los que un usuario es capaz de distinguir. Estos deben ser lo más diferentes posible, para aumentar la capacidad de distinguirlos fácilmente por un usuario. Además, se requiere de colores más intensos cuanto menor es el tamaño de la representación. Debido a lo anterior, para la selección de la paleta de color se optó por utilizar una que ya ha sido probada con éxito en la visualización de mapas cartográficos [117], ya que es uno de los campos donde la codificación por color tienen mayor impacto.

Figura 8: La paleta de color elegida para representar a los temas

Para comparar cómo afectan las características de los usuarios a los propios temas, se decidió seguir utilizando el mismo color para cada tema, usando una escala de degradados para representar los valores de estos atributos, como se puede apreciar en la Fig. 9.

Figura 9: Gradación seleccionada para comparar un tema según un atributo concreto

Respecto a las capacidades de interacción y adaptación del gráfico, se proponen usar las siguientes:

- Capacidad de seleccionar los temas a representar.
- Posibilidad de realizar comparaciones, eligiendo el atributo de los usuarios por el cual se compara.
- Soporte para reordenar los temas, ya que es más sencillo comparar aquellos que están más próximos entre sí.
- Capacidad de conocer el nivel de relevancia de ese tema en un instante concreto.
- Posibilidad de hacer *zoom* de manera automática.
- Capacidad de restringir la selección a un periodo temporal.

5.4. Módulo para la detección de comunidades

Los usuarios al utilizar una comunidad, implícitamente empiezan a tejer pequeñas redes, debido a las interacciones mutuas (mensajes, seguimientos, contestaciones, etc.). Estas redes suelen proporcionar una gran cantidad de información a los investigadores, una vez que se analizan su topología y su contenido, por lo tanto, el análisis y la detección de las distintas redes o subcomunidades es una de las tareas de investigación que aparecen de forma más recurrente en este campo.

Para la detección de comunidades, se han utilizado multitud de técnicas [20]:

- *Clustering* jerárquico para agrupar a los usuarios más comunes y detectar los grupos más compactos.
- Detección de los nodos centrales de los grupos, mediante el cálculo de los grados de entrada y salida y el coste de alcanzar a todos los usuarios desde cada uno de ellos.
- Detección de comunidades mediante el cálculo de la centralidad.

Sin embargo, estas se basan en la ejecución de complejos algoritmos, que requieren un elevado tiempo de computación, para calcular cómo se relaciona un nodo con el resto. Además de estos métodos numéricos, esta tarea también puede ser resuelta de manera efectiva mediante una visualización.

5.4.1. Visualización propuesta

La principal tarea que se quiere abordar con esta visualización es descubrir cómo interaccionan los usuarios, para poder intuir las comunidades implícitas que los mismos forman. Por esta razón, la representación mediante grafos se ha destacado como el mejor sistema para visualizar este tipo de datos. Esta representación está compuesta de dos componentes principales, los nodos o vértices y los arcos o enlaces, que representan a los usuarios y sus relaciones, respectivamente. En el caso concreto de la visualización que se propone, las relaciones hacen referencia al número de comentarios que intercambian.

Elegir como sistema de presentación los grafos, tiene diversas ventajas:

- Es una representación que ha sido muy estudiada y se conoce en profundidad.
- Es el mecanismo usual de representación de este tipo de datos [15].
- Permite de manera rápida identificar grupos de nodos.
- Es capaz de detectar nodos con comportamientos anómalos, que no tienen relación con la comunidad [118].
- Es una representación fácilmente interpretable.

- Es un modelo muy flexible.

Aprovechando esta última característica, se plantea codificar cada nodo con un tamaño relativo al número de mensajes que ha publicado en la plataforma. Respecto a los enlaces, su longitud debe representar la cercanía o lejanía de un nodo respecto a otro, atendiendo a las interacciones que han tenido. En una primera aproximación, se podría calcular la longitud o distancia (d) entre dos nodos A y B como un valor proporcional a la inversa de los enlaces que comparten:

$$d(A, B) = k * E(A, B)^{-1} \tag{3}$$

Siendo k la constante de proporcionalidad y E(A,B) la función que, para dos nodos cualesquiera (A y B), calcula la suma del número de enlaces que van de A a B y viceversa.

Sin embargo, esta propuesta no refleja fielmente lo cercano que está un nodo de otro, ya que favorece a los usuarios que más usan la plataforma, al utilizar el número de enlaces absolutos entre dos nodos. Esto queda demostrado en la Fig. 10, ya que los nodos A y B están más relacionados que los nodos C y D, al compartir una mayor proporción de enlaces.

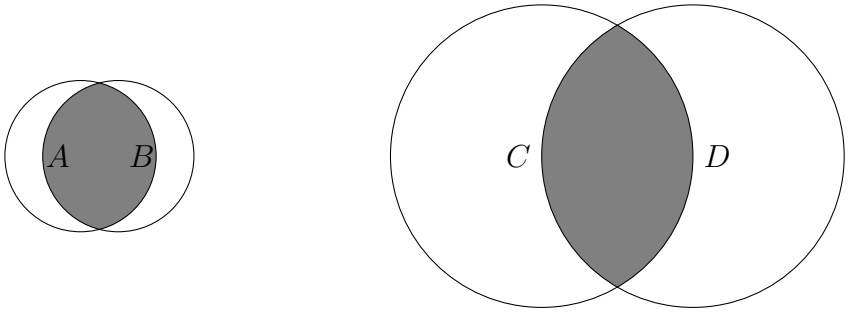


Figura 10: Cercanía de los nodos según sus relaciones

Para corregir este problema, se ha introducido el concepto de distancia relativa d_r entre dos nodos A y B, como un valor proporcional al número de enlaces totales entre el número de enlaces que comparten:

$$d_r(A, B) = k * \frac{g(A) + g(B)}{E(A, B)} \tag{4}$$

Siendo $g(A)$ el grado de A, es decir, el número de enlaces que tienen como origen o destino A.

Para la elección de los colores, se ha optado por utilizar un color azul intenso para los nodos, lo cual, como se ha comentado en el apartado anterior, es necesario cuando el área es pequeña, combinándolo con un marrón que colorea los enlaces. Esta combinación, la cual se puede apreciar en la Fig. 11, ha sido obtenida gracias

al proyecto *I want hue* [119], el cual realiza sugerencias acerca de combinaciones de colores que proporcionan un buen índice de contraste y legibilidad.

Figura 11: La paleta de color escogida para los grafos

Respecto a las características de interacción implementadas, para ayudar a resolver de manera sencilla y efectiva la tarea de la detección de subcomunidades, se han establecido las siguientes:

- Posibilidad de conocer en detalle las características de un usuario, al visitar un nodo.
- Capacidad de hacer *zoom* para poder analizar en mayor profundidad el grafo y ayudar a que esta visualización pueda seguir siendo útil con un número elevado de nodos.
- Soporte para seleccionar un conjunto de nodos y conocer el valor medio de sus atributos.
- Posibilidad de mover y analizar en detalle cada una de las comunidades que se formen.

5.5. Módulo para la exploración de los usuarios

Otra de las cuestiones que los investigadores han manifestado que quieren hacer con la plataforma, gira en torno al análisis de los usuarios a través del valor de sus atributos característicos. Esto es necesario para poder analizar cómo afecta uno de ellos a su comportamiento en la plataforma, lo cual es el origen de la mayor parte de estudios sociológicos y de uso que se van a desarrollar.

5.5.1. Visualización propuesta

El problema de representar los atributos de los usuarios de una plataforma es bastante complejo, debido a las siguientes características:

- El número de usuarios puede ser muy elevado.
- El cantidad de características a analizar es alta.
- El uso de la plataforma puede presentar una alta variabilidad a lo largo del tiempo.
- Al ser una comunidad abierta a un público amplio, las características del mismo pueden no estar claramente definidas.

Por estas razones, se hace necesario el uso de una visualización que permita gestionar un amplio número de usuarios y características de los mismos, y que permita que escale bajo demanda. Además, al ser destinada a un público que no es experto en analítica visual, la propuesta que se presente debe ser compacta y sencilla de ser interpretada.

Después de realizar el análisis sistemático de la literatura, se detectó que había una técnica que se usaba de manera recurrente para afrontar estos desafíos. Esta tipo de visualización eran las coordenadas paralelas, las cuales son aplicadas, por ejemplo, para explorar las características de un conjunto de documentos [71].

Las coordenadas paralelas son un sistema de representación propuesto por Alfred Inselberg, capaz de representar n dimensiones en un contexto bidimensional [70] [120]. Para ello, se disponen cada uno de los n ejes de manera vertical y paralela entre ellos. Una vez realizado esto, cada entidad n-dimensional será representada por una polilínea compuesta por cada uno de los segmentos que une el punto del eje que representa el valor de la dimensión actual, con el punto del siguiente eje que representa el valor de la dimensión contigua.

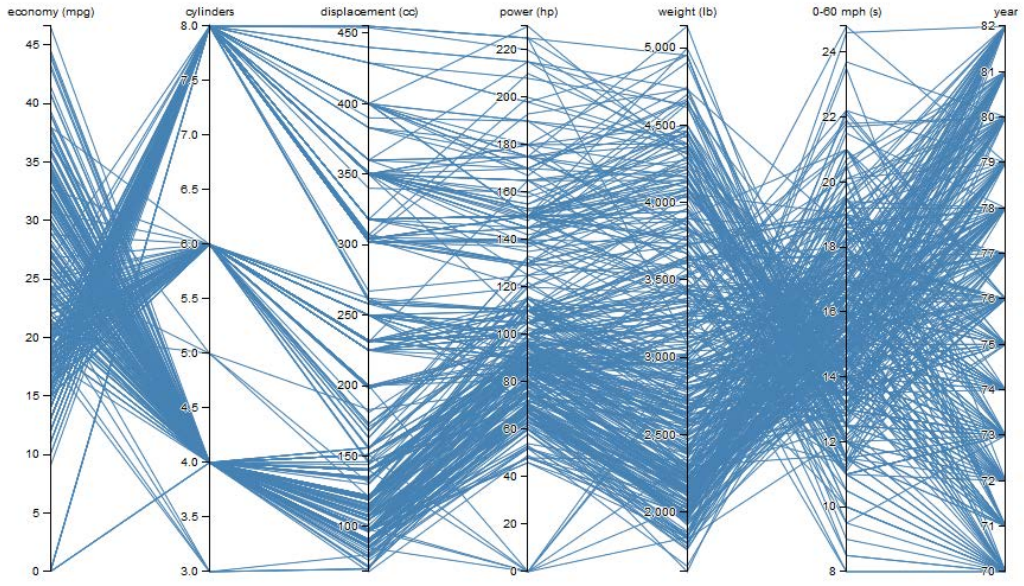


Figura 12: Ejemplo de uso de las coordenadas paralelas[3]

La aplicación de este sistema, también proporciona ventajas adicionales al utilizar esta distribución específica de los ejes. Algunas de ellas son: la facilidad para identificar grupos de entidades con valores similares de los atributos, la detección de desviaciones o entidades con valores anómalos y la capacidad de conocer si hay una correlación entre los valores de dos de los atributos [121].

Respecto a las características de interacción, se proponen las siguientes:

- Posibilidad de reordenar los atributos que van a ser visualizados, para así poder detectar si hay correlación entre ellos o no.
- Capacidad de poder filtrar por cada uno de los atributos, soportando el filtrado múltiple.
- Posibilidad de restringir el periodo temporal a estudiar.

[3]Este ejemplo visualiza las características de los coches de las décadas 70 y 80, estando accesible en `https://bl.ocks.org/jasondavies/1341281`

5.6. Módulo para la exploración geográfica del proyecto

Este módulo se encarga de dar respuesta a la necesidad de conocer cuáles son los países más activos y cómo afecta esta dimensión al análisis de los datos de la plataforma.

5.6.1. Visualización propuesta

En este caso, elegir una visualización era bastante más sencillo que en los casos anteriores, ya que al tratarse de exploración geográfica, la visualización que mejor representa este concepto es el mapa. Esta representación cuanta con las siguientes ventajas:

- Es la forma usual de representar este tipo de datos, así que el usuario está acostumbrado a utilizarlos.
- Permite, de manera sencilla, visualizar un valor asignado a cada país.
- Es capaz de trabajar con distintos niveles de abstracción (países, comunidades, provincias, etc.).
- Es capaz de representar la proximidad entre territorios, lo cual con una tabla es difícil de lograr. Esto es interesante a la hora de comprobar si hay zonas con características comunes.

Por las razones anteriores, queda demostrado que el uso de un mapa es una buena opción, sin embargo, hay multitud de tipos de mapas. La mayoría de ellos utilizan las fronteras entre los distintos países y otras demarcaciones de ámbito estatal, pero en algunos trabajos [122] [123] se prefiere obviar lo anterior y dividir el territorio en otro tipo de zonas.

Además de la manera en la cual se dividen las distintas zonas del mapa, la creación del mismo plantea otra difícil elección, la selección del tipo de proyección que va a ser usada. Esto es debido a que hay que trasladar una superficie esférica en un plano, y para ello es necesario que haya deformaciones, lo que implica que no todas las zonas van a tener la misma forma y tamaño que tienen en la realidad, como se puede ver en la Fig. 13. En la propuesta, se ha elegido utilizar la proyección Mercator, porque esta es a la que más familiarizados están los usuarios al ser utilizada en servicios como Google Maps, Bing Maps u OpenStreetMap.

(a) Proyección cónica de igual área (b) Proyección Mercador

Figura 13: Distintos tipos de proyecciones para realizar un mapa

En este caso, se propone utilizar un mapa que tenga como forma de división principal los países, ya que esto va a permitir comprender mejor los datos. La información que se va a mostrar en el mismo, es el número de mensajes que han sido generados por los usuarios de cada uno de los territorios. Para transmitir esta información, se va a utilizar el color, con el cual va a estar pintado cada uno de los países. La paleta de color elegida, es la propuesta en el módulo de visualización de comunidades, siendo utilizado el marrón para los países cuyos usuarios no han escrito mensajes, y una gradación del azul (más oscuro cuantos más mensajes hay), para los países donde sí ha habido actividad.

Respecto a las características de interacción, se proponen las siguientes:

- Posibilidad de moverse por el mapa.
- Capacidad de tener *zoom* semántico, de tal manera que cuando el nivel de acercamiento sea alto, el mapa deje de representar los países y pase a mostrar sus provincias.
- Soporte para volver a centrar el mapa.
- Capacidad de conocer el número de mensajes exacto en cada país.
- Posibilidad de filtrar los datos por país.

6. Resultados

En esta sección del trabajo se va a analizar cómo se ha llevado a cabo la arquitectura propuesta anteriormente y cómo la misma, junto con los módulos descritos, podrían ser usados para responder algunas preguntas de investigación concretas.

6.1. Realización de la arquitectura propuesta

Para desarrollar la arquitectura propuesta, se ha recurrido a utilizar tecnologías y lenguajes de programación web. Esta decisión, como ya quedó patente en el estado del arte, permite principalmente dos cosas: conseguir que los desarrollos sean accesibles a un mayor público y dotarlos de un mayor grado de interactividad. Dentro de las tecnologías web, se van a destacar especialmente D3 [33], la librería para crear visualizaciones interactivas, y el formato SVG, que posibilita realizar gráficos escalables, ambos aspectos se abordan en el Apéndice C.

Al realizar el desarrollo de cada uno de los módulos, hay un aspecto que ha tomado gran importancia, la capacidad de poder filtrar los datos que se quieren estudiar. Para ello, se ha establecido en la parte superior unos controles destinados a tal fin, lo que ayuda a cumplir con el mantra de la visualización interactiva enunciado por Ben Shneiderman "*Overview first, zoom and filter, then details-on-demand* " [124] y ampliado por Keim, como el mantra de la analítica visual: *Analyze first, show the important, zoom, filter and analyze further, details on demand* [125].

A continuación, se muestra el resultado de la implementación de los módulos de análisis de temas (Fig. 14), exploración de comunidades (Fig. 15), exploración de usuarios (Fig. 16) y exploración geográfica (Fig. 17):

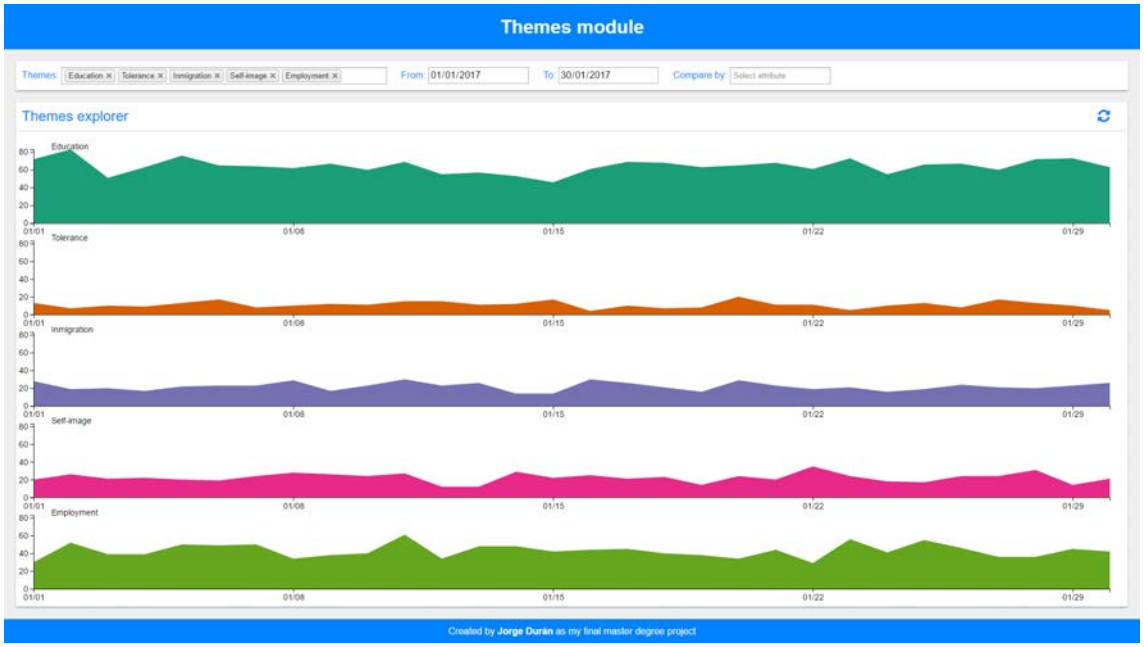


Figura 14: Módulo para el análisis de los temas más frecuentes[4]

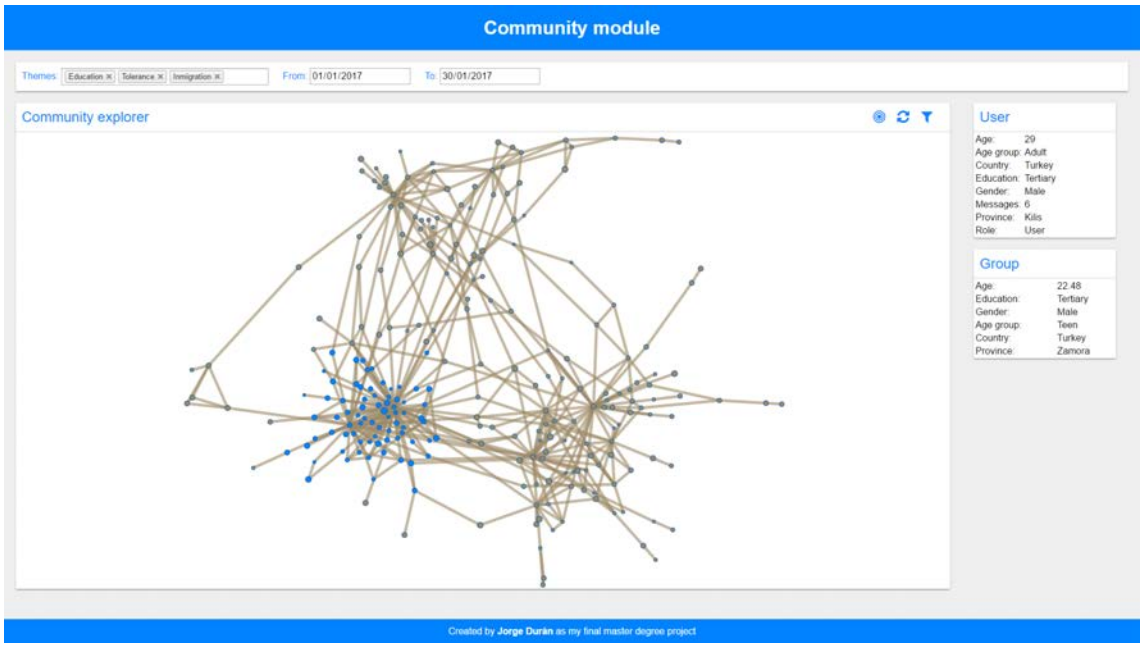


Figura 15: Módulo para la detección de comunidades[5]

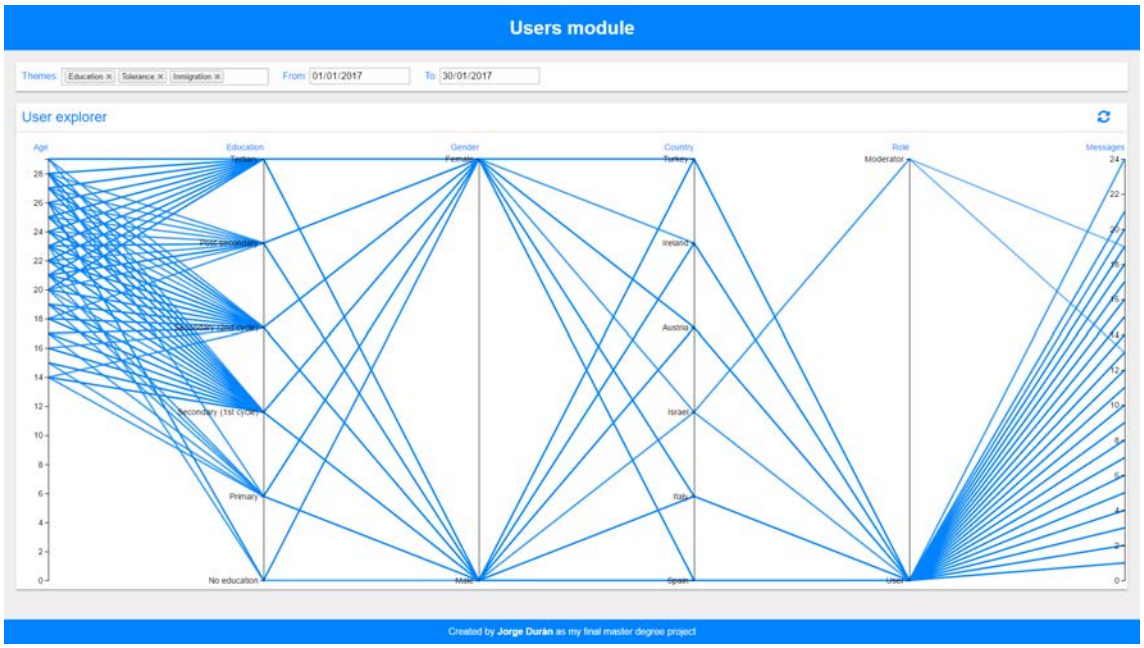


Figura 16: Módulo para la exploración de los usuarios[6]

[4] Accesible en `https://jorge-duran.com/research/tfm/themes/`
[5] Accesible en `https://jorge-duran.com/research/tfm/graph/`
[6] Accesible en `https://jorge-duran.com/research/tfm/parallel/`

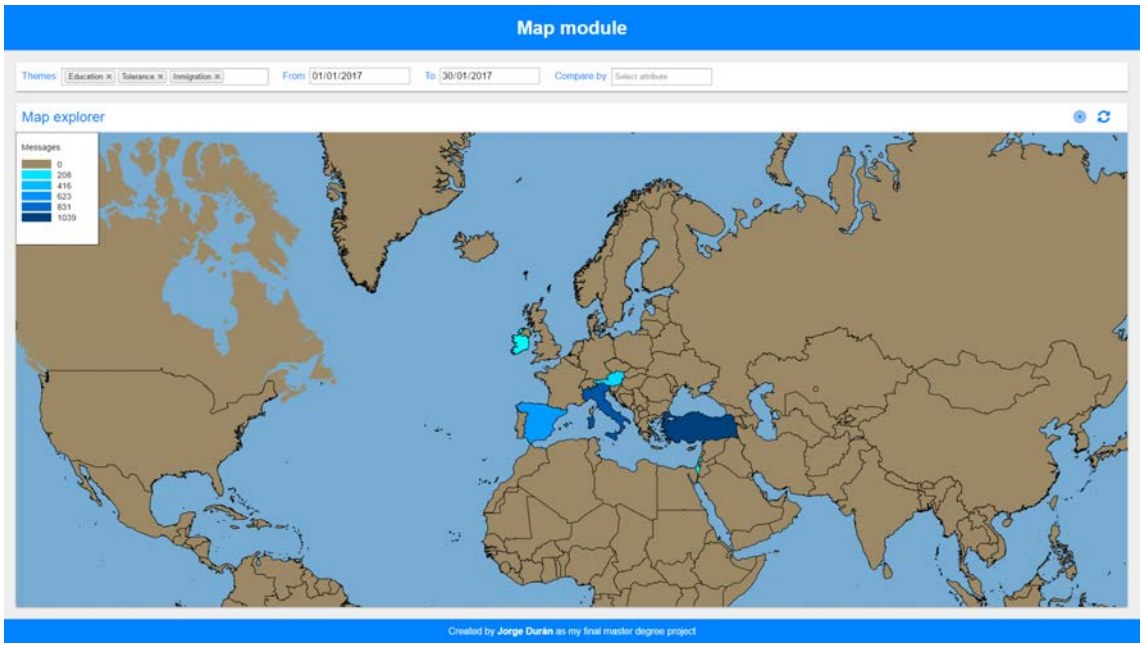


Figura 17: Módulo para la exploración geográfica del proyecto[7]

La utilización de una arquitectura modular no implica, necesariamente, el uso de cada uno de los componentes por separado, por ello han sido combinados mediante la técnica de vistas enlazadas [126], para constituir un panel de monitorización que permita explorar todas las facetas del proyecto, al mismo tiempo.

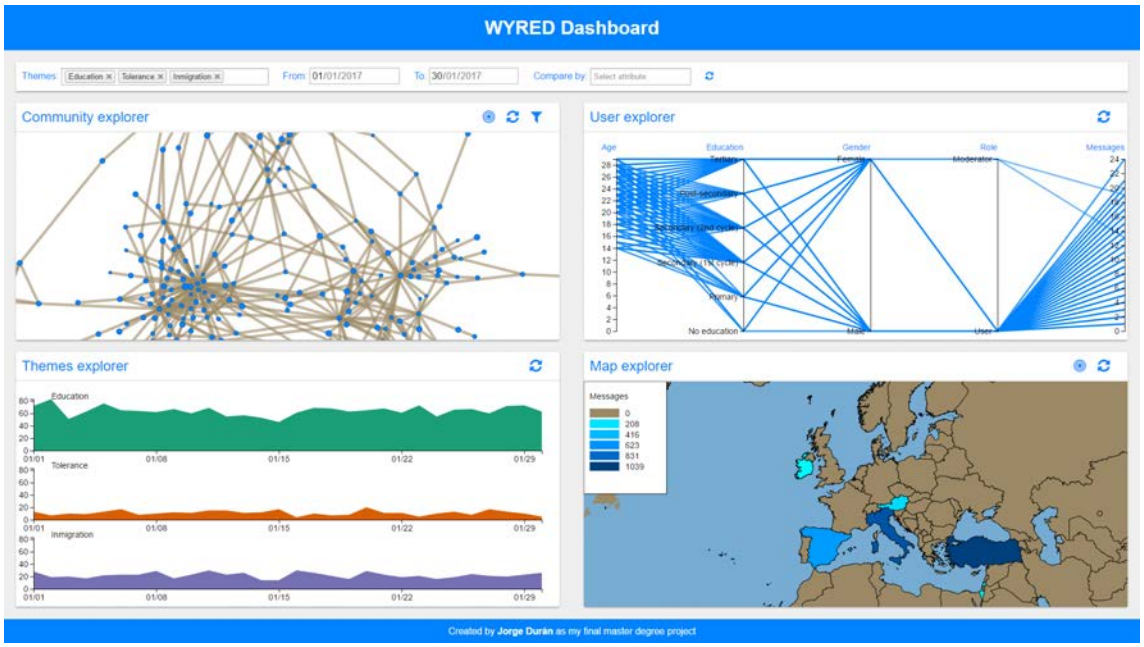


Figura 18: Panel de monitorización del proyecto[8]

[7] Accesible en `https://jorge-duran.com/research/tfm/map/`
[8] Accesible en `https://jorge-duran.com/research/tfm/dashboard/`

6.2. Casos de uso

Bajo este epígrafe se va a describir paso a paso cómo deberían utilizarse las visualizaciones propuestas, para responder a las siguientes preguntas de investigación extraídas de la consulta a expertos realizada anteriormente:

1. ¿Cuáles son las principales comunidades sobre educación y empleo y qué características tienen?
2. ¿Quiénes son los usuarios más activos de Turquía hablando sobre privacidad?
3. ¿Cómo influye el género a la hora de hablar sobre la tolerancia y la inmigración?
4. ¿Cuál es la evolución temporal de las discusiones sobre acoso, según los países participantes?

6.2.1. ¿Cuáles son las principales comunidades sobre educación y empleo y qué características tienen?

La primera pregunta versa sobre dos temas educación y empleo, por ello lo primero que hay que hacer es seleccionar estos temas para ser analizados en nuestra visualización, como se puede observar en el extremo izquierdo de la Fig. 19.

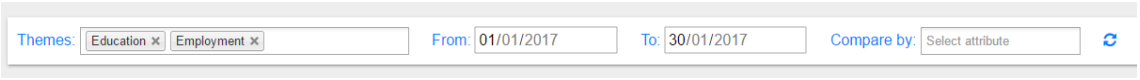


Figura 19: Selección de los temas para la pregunta de investigación 1

Una vez seleccionados los temas, en el explorador de comunidades se pueden identificar las principales comunidades formadas al ver los grupos más cohesionados. Como se puede apreciar en la Fig. 20, las mismas han sido enmarcadas y numeradas para su posterior análisis.

Para analizar una de las comunidades en profundidad, sólo hay que seleccionar los usuarios que pertenecen a la misma, haciendo clic y extendiendo el área de selección hasta cubrirlos a todos. Según cambian los seleccionados, el resto de visualizaciones se actualizan para mostrar la información de estos usuarios, como se puede ver en la Fig. 21, donde figuran en azul los nodos seleccionados en el explorador de comunidades y sus respectivos valores en el resto de visualizaciones enlazadas. Esta primera comunidad está formada principalmente por turcos que se decantan por hablar más de educación que de empleo. Esto se aprecia en la Fig. 21 al ser el color de Turquía más oscuro y alcanzar un mayor número de mensajes sobre educación, en el explorador de temas.

En el caso de la comunidad 2, como se puede consultar en la Fig. 22, los usuarios son principalmente españoles, lo cual se aprecia por el tono oscuro que presenta España en el mapa, que en momentos puntuales se centran en el empleo, pero de

manera más continua, hablan más de educación, como se aprecia en el explorador de temas.

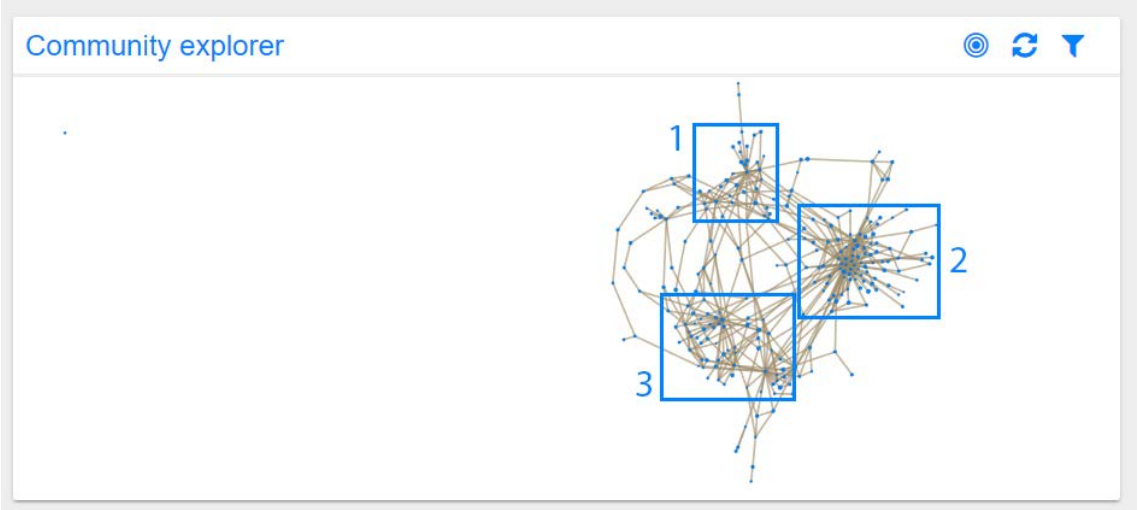


Figura 20: Identificación de las principales comunidades para la pregunta de investigación 1

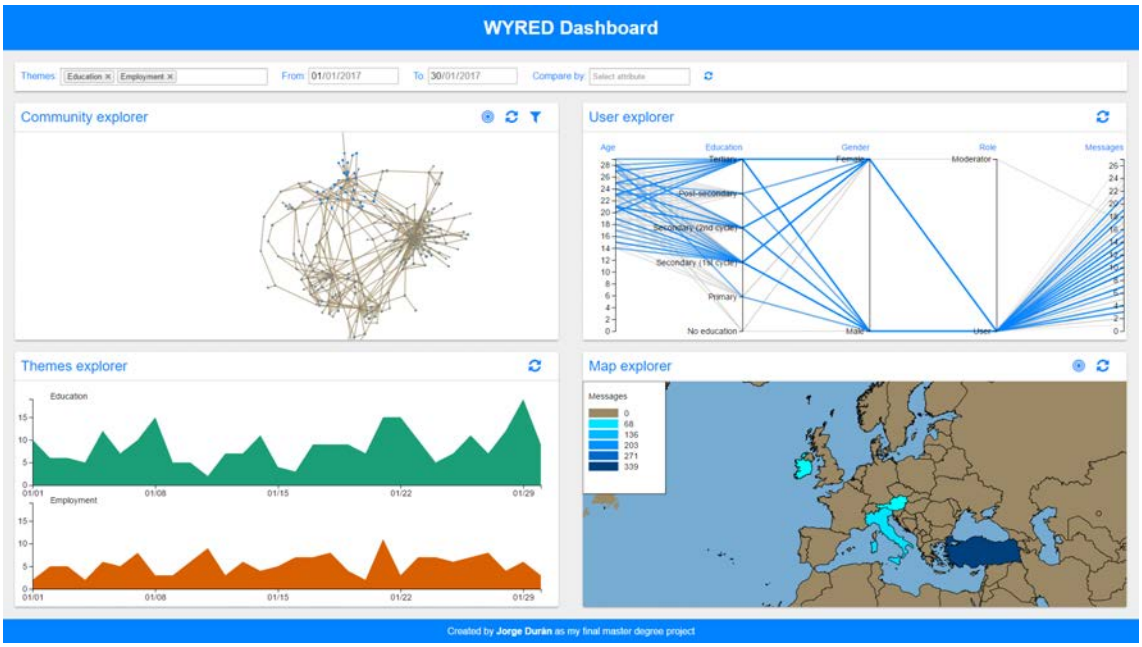


Figura 21: Visualización de los datos de los usuarios de la primera comunidad para la pregunta de investigación 1

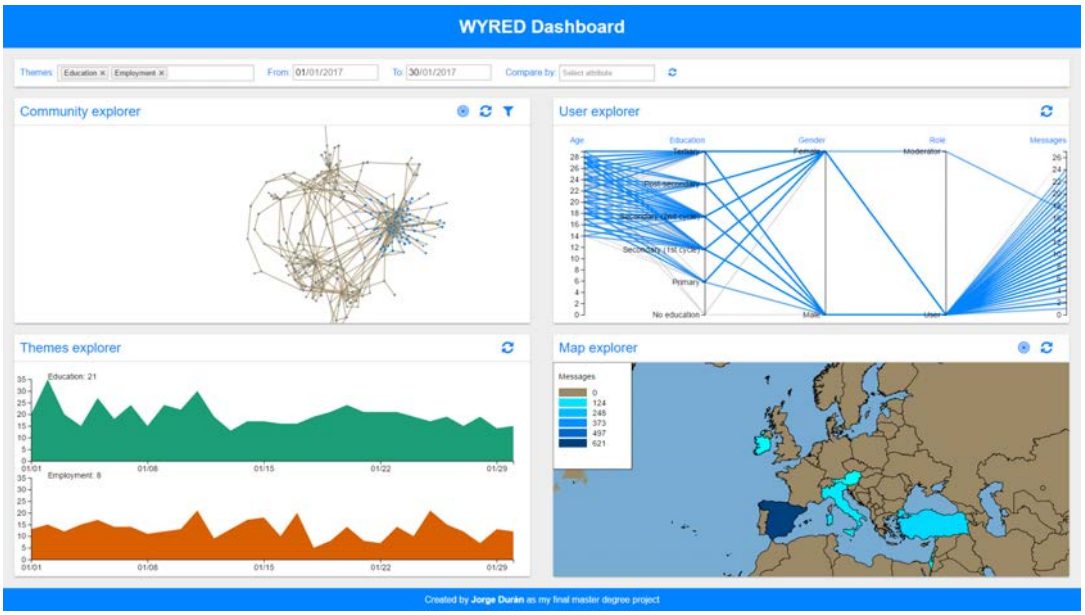


Figura 22: Visualización de los datos de los usuarios de la segunda comunidad para la pregunta de investigación 1

Si se analiza la comunidad 3, sus usuarios son principalmente italianos, con un comportamiento similar a los españoles en el uso de ambos temas. Sin embargo hay un valor anómalo que identifica a la comunidad, el reducido número de estudiantes que cuentan con educación postsecundaria, lo cual se puede aprecia en la visualización de los datos de los usuarios (Fig. 23), al presentar la mayor parte de las líneas con destino a educación postsecundaria en gris, lo que denota que estos usuarios no forman parte de esa comunidad.

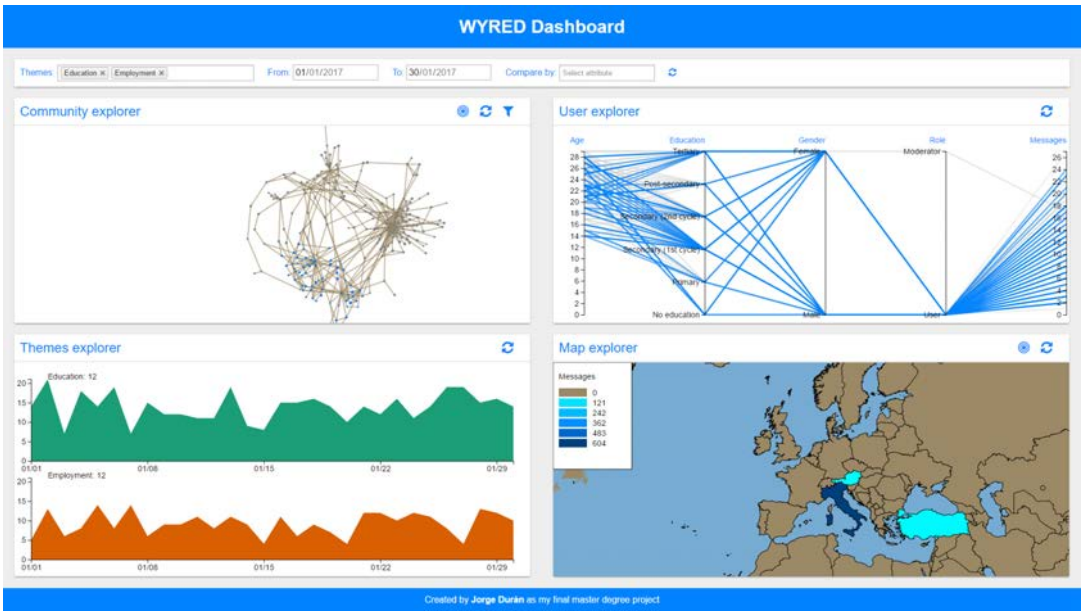


Figura 23: Visualización de los datos de los usuarios de la tercera comunidad para la pregunta de investigación 1

Para apreciar las características interactivas y cómo utilizaría un investigador estas visualizaciones para resolver este caso de uso, se puede visualizar el siguiente vídeo `https://jorge-duran.com/research/tfm/videos/mainCommunities.mp4`.

6.2.2. ¿Quiénes son los usuarios más activos de Turquía hablando sobre privacidad?

Para dar respuesta a esta pregunta, lo primero que hay que hacer es centrar el estudio sobre este tema (la privacidad) y descartar los demás, como se aprecia en la zona izquierda de la Fig 24.

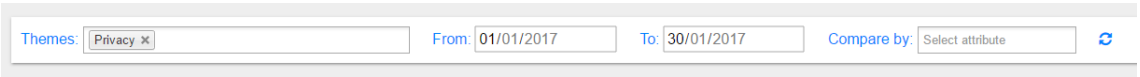


Figura 24: Selección del tema para la pregunta de investigación 2

El siguiente paso es descartar los datos de los usuarios que no son de Turquía, para ello se hace clic dentro de este país, Fig. 25.

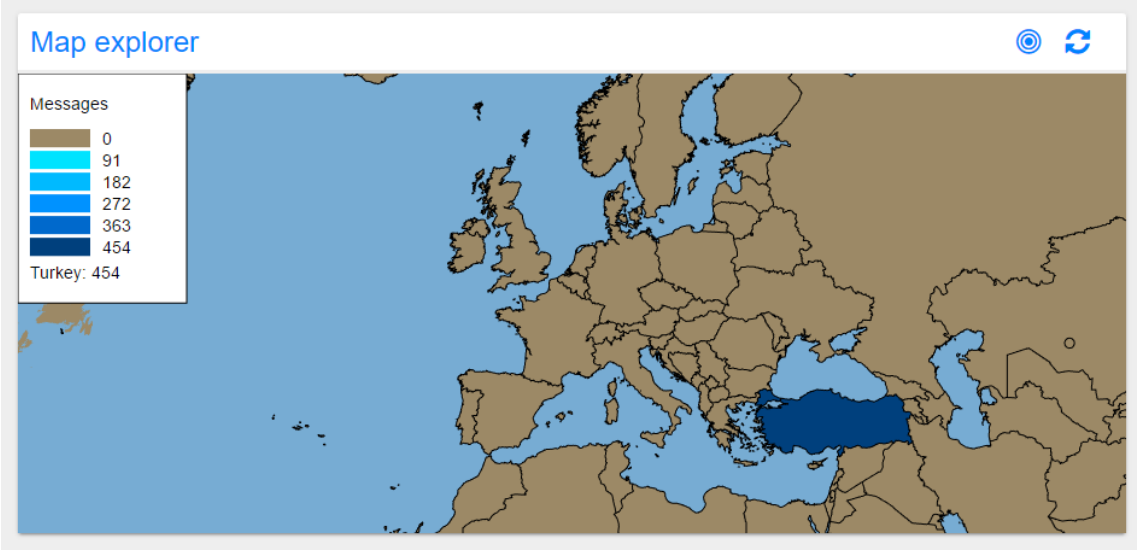


Figura 25: Selección del país para la pregunta de investigación 2

A continuación, se restringen los usuarios a los más activos en este campo, por ejemplo, indicando que tengan 10 mensajes o más sobre privacidad, para ello, y tal y como se aprecia en la Fig. 26, es necesario seleccionar en el eje de mensajes del explorador de usuarios los valores iguales o superiores a 10. Sobre sus características podemos decir que principalmente son mujeres, mayores de edad y que ninguno de ellos actúa como moderador. Para extraer estas conclusiones hay que fijarse en las líneas azules que representan a los usuarios seleccionados.

Finalmente, se puede ver cómo se relacionan con la comunidad al visualizarlos resaltados en el grafo. En la Fig. 27 se puede apreciar como uno de ellos forma parte

de un pequeño grupo muy cohesionado.

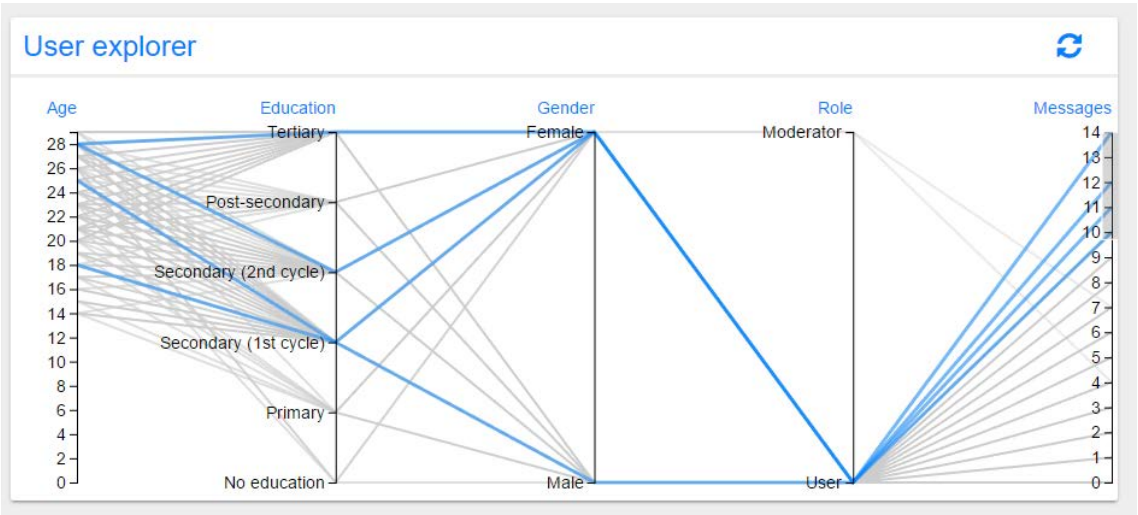


Figura 26: Selección de los usuarios más activos para la pregunta de investigación 2

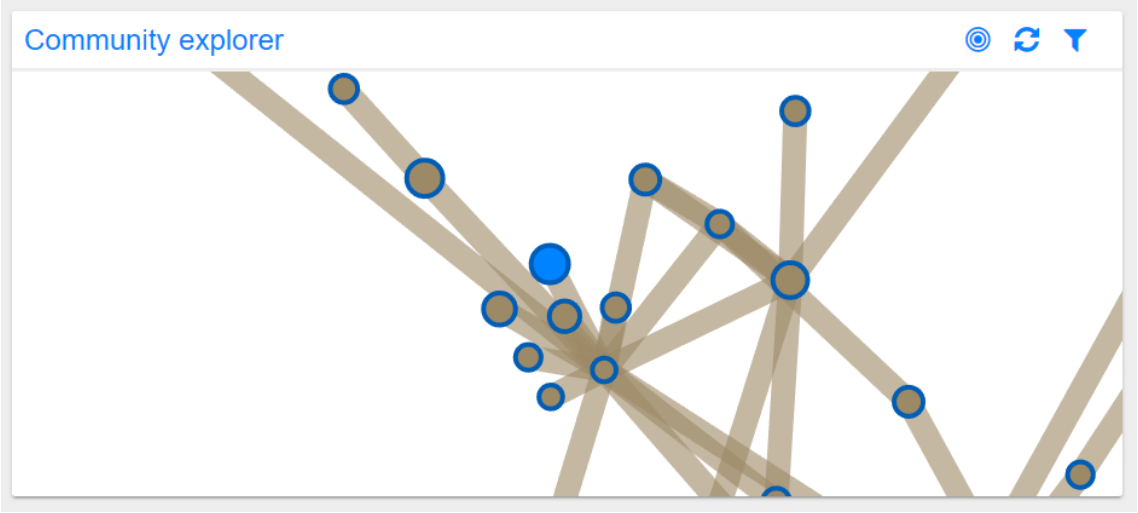


Figura 27: Identificación de uno de los usuarios de la pregunta de investigación 2

Para apreciar las características interactivas y cómo utilizaría un investigador estas visualizaciones para resolver este caso de uso, se puede visualizar el siguiente vídeo `https://jorge-duran.com/research/tfm/videos/privacyTurkey.mp4`.

6.2.3. ¿Cómo influye el género a la hora de hablar sobre la tolerancia y la inmigración?

A la hora de dar respuesta a esta pregunta, lo primero que hay que hacer es elegir como temas a estudiar la tolerancia y la inmigración, como se puede ver en la zona izquierda de la Fig. 28.

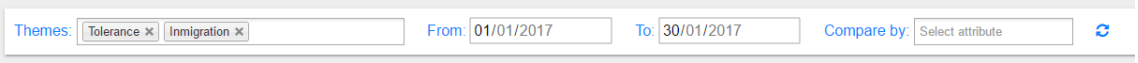


Figura 28: Selección de los temas a estudiar para la pregunta de investigación 3

El siguiente paso es seleccionar el atributo género, como campo por el que se van a dividir los datos, para su comparación. Esta opción se encuentra en el extremo derecho de la Fig. 29.

Figura 29: Selección del atributo por el que comparar para la pregunta de investigación 3

Para finalizar, en el explorador de temas se puede apreciar los distintos comportamientos de los usuarios según su género para ambos temas, Fig. 30. Respecto a la tolerancia, se puede afirmar que este tema, en general, está menos presente en las conversaciones, debido a que es poco usado por el género femenino, el cual se representa con un tono más oscuro. En el caso de la inmigración, se constata que es un tema tratado de manera similar por ambos géneros y que su presencia en las discusiones es constante.

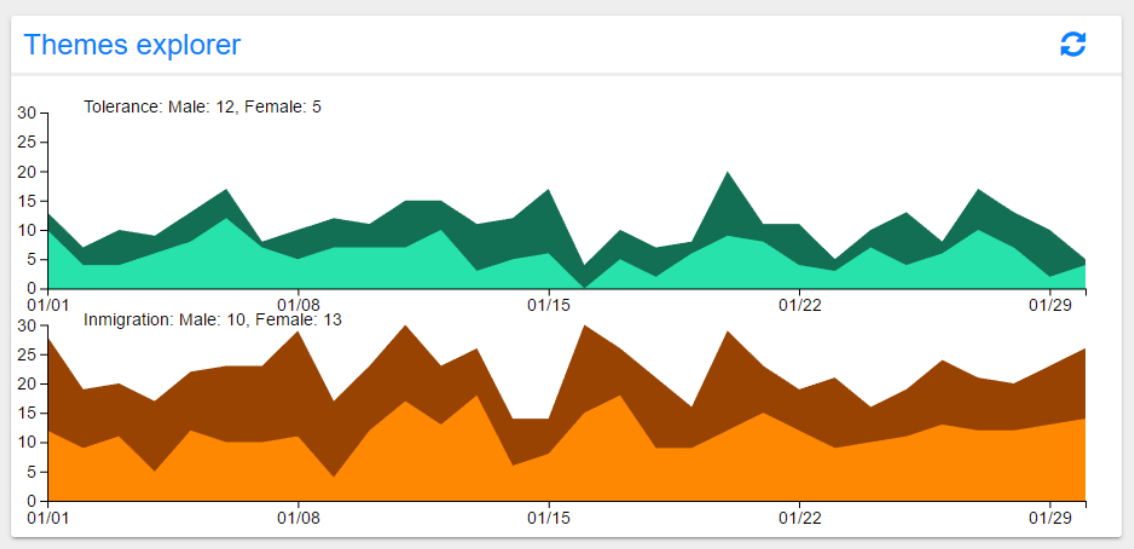


Figura 30: Uso de los temas según el género para la pregunta de investigación 3

Para apreciar las características interactivas y cómo utilizaría un investigador estas visualizaciones para resolver este caso de uso, se puede visualizar el siguiente vídeo `https://jorge-duran.com/research/tfm/videos/gender.mp4`.

6.2.4. ¿Cuál es la evolución temporal de las discusiones sobre acoso, según los países participantes?

Para responder a la última pregunta de investigación, como en los casos anteriores, se debe empezar por la selección del tema objeto de estudio. Esta opción se encuentra en la zona izquierda de la Fig. 31.

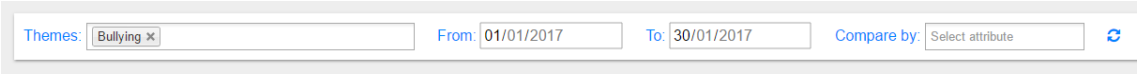


Figura 31: Selección del tema para la pregunta de investigación 4

El siguiente paso, es establecer por qué atributo se va a comparar la evolución del tema objeto de estudio, en este caso, por el país. Esta selección del atributo por el cual se compara está localizada en el extremo izquierdo de la Fig. 32.

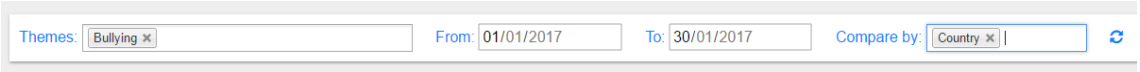


Figura 32: Selección del atributo por el que comparar para la pregunta de investigación 4

En primer lugar, en el mapa podemos ver que los países cuyos usuarios incluyen este tema más frecuentemente en sus comentarios son Italia y Turquía, Fig. 33. Esto es apreciable a simple vista, al estar coloreados ambos países con un tono más oscuro de azul que el resto.

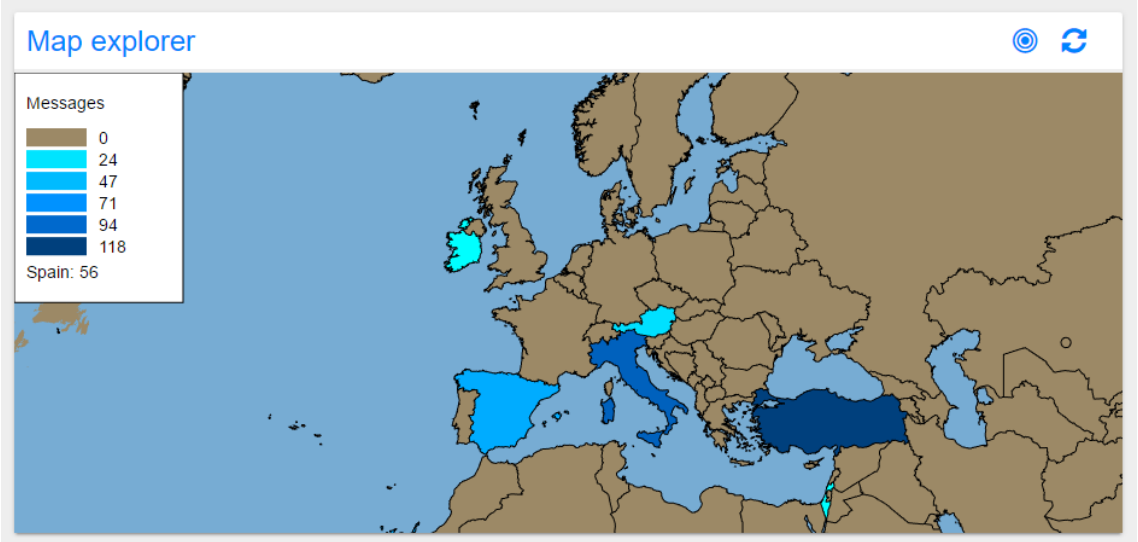


Figura 33: Mapa de uso para la pregunta de investigación 4

A continuación, en el explorador de temas se puede visualizar cómo afecta el país del usuario, a la hora de hablar sobre el acoso, Fig. 34. De la imagen anterior se

pueden extraer dos conclusiones principales: la irregularidad del uso de este tema, por el comportamiento escarpado que presenta el gráfico, y la baja aportación de los usuarios de Israel, Irlanda y Austria, debido a que en muchos instantes su color no está presente, indicando una aportación nula.

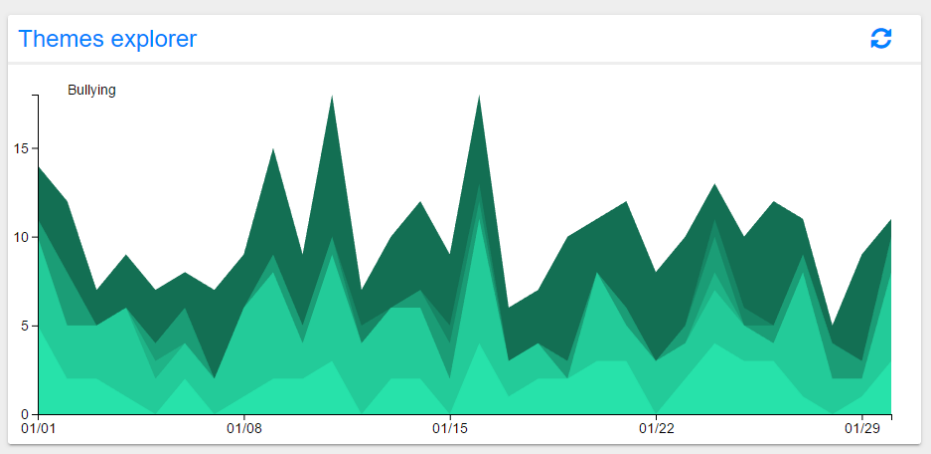


Figura 34: Influencia del país para la pregunta de investigación 4

Finalmente, haciendo clic en uno de los países se puede analizar en mayor profundidad el comportamiento que presentan sus usuarios. Para ello primero hay que utilizar la vista geográfica y seleccionar el país, para después pasar a analizar el uso de ese tema en ese país concreto. En el caso de Turquía, sus usuarios presentan también un comportamiento irregular en el uso del tema (Fig. 35).

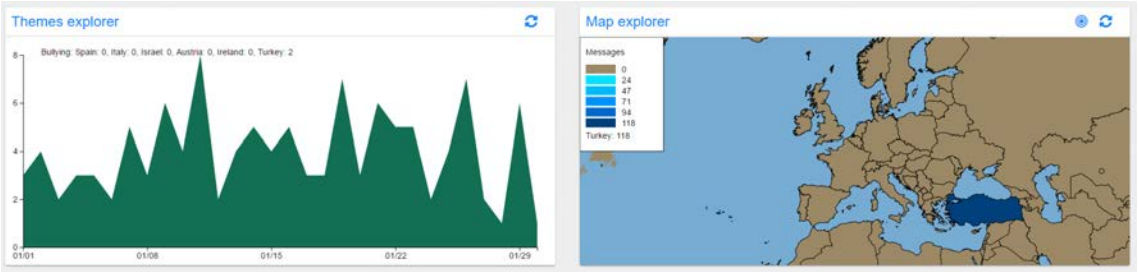


Figura 35: Uso del tema acoso, en Turquía, para la pregunta de investigación 4

Para apreciar las características interactivas y cómo utilizaría un investigador estas visualizaciones para resolver este caso de uso, se puede visualizar el siguiente vídeo `https://jorge-duran.com/research/tfm/videos/themesEvolution.mp4`.

7. Conclusiones y futuras líneas de investigación

En este trabajo se han analizado los principales estudios de otros investigadores sobre comunidades, privacidad y analítica visual, siendo este último aspecto abordado mediante una revisión sistemática de la literatura ya publicada sobre el tema.

Debido a la actual ausencia de datos generados por el proyecto WYRED se ha analizado la generación automática de los mismos, y se ha desarrollado una propuesta para construir un conjunto de prueba lo más similar posible a los datos reales del proyecto. Para ello se han seleccionado los distintos atributos, se han dividido según si son dependientes o no y se les ha asignado valor convenientemente.

En este trabajo también se ha presentado la propuesta de arquitectura para diseñar y desplegar un conjunto de visualizaciones interactivas que permitan explorar los datos del proyecto WYRED. Esta arquitectura modular basada en la arquitectura de micronúcleo, consta de 2 capas básicas: adquisición y anonimización de datos, y de 4 módulos: exploración de los temas principales, representación de las comunidades, visualización de las características de los usuarios y exploración geográfica. Estos a su vez están formados por una o varias capas encargadas del tratamiento y preprocesamiento de los datos, y de la construcción y gestión de la propia visualización. Finalmente, en los resultados se presenta cómo se puede explotar esta propuesta aquí realizada y las distintas formas de utilizar las visualizaciones interactivas que se han creado para responder a algunas de las preguntas de investigación realizadas por los expertos, de manera visual y sin necesidad de utilizar complejas técnicas numéricas. Por ello, se puede afirmar que este trabajo ha cumplido los objetivos planteados en un principio:

- Proponer un sistema que permita explorar los datos del proyecto de manera visual.
- Mantener en todo momento salvaguardada la identidad de los usuarios participantes.
- Prever los posibles problemas que podrían aparecer al tratar los datos del proyecto, de manera anticipada.
- Dar lugar a un modelo flexible que permita que se pueda adaptar con la evolución del proyecto.
- Demostrar cómo se puede extraer conocimiento, mediante visualizaciones interactivas, de los datos generados por el propio proyecto.

Respecto a las futuras líneas de investigación, se considera que hay algunos aspectos en los que se podría seguir trabajando para potenciar y ampliar este trabajo:

- Realizar un estudio de la usabilidad del sistema propuesto. Para ello habría que seleccionar usuarios, que podrían limitarse a 5 según el estudio de Nielsen [127], establecer una serie de pequeñas tareas a realizar, así como un modelo para medir y valorar el desempeño de los mismos al llevarlas a cabo.

- Estudiar e implementar el uso colaborativo de las visualizaciones, para que así distintos investigadores puedan cooperar en la realización de análisis tanto de forma síncrona, como asíncronamente [128].
- Abordar la integración del sistema propuesto con otros sistemas, para favorecer la labor investigadora [128].

A. Apéndice A - El proyecto WYRED

El proyecto WYRED [4] consiste en el desarrollo de un ecosistema tecnológico, para poder conocer en mayor profundidad los intereses y problemas de los jóvenes, la manera que tienen de afrontarlos y, en definitiva, para ser un lugar donde su voz sea escuchada y tomada en cuenta.

Un ecosistemas tecnológico es un conjunto de elementos tecnológicos que permiten cubrir todas las necesidades de un proyecto, para ello es necesario la gestión de los usuarios y la información generada, el soporte para la difusión de estos datos, la integración con otros ecosistemas tecnológicos y la capacidad de que cada uno de estos aspectos pueda evolucionar para adaptarse al proyecto [129]. En el caso del proyecto WYRED, como se puede apreciar en la Fig. 36, este ecosistema está formado por 4 partes bien diferenciadas: un servicio que se encarga de anonimizar a los usuarios, una plataforma privada donde tienen lugar los diálogos con los jóvenes, un sistema para la difusión en redes sociales y una web pública para conocer el proyecto.

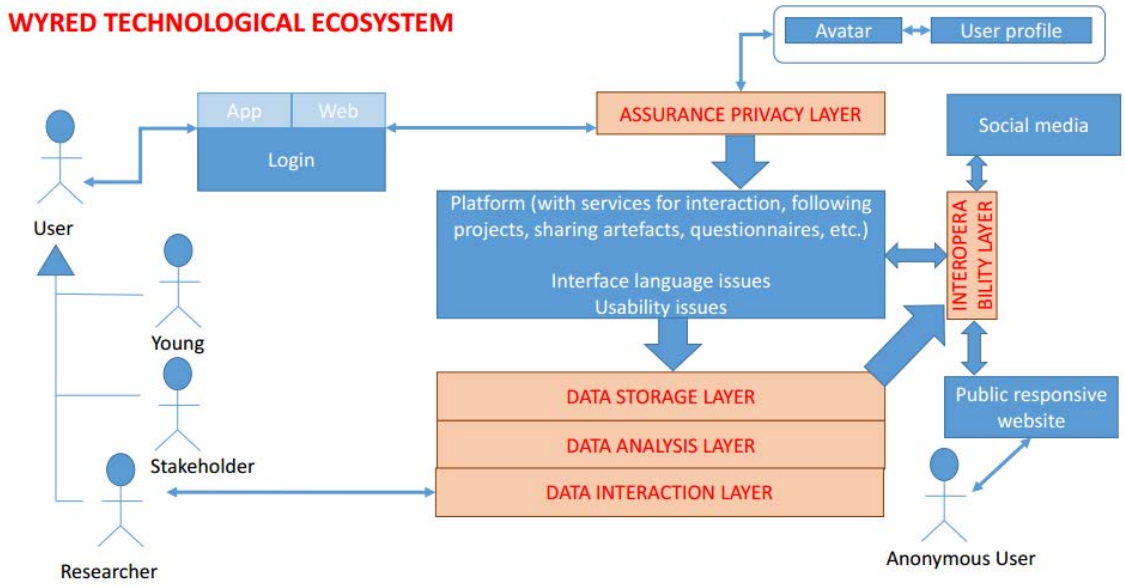


Figura 36: Arquitectura del ecosistema tecnológico del proyecto WYRED

El servicio de registro de los usuarios se encarga de separar los datos privados de los públicos, para anonimizar al usuario y evitar, en la medida de lo posible, su posterior identificación. Para ello, este servicio se implementa en un servidor externo que se encarga de la autentificación y salvaguarda de los datos privados. Como datos públicos, únicamente se utilizan un seudónimo, un avatar, los lenguajes en los que un usuario puede comunicarse y algunos de sus intereses.

La plataforma privada es el lugar donde se ponen en contacto los jóvenes, con distintos expertos que serán los encargados de guiar los diálogos, para obtener in-

formación sobre cómo afectan y qué impresiones tiene la juventud sobre distintos temas de actualidad. Esta comunidad en línea en formato de foro, será analizada para poder extraer conclusiones bajo distintos acercamientos:

- Se analizará la interacción de los usuarios con la plataforma, para ver cuantas veces la visitan, el número de páginas que consultan, el tiempo que permanecen en la misma, etc. tratando de comprobar si esto afecta de alguna manera a sus respuestas y su interés por el proyecto.
- Se medirá el grado de comunicación de los usuarios con los expertos que guíen los diálogos, así como su contacto con otros usuarios del proyecto.
- Se analizarán de manera automática o semiautomática las distintas respuestas obtenidas, para buscar patrones, localizar los temas más importantes, extraer las más relevantes, etc.
- Se comprobará la influencia del contexto sociológico a la hora de opinar sobre distintas cuestiones.

Además de estas cuestiones, esta plataforma contará con un área destinada a ser el punto de reunión de distintos expertos, principalmente del campo de la sociología, para analizar los datos del proyecto. Estos serán apoyados por distintas herramientas tecnológicas como sistemas de colaboración en línea, de compartición de documentos o de análisis de los datos, como es la propuesta que se presenta en este trabajo. Estos expertos podrán ser apoyados por otros que no formen parte del proyecto, ya que los datos generados se podrán consultar de manera pública y abierta por cualquier persona interesada en el tema.

Otro aspecto relevante es la página web del proyecto[9] en la cual se puede conocer el proyecto en profundidad, así como las últimas noticias e investigaciones fruto del trabajo de los miembros de este consorcio. Además, este sitio forma parte de la estrategia de difusión al permitir tanto a investigadores como a miembros de colectivos juveniles unirse al proyecto.

El último servicio relevante es el dedicado a redes sociales, con este servicio se tiene presencia en las redes sociales más relevantes (Facebook, Twitter, YouTube e Instagram) con el fin de publicitar los avances del proyecto, dar a conocer el mismo y captar nuevas personas interesadas en participar de los diálogos sociales.

Debido a la complejidad de este ecosistema tecnológico, hay multitud de requisitos y funcionales a soportar, siendo los más importantes [6]:

- El tratamiento automático de toda la información generada.
- La publicación de la misma, de manera accesible para otros investigadores.
- El mantenimiento de la privacidad de los usuarios en todo momento.

[9] La página web de WYRED es `https://wyredproject.eu/`

- La gestión de la información en multitud de lenguajes.
- La adaptación del proyecto a un variado número de contextos sociales y edades de los usuarios.
- La construcción del ecosistema mediante tecnologías y plataformas *Open Source*.

B. Apéndice B - SLR

B.1. Introducción

La representación gráfica de los datos, con el objetivo de facilitar su compresión, se ha utilizado de manera intensiva desde hace varias décadas. En un primer momento, la visualización se utilizó como una herramienta para transmitir una visión general del conjunto de datos. Sin embargo, es ahora cuando su uso se ha vuelto fundamental, debido a la necesidad de trabajar con conjuntos de datos muy extensos [3].

La visualización de los datos es un campo que se ha venido estudiando durante décadas, con el objetivo de generar nuevos modelos para representar los datos, que se adapten mejor a las restricciones de su contexto de aplicación. De esta manera, han surgido multitud de trabajos centrados en la representación de la información por medio de grafos, líneas, puntos o áreas. Además de lo anterior, a lo largo de este tiempo se han ido desarrollando nuevas ramas de la ciencia como la visualización científica, la ciencia de datos, o la analítica visual, las cuales han demostrado las ventajas de su aplicación para el entendimiento de los propios datos y la resolución de los problemas.

B.1.1. Descripción del problema y motivación

Debido al auge de este campo y la importancia del tratamiento masivo de datos para extraer su conocimiento implícito, algunos autores como Chen, Mao y Liu [130] ya han realizado análisis detallados del estado del arte. En los últimos tiempos, la visualización de los datos, ha demostrado ser una de las herramientas más útiles para analizarlos y comprenderlos, aplicándose de manera satisfactoria en distintas áreas como la medicina [131], la educación [132] o las redes sociales [68].

Gracias a la evolución de la informática, las primitivas visualizaciones propuestas han evolucionado en poderosas visualizaciones interactivas. Para ello ha sido fundamental el desarrollo en un primer momento de Java [32] , y después de las tecnologías web, especialmente del *software* D3 [33]. El estado del arte en este campo, queda recogido en artículos como el de Kucher y Kerren [2] donde ambos autores presentan una taxonomía para categorizar las distintas propuestas de visualizaciones de textos interactivas, la cual se puede consultar en la Fig. 37.

A pesar de existir, como se ha mostrado en este apartado, algunos trabajos que recopilan el estado del arte en el campo de la analítica de datos y de la visualización interactiva, estos son muy generalistas. Lo que implica que no se adaptan bien a un contexto donde se tiene una comunidad [133], cuyo contenido y datos de uso se quieren visualizar, para incrementar el conocimiento que se tiene de la misma.

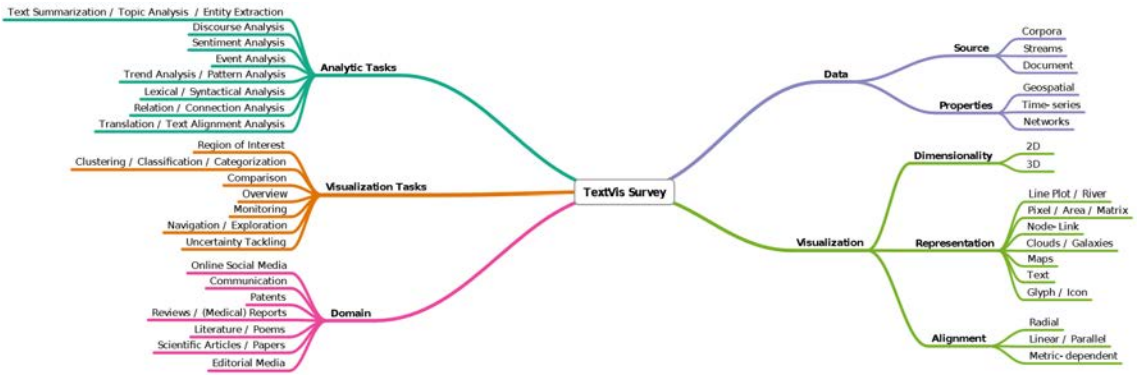


Figura 37: Taxonomía de las propuestas de visualización de textos [2] 2015 IEEE

B.1.2. El enfoque de la investigación

A fin de conseguir una investigación exhaustiva sobre el estado del arte en este campo, se propuso la realización de una revisión sistemática de la literatura. Esto permitirá conocer:

- Cuáles son los autores más prolíficos e importantes en este sector.
- Qué evolución ha tenido el mismo en los últimos años.
- Dónde se publican los avances en este campo.
- Cuáles son las propuestas de visualización más utilizadas y qué problemas resuelven.

Para la realización de esta revisión sistemática de la literatura, se ha partido de 189 artículos publicados en revistas y congresos, los cuales mediante sucesivos cribados que se expondrán más adelante en este documento, han dado lugar a un corpus documental formado por 9 artículos.

B.1.3. Organización

Este documento está organizado de la siguiente manera: la Sección B.2 presenta cómo se ha realizado el proceso de revisión, la Sección B.3 se centra en el *mapping* de la literatura, la Sección B.4 aborda la revisión de los artículos seleccionados y la Sección B.5 resume las principales conclusiones de este trabajo.

B.2. El proceso de revisión

Para realizar la revisión sistemática de la literatura, se ha tomado como referencia la guía propuesta por Kitchenham [134], la cual define una estructura de 3 etapas: planteamiento, realización y presentación. Para ello se realizan las siguientes fases:

1. Establecer las preguntas de investigación que se quieren contestar con este sistemático.
2. Describir la investigación mediante el método PICOC.
3. Definir los criterios de inclusión y exclusión que deben cumplir los distintos trabajos para ser tomados en cuenta.
4. Seleccionar las fuentes bibliográficas que van a ser consultadas.
5. Concretar las consultas que se van a realizar en las distintas fuentes bibliográficas.
6. Definir el protocolo de aseguramiento de la calidad.
7. Realizar un análisis detallado de los trabajos que hayan superado todas las fases anteriores.

Además de la revisión, se ha aprovechado este proceso para complementarlo con un *mapping* de la literatura escrita sobre este campo, siguiendo el modelo usado por Weidt Neiva [135].

B.2.1. Preguntas de investigación

El primer paso para la realización del sistemático de literatura y su *mapping*, es el planteamiento de las preguntas que se quieren resolver con esta investigación. A continuación se plantean dichas preguntas[10]:

MQ1: ¿Quiénes son los autores más importantes?

MQ2: ¿Cuál ha sido la evolución de este campo?

MQ3: ¿En qué medios están publicados los documentos más relevantes de este campo?

MQ4: ¿Cuáles son las principales técnicas de visualización utilizadas?

RQ1: ¿Qué soluciones ofrece la visualización interactiva de datos para explorar el contenido y los datos de uso de una plataforma?

B.2.2. PICOC

El método PICOC [136] (*Population, Intervention, Comparation, Outcome* y *Context*), permite describir los 5 elementos de una pregunta de investigación, que el caso de este trabajo son:

[10]Bajo las siglas MQ se agrupan las preguntas relativas al *mapping* y bajo RQ, las preguntas de investigación

- *Population:* Soluciones de visualización interactiva.
- *Intervention:* Visualización de comunidades.
- *Comparation:* Visualización estática.
- *Outcome:* Soluciones de visualización interactiva de comunidades.
- *Context* Soluciones implementables computacionalmente.

B.2.3. Criterios de inclusión y exclusión

Una vez que se tiene claro qué se quiere buscar, el siguiente paso que se ha dado ha sido definir los criterios de inclusión (IC) y exclusión (EC). Estos criterios limitan el tipo de documentos que van a ser analizados, atendiendo a su idioma, formato u otros aspectos. Así, los criterios de inclusión son los siguientes:

IC1: Documentos en inglés *AND*

IC2: Publicados en revistas *OR* congresos *AND*

IC3: Sobre visualización interactiva de datos *AND*

IC4: Que trabajen con estadísticas de uso *OR* datos del contenido *AND*

IC5: Centrados en comunidades *online*

Los criterios de exclusión son los siguientes:

EC1: Documentos escritos en un idioma distinto al inglés *AND*

EC2: Artículos no revisados por pares *OR* publicados en otros formatos *AND*

EC3: Con una propuesta que no utilice la visualización interactiva *AND*

EC4: Artículos que no tratan con datos del uso *OR* no utilizan datos del propio contenido *AND*

EC5: Artículos que no se centran en la información generada por una comunidad online

B.2.4. Fuentes bibliográficas

El siguiente paso en la realización del sistemático de literatura, ha sido elegir las fuentes de referencia que se van a usar para buscar los artículos. Las tres fuentes principales son Scopus, Web of Science y Google Scholar. Todas ellas ofrecen unos resultados similares, al tener un nivel de cobertura amplio de las publicaciones realizadas. Por ello, se ha escogido Scopus como fuente bibliográfica para la realización de este estudio, al cubrir mejor esta rama del conocimiento [137].

B.2.5. Consulta de búsqueda

Una vez definidas las preguntas que se quieren resolver y las fuentes bibliográficas a consultar, el siguiente paso dado es definir de manera precisa la consulta de búsqueda que se va a utilizar. Para la construcción de la misma, se han utilizado las palabras claves extraídas de la pregunta de investigación, combinadas con los operadores *AND* y *OR*. Además, se ha tenido presente la utilización de los términos: visualización interactiva y análisis interactivo como sinónimos. Debido a lo anterior, la consulta efectuada en Scopus es la siguiente:

TITLE-ABS-KEY(("interactive visualization" OR "interactive analysis") AND "community")

El resultado de ejecutar la consulta anterior son 189 trabajos sobre comunidades y métodos interactivos de visualización o análisis. Los cuales van a ser analizados en los siguientes apartados.

B.2.6. Comprobación de la calidad

Una de las partes fundamentales para la realización de una revisión de la literatura correcta, es el establecimiento de un sistema de comprobación de la calidad [134]. Para realizar esta valoración, se ha decidido desarrollar la siguiente batería de preguntas, de acuerdo con la pregunta de investigación:

QQ1: ¿Los artículos utilizan datos del uso (estadísticas) de una plataforma o comunidad?

QQ2: ¿Los artículos utilizan datos extraídos del contenido de una plataforma o comunidad?

QQ3: ¿Los artículos justifican la decisión de utilizar una visualización interactiva?

QQ4: ¿Los artículos presentan una visualización interactiva satisfactoria?

QQ5: ¿Los artículos explican de manera clara cómo se construye la visualización interactiva?

QQ6: ¿Los artículos contienen referencias a la realización de un estudio con usuarios?

Estas pueden ser respondidas de manera afirmativa, negativa o parcialmente, teniendo como valor asociado, 1, 0 y 0,5 puntos, respectivamente. Este mecanismo permite aportar mayor flexibilidad, si se compara con el uso de respuestas binarias. Como punto de corte para aceptar un documento, se ha establecido el primer tercil (4 puntos), quedando todos los artículos con un valor inferior a 4 excluidos.

B.2.7. La revisión

Una vez ejecutada la consulta, se ha procedido a agregar a EndNote[11] los 189 resultados obtenidos. Después, se ha diseñado un proceso de revisión formado por los siguientes pasos:

1. Se ha procedido a comprobar que no hubiera ningún documento duplicado con la herramienta EndNote.
2. Se ha leído el título y el resumen de cada artículo para, junto con los criterios de inclusión y exclusión, realizar el primer filtrado. En este paso, sólo se han eliminado aquellos documentos donde se tenía la certeza de que deberían ser filtrados.
3. Se ha accedido a los documentos completos y se han leído enteros, para valorar su aportación a la pregunta de investigación. En este paso se han eliminado todos los que no agregaban información al respecto o no cumplían los criterios de inclusión.
4. Se han añadido algunos documentos al consultar la bibliografía de los artículos seleccionados en el paso anterior.
5. Se ha procedido a analizar la calidad de cada uno de los documentos seleccionados, eliminando todos aquellos que no han pasado el corte.

Siguiendo estos pasos, se han obtenido los resultados recogidos en la Fig. 38. En el paso 1, no se ha detectado ningún documento duplicado así que se mantuvieron los 189 documentos. En el paso 2, en el que se aplica el filtrado por título y resumen, sólo conservo 39 documentos (20.63 %). En el paso 3, al aplicar el filtrado por contenido, únicamente se conservaron 13 documentos (6.88 %). En el paso 4, se incrementa en 2 el número de documentos, teniendo un total de 15 (7.85 %), al revisar la bibliografía de los documentos anteriormente aceptados. Finalmente, en el paso 5 después de filtrarlos, según los criterios de calidad, quedaron finalmente 9 documentos (4.71 %). Como se puede apreciar, el número final de documentos aceptados es muy bajo respecto del total, esto es debido a multitud de razones: artículos genéricos, documentos no relacionados con el tema, que no explican con calidad los métodos utilizados, etc.

En la siguiente dirección web (`https://goo.gl/gUCgwc`), se encuentra recogido cada uno de los pasos dados, así como los documentos que forman parte de los mismos.

B.3. El *mapping* sistemático de la literatura

El *mapping* sistemático de la literatura, toma como fuente los 9 artículos analizados anteriormente. Con ellos se va a dar respuesta a las tres preguntas de *mapping* planteadas.

[11] Un *software* para la gestión de referencias, http://endnote.com/

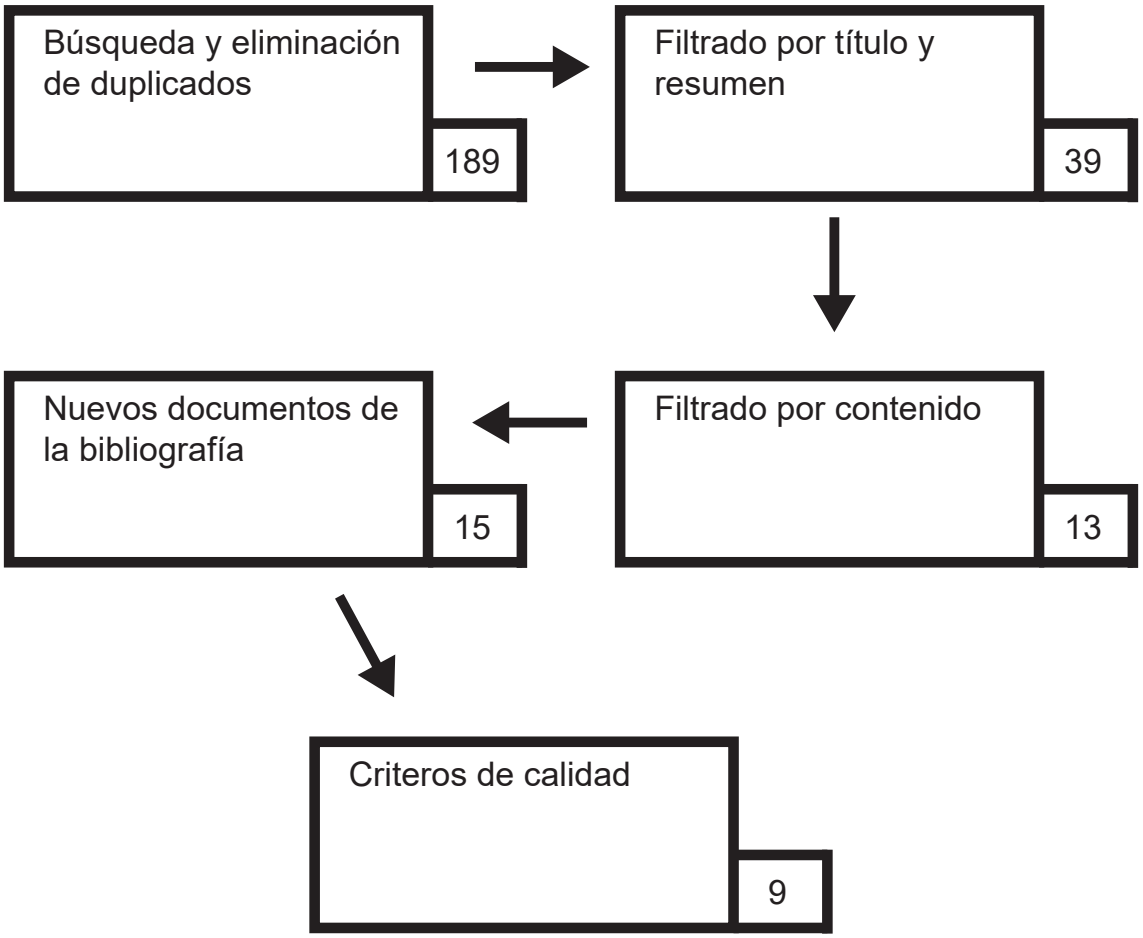


Figura 38: Representación del proceso de revisión de la literatura

En la Tabla 9, se puede encontrar la respuesta a la primera pregunta del *mapping* (*¿Quiénes son los autores más importantes?*). Al partir de un número reducido de artículos, no hay muchos autores que destaquen con más de una publicación en este campo.

Nombre	**Número de documentos**
Dou, W.; Ribarsky, W	2
Abdelsadek, Y.; Ahmed, N. K.; Berger-Wolf, T.; Chang, R.; Chelghoum, K.; Conglei, S.; Guo, D.; Havre, S.; Herrmann, F.; Hetzler, E.; Huamin, Q.; Johnson, A.; Kacem, I.; Leigh, J.; Melançon, G.; Nowell, L.; Otjacques, B.; Pich, C.; Qing, C.; Reda, K.; Rossi, R. A.; Sallaberry, A.; Siwei, F.; Tantipathananandh, C.; Wang, X.; Whitney, P.; Xiaoyu Wand, D.; Zaidi, F.	1

Tabla 9: Autores y su número de publicaciones

Respecto a la pregunta 2 (*¿Cuál ha sido la evolución temporal de este campo?*), la misma queda esbozada en la Fig. 39. En lineas generales, se puede decir que es un ámbito bastante reciente, ya que la mayoría de las publicaciones son posteriores a 2008. Además, tal y como atestigua la gráfica, es una temática que todavía mantiene el interés de los investigadores.

En referencia a la pregunta 3 (*¿En qué medios están publicados los documentos más relevantes de este campo?*), la Fig. 40 contiene la respuesta. Como se puede apreciar, los documentos que se han seleccionado provienen principalmente de revistas y conferencias.

Respecto a la pregunta 4 (*¿Cuáles son las principales técnicas de visualización utilizadas?*), hay 3 tipos de visualizaciones que destacan frente al resto: los gráficos de área, los grafos y las coordenadas paralelas [70], como se puede ver en la Fig. 41. Es importante destacar, que hay un mayor número de visualizaciones que documentos examinados, esto es debido a que en algunos artículos se plantea el uso de varias visualizaciones. Cada una de ellas se utiliza con un fin:

- Los gráficos de área son utilizados para mostrar la evolución temporal de un tema o un evento.
- Los grafos son usados para visualizar la relación entre los distintos usuarios, lo que da lugar a conocer las comunidades qué forman y el modo de interacción.
- Las coordenadas paralelas se utilizan para representar un grupo amplio de características de una persona o suceso.

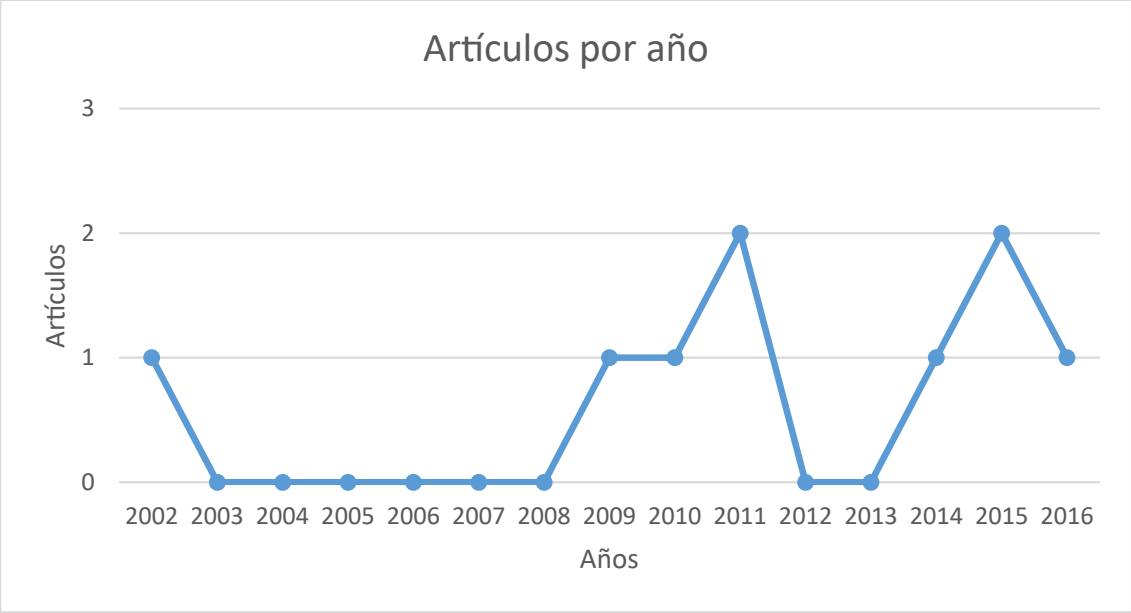


Figura 39: Evolución temporal del tema de investigación

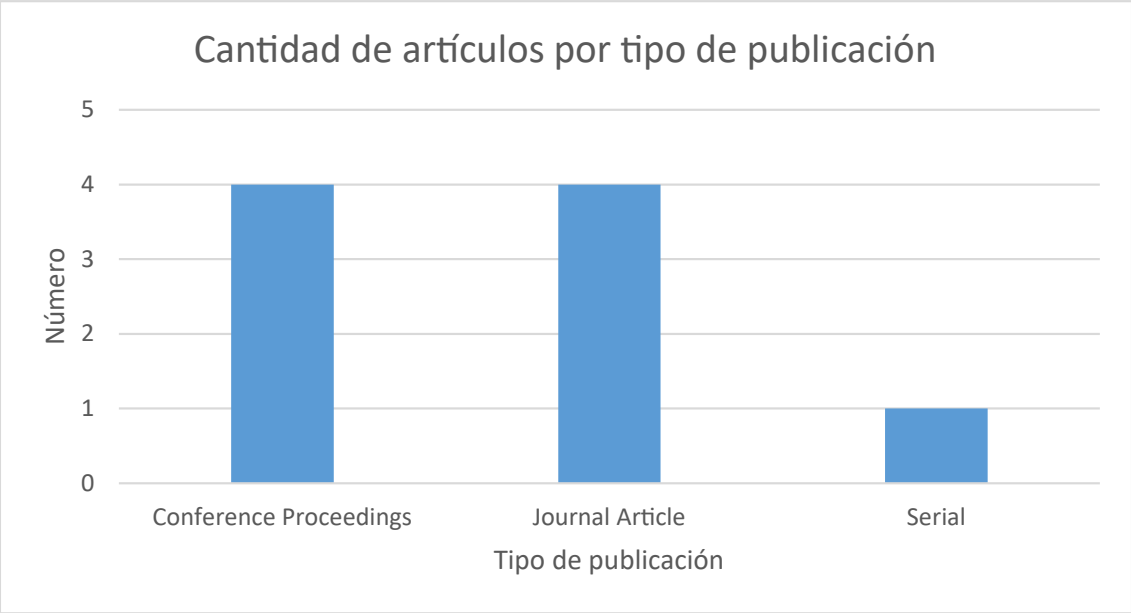


Figura 40: Los medios de publicación en relación al número de publicaciones

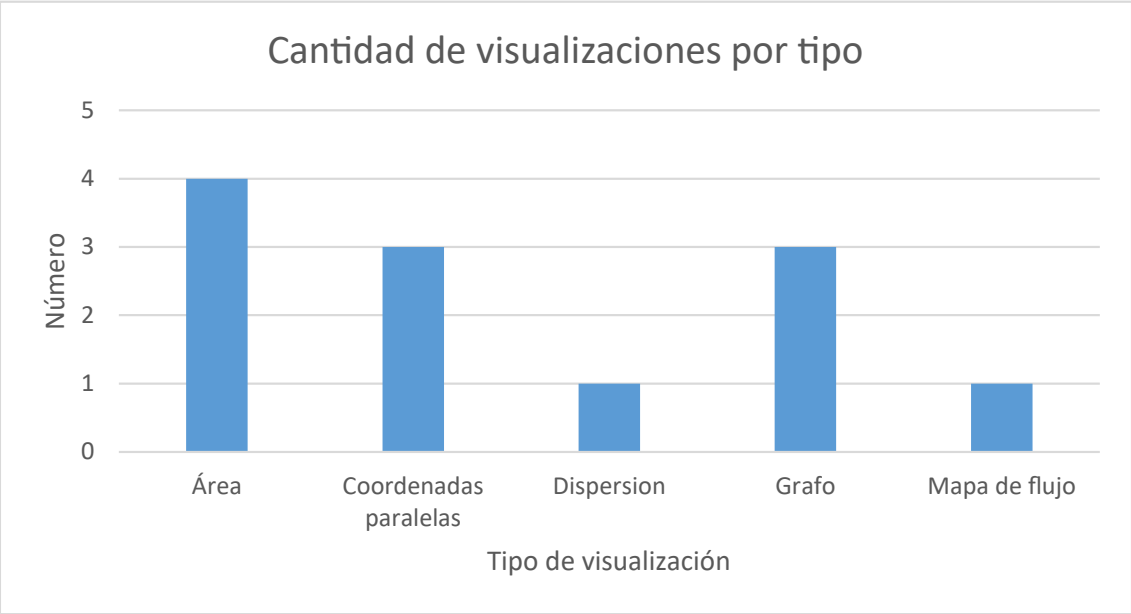


Figura 41: Los tipos de visualización más usados

B.4. La revisión de la literatura

El estudio está centrado en visualizaciones interactivas, es decir, aquellas cuya representación cambia con la interacción del usuario (gestos, movimientos, selecciones, etc.), para adaptarse al mismo y poder transmitir mejor la información. Además, es necesario tener en cuenta que no se limita sólo a esto último, sino que también incluye conceptos como colaboración, integración, descubrimiento del conocimiento,

etc. [128].

Los documentos seleccionados, al haber superado un proceso de comprobación de la calidad, nos aseguran que plantean propuestas de visualizaciones interactivas satisfactorias. Esto es una cuestión muy importante, ya que tal y como demuestran Kandel et al. en su estudio *Research directions in data wrangling: Visualizations and transformations for usable and credible data* [34], tratar con conjuntos de datos reales genera multitud de problemas.

Para conseguir responder a la pregunta de investigación (*¿Qué soluciones ofrece la visualización interactiva de datos para explorar el contenido y los datos de uso de una plataforma?*), los documentos han sido analizados atendiendo a 4 factores:

- La utilización de estadísticas del uso de la plataforma.
- El uso del contenido textual.
- El análisis de la componente temporal de los datos.
- La detección de grupos de usuarios que forman comunidades.

La relación entre estos factores y los artículos revisados, se puede consultar en la Tabla 10.

Documento	**Estadísticas de uso**	**Textos**	**Análisis temporal**	**Detección de comunidades**
[69]		X	X	
[123]	X			
[67]		X		X
[71]		X	X	
[138]	X		X	X
[114]		X	X	
[139]	X			X
[140]	X		X	X
[66]	X			X

Tabla 10: Artículos revisados y su relación con los 4 factores de revisión

B.4.1. Estadísticas de uso

De los 9 documentos analizados, 5 abordan el problema de utilizar como datos, las estadísticas de uso de una plataforma. Sin embargo, los autores presentan distintas soluciones según el problema a resolver.

La gran mayoría de los autores, utiliza estos datos para representar los usuarios y las comunidades que estos forman. Para ello suelen utilizar los grafos [139] [66]. Sin embargo, otros autores prefieren usar los gráficos de líneas [138] o los gráficos

de áreas agrupadas [140].

Otro problema afrontado con estos datos, es la interacción espacial de los usuarios. Aquí los autores proponen el uso de un mapa de flujo [123], ya que la presencia del mapa, ayuda a reducir la carga cognitiva.

B.4.2. Corpus textuales

4 de los 9 documentos analizados, utilizan como fuente de datos conjuntos de textos. Estos textos provienen de multitud de fuentes, siendo la más frecuente, las redes sociales. Además, es muy común que los artículos cuyo conjunto de datos está formado por textos, también aborden el análisis temporal de los mismos. Para la extracción de los principales temas, varios de los artículos sugieren usar la técnica LDA (*Latent Dirichlet Allocation*) [109], la cual ha sido usada de forma satisfactoria en multitud de investigaciones.

El modelo de visualización más usado en los mismos, es el gráfico de áreas apiladas [69] [71]. El cual permite de manera sencilla conocer cual es el tema más popular en un momento dado, sin embargo, otros autores [114] prefieren no apilar los gráficos y que cada tema tenga su propia visualización. Lo que permite poder apreciar de manera más clara la evolución de cada tema.

En contraposición al modelo anterior, hay autores [67] que prefieren utilizar los corpus documentales para buscar grupos entre sí. Por ello, para la representación de los mismos utilizan como modelo de visualización, los grafos.

B.4.3. Análisis temporal

Más de la mitad de los artículos analizados (5 de 9) contemplan alguna visualización que tenga en cuenta la evolución de los datos a través del tiempo. Lo cual denota que hoy en día, con el volumen de datos que se genera, es necesario trabajar con representaciones dinámicas que no solo muestren los datos actuales, sino que sean capaces de representar los anteriores, para poder realizar comparaciones.

Respecto a los modelos de visualización principales, el más destacado es el gráfico de áreas asociado a datos textuales, como se ha comentado en el punto anterior. El componente temporal en estos casos, no es un elemento estático, sino que es dinámico, al permitir al usuario de manera interactiva modificar la ventana temporal que se representa.

Otros autores se desmarcan de este modelo de visualización, optando por utilizar gráficos de líneas [138] [140], de tal manera que aprovechan uno de los ejes para representar el tiempo.

B.4.4. Detección de comunidades

La detección de comunidades, mediante algoritmos y/o visualizaciones, es un tema que está muy presente en los documentos analizados (5 de 9). En la mayoría de los casos, está muy relacionado con aquellos documentos que utilizan estadísticas de uso como fuente de datos, aunque otras veces lo que se busca es agrupar conceptos [67].

El tipo de visualización elegida por la mayoría de los documentos son los grafos. Partiendo de este modelo, algunos autores añaden la posibilidad de mover cada uno de los nodos, descubrir nueva información al hacer clic en ellos o la posibilidad de poder modificar la manera en que se agrupan los mismos [139].

Otros autores [138] utilizan un enfoque distinto para descubrir comunidades. El mismo, se basa en representar mediante gráficos de líneas a cada usuario, de tal manera que automáticamente se generan zonas compactas, que agrupan a usuarios que han tomado la misma decisión en un instante concreto de tiempo.

B.5. Conclusiones

En este anexo se presenta una revisión sistemática de la literatura y un mapeo de la misma, para identificar, clasificar y analizar las visualizaciones interactivas propuestas para representar las estadísticas de uso y el contenido de una comunidad. Por ello, se ha pasado por una serie de fases, donde los documentos se han ido filtrando para escoger finalmente aquellos de mayor calidad y más relevantes de acuerdo con la pregunta de investigación. Esto ha dado lugar, a que sólo fueran seleccionados 9 artículos de los 189 recuperados en un primer momento.

Los resultados obtenidos permiten afirmar que este campo es bastante moderno, ya que los artículos analizados son posteriores al año 2000 y que además, sigue en desarrollo, al tener varios artículos muy recientes. Respecto a la manera de presentar las investigaciones, los autores se decantan de manera proporcional por artículos de revistas y congresos.

Sobre el origen de los datos, la mayoría de los artículos se centran en trabajar con estadísticas de uso o con el contenido generado (corpora documentales), sin embargo, algún autor si propone un conjunto integral de visualizaciones donde trabaja con ambos al mismo tiempo.

Las visualizaciones interactivas más utilizadas son los grafos, cuando se quiere analizar las estadísticas de uso con el fin de mostrar o agrupar los usuarios en comunidades, y los gráficos de área, para mostrar la evolución temporal de los temas más frecuentes en el contenido. Además de estas dos representaciones, también son utilizadas de forma recurrente, las coordenadas paralelas, a fin de representar las múltiples características de un individuo o tema.

Finalmente, es importante destacar que varios de los artículos defienden el uso de visualizaciones interactivas, como un buen sistema para representar y descubrir

el conocimiento implícito en los datos. Quedando esto también demostrado en los estudios realizados con usuarios, presentes en algunos documentos analizados. En los mismos, se llega a la conclusión de que se han desarrollado visualizaciones interactivas satisfactorias.

C. Apéndice C - Tecnologías Web para la generación de gráficos interactivos

Las tecnologías y lenguajes de programación Web se han convertido en la mejor manera para representar la información, debido a que están ampliamente soportados en todos los sistemas operativos y dispositivos. Además, estos han sufrido una rápida evolución con la que han conseguido afrontar el tratamiento y manipulación de amplios volúmenes de datos y de multitud de elementos visuales de manera efectiva. A continuación, se describen dos de las soluciones tecnológicas que mayor peso han tenido en la creación de gráficos interactivos para la Web.

C.1. D3.js

D3.js [33] es una biblioteca de código abierto escrita en *JavaScript*, que permite crear visualizaciones interactivas utilizando únicamente las tecnologías Web (HTML, CSS, *JavaScript*). Mike Bostock, su creador, utilizó las tecnologías anteriores con el fin de evitar los sistemas propietarios para la representación visual de los datos, como *Flash*. Además, llamó a su biblioteca D3 (*Data Driven Documents*), porque la creó con el fin de que sirviera de base para el desarrollo de documentos dirigidos por los datos [141]. La diferencia entre estos últimos, y los documentos textuales, es que en los primeros, la importancia está en los propios datos y sus características.

D3.js tiene multitud de características, ya que es un desarrollo que hoy en día sigue evolucionando[12], pero las principales son las siguientes [142]:

- Un sistema eficiente para seleccionar elementos del DOM de HTML.
- Un mecanismo para enlazar los datos con los elementos visuales.
- Soporte para la gestión de la creación y eliminación de elementos de datos.
- Capacidad de aplicar estilos de manera dinámica a los elementos del DOM.
- Un mecanismo para definir un modelo de interacción entre el usuario y los datos.
- Un sistema para definir transiciones, basado en el cambio dinámico de los datos.

El flujo de trabajo que debe seguir todo desarrollo realizado utilizando esta herramienta es el siguiente:

1. Diseñar una visualización y su modelo de interacción.
2. Generar un conjunto de datos para ser visualizado.

[12]El desarrollo es accesible mediante la siguiente dirección: https://d3js.org/

3. Hacer accesible esos datos, mediante un fichero, una API, etc.

4. Procesar esos datos, según la visualización elegida.

5. Desarrollar la visualización mediante código en *JavaScript*.

6. Enlazar el conjunto de datos, para que sea visualizado.

7. Añadir los gestores de cada una de las interacciones soportadas y su comportamiento.

8. Optimizar la visualización, si procede.

Como se puede apreciar, D3.js no es una biblioteca pensada para generar un conjunto de gráficos sencillos, como *Google Charts* [143], sino una compleja, pero poderosa herramienta capaz de generar cualquier tipo de representación. Debido a la complejidad de la misma, es muy común reutilizar parte o la totalidad de las visualizaciones realizadas por otros usuarios, que se pueden encontrar en algunas páginas Web [144]. Otra recomendación asociada al uso de esta biblioteca, es utilizar el navegador Chrome debido a que su motor de renderizado soporta de manera más eficiente, las interacciones generadas por esta biblioteca.

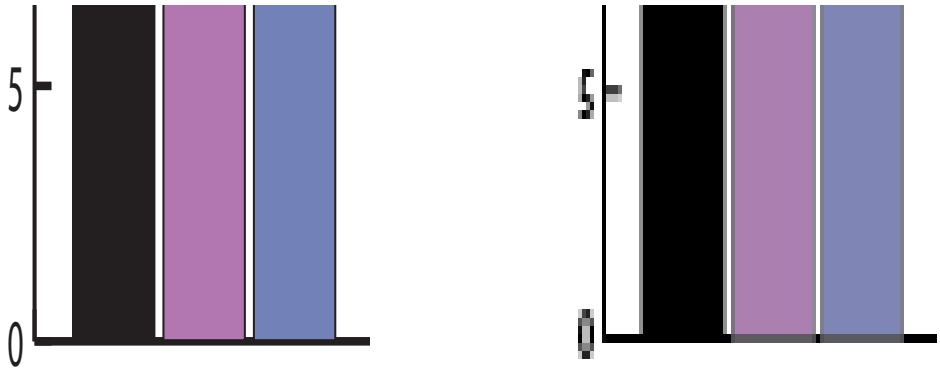
C.2. SVG

El formato SVG, acrónimo de *Scalable Vector Graphics*, es un estándar creado por el W3C[13] en el cual se define una especificación para, mediante XML, describir vectores bidimensionales y la integración de contenido vectorial y rasterizado [145].

El uso de este formato se ha popularizado mucho en la Web debido a que es una tecnología abierta diseñada para trabajar con otros estándares del W3C, como CSS, DOM o HTML. Si bien, el soporte de todas sus características todavía no es completo en todos los navegadores [146].

Dentro de las ventajas que tiene este formato, podemos destacar 2, la facilidad de ser tratados, y por ende la posibilidad de aplicarle estilos, como un componente HTML y la capacidad de no perder calidad al ser redimensionado, como se puede apreciar en la Fig. 42.

[13]W3C (*World Wide Web Consortium*) es una institución encargada de fijar los principales estándares de la web https://www.w3.org/



(a) Fichero SVG (b) Fichero PNG

Figura 42: Comparación de una misma visualización en dos formatos distintos

Por estas características, se puede llegar a la conclusión de que la elección de usar el formato SVG en D3.js no fue arbitraria.

Referencias

[1] A. Parker and H. Beach. Theme river. [Online]. Available: http://www.cs.middlebury.edu/~candrews/showcase/infovis_techniques_s16/themeriver/themeriver.html [*Citado en págs.* III *y 31.*]

[2] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 2015, Conference Proceedings, pp. 117–121. [*Citado en págs.* IV, *30, 59 y 60.*]

[3] Micro Focus. How much data is created on the internet each day? [Online]. Available: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/ [*Citado en págs. 1 y 59.*]

[4] F. J. García-Peñalvo and N. A. Kearney, "Networked youth research for empowerment in digital society. the wyred project," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'16)*. ACM, 2016, Conference Proceedings, pp. 3–9. [*Citado en págs. 1 y 55.*]

[5] F. J. García-Peñalvo, "The wyred project: A technological platform for a generative research and dialogue about youth perspectives and interests in digital society." *Journal of Information Technology Research*, 2016. [*Citado en pág. 1.*]

[6] WYRED, "Requirements document wp3_d3.1," 2017. [Online]. Available: https://doi.org/10.5281/zenodo.292978 [*Citado en págs. 2 y 56.*]

[7] J. Cruz-Benito. Systematic literature review & mapping. [Online]. Available: http://repositorio.grial.eu/handle/grial/685 [*Citado en pág. 7.*]

[8] A. Booth, A. Sutton, and D. Papaioannou, *Systematic approaches to a successful literature review*. Sage, 2016. [*Citado en pág. 7.*]

[9] C. Maher, K. Baessler, C. M. Glazener, E. J. Adams, and S. Hagen, "Surgical management of pelvic organ prolapse in women: a short version cochrane review," *Neurourology and urodynamics*, vol. 27, no. 1, pp. 3–12, 2008. [*Citado en pág. 7.*]

[10] O. Olsen and P. C. Gøtzsche, "Cochrane review on screening for breast cancer with mammography," *The Lancet*, vol. 358, no. 9290, pp. 1340–1342, 2001. [*Citado en pág. 7.*]

[11] L. Codina. Revisiones sistematizadas y cómo llevarlas a cabo con garantías: systematic reviews y salsa framework. [Online]. Available: https://www.lluiscodina.com/revision-sistematica-salsa-framework/ [*Citado en pág. 7.*]

[12] F. García-Peñalvo. Taller de revisión sistemática de literatura. [Online]. Available: https://repositorio.grial.eu/handle/grial/771 [*Citado en pág. 7.*]

[13] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online social networks," *Journal of Computer-Mediated Communication*, vol. 3, no. 1, 1997. [*Citado en pág. 9.*]

[14] C. Licoppe and Z. Smoreda, "Are social networks technologically embedded?: How networks are changing today with changes in communication technology," *Social networks*, vol. 27, no. 4, pp. 317–335, 2005. [*Citado en pág. 9.*]

[15] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8. [*Citado en págs. 9 y 33.*]

[16] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978. [*Citado en pág. 10.*]

[17] R. Kumar, J. Novak, and A. Tomkins, *Structure and Evolution of Online Social Networks*. New York, NY: Springer New York, 2010, pp. 337–357. [*Citado en pág. 10.*]

[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, Conference Proceedings, pp. 29–42. [*Citado en pág. 10.*]

[19] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *First Monday*, 2008. [*Citado en pág. 11.*]

[20] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002. [*Citado en págs. 11 y 33.*]

[21] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010. [*Citado en pág. 11.*]

[22] Alexa. Top sites in spain. [Online]. Available: http://www.alexa.com/topsites/countries/ES [*Citado en pág. 11.*]

[23] R. M. Marra, J. L. Moore, and A. K. Klimczak, "Content analysis of online discussion forums: A comparative analysis of protocols," *Educational Technology Research and Development*, vol. 52, no. 2, p. 23, 2004. [*Citado en pág. 12.*]

[24] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, 2010. [*Citado en pág. 12.*]

[25] M. Mazzolini and S. Maddison, "When to jump in: The role of the instructor in online discussion forums," *Computers & Education*, vol. 49, no. 2, pp. 193–213, 2007. [*Citado en pág. 12.*]

[26] G. Eysenbach and J. E. Till, "Ethical issues in qualitative research on internet communities," *BMJ*, vol. 323, no. 7321, pp. 1103–1105, 2001. [*Citado en pág. 13.*]

[27] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005, Conference Proceedings, pp. 71–80. [*Citado en pág. 13.*]

[28] A. Lenhart and M. Madden, "Teens, privacy and online social networks," Pew Research Center, Report, 2007. [*Citado en pág. 14.*]

[29] H. Chen, C. E. Beaudoin, and T. Hong, "Teen online information disclosure: Empirical testing of a protection motivation and social capital model," *Journal of the Association for Information Science and Technology*, vol. 67, no. 12, pp. 2871–2881, 2016. [*Citado en pág. 14.*]

[30] E. Auschitzky, M. Hammer, and A. Rajagopaul. How big data can improve manufacturing. [Online]. Available: http://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing [*Citado en pág. 14.*]

[31] E. D. Keim, J. Kohlhammer, and G. Ellis, "Mastering the information age: Solving problems with visual analytics, eurographics association," 2010. [*Citado en págs. 14 y 15.*]

[32] Oracle. Conozca más sobre la tecnología java. [Online]. Available: https://www.java.com/es/about/ [*Citado en págs. 15 y 59.*]

[33] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011. [*Citado en págs. 15, 41, 59 y 73.*]

[34] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. Van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Information Visualization*, vol. 10, no. 4, pp. 271–288, 2011. [*Citado en págs. 15 y 68.*]

[35] J. Thomas and K. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005. [*Citado en pág. 15.*]

[36] T. Bernardin, E. Cowgill, O. Kreylos, C. Bowles, P. Gold, B. Hamann, and L. Kellogg, "Crusta: A new virtual globe for real-time visualization of sub-meter digital topography at planetary scales," *Computers and Geosciences*, vol. 37, no. 1, pp. 75–85, 2011. [*Citado en pág. 16.*]

[37] F. Poulet, C. Quantin-Nataf, H. Ballans, K. Dassas, J. Audouard, J. Carter, B. Gondet, L. Lozac'h, J. C. Malapert, C. Marmo, L. Riu, and A. Séjourné,

"Psup: A planetary surface portal," *Planetary and Space Science*, 2017. [*Citado en pág. 16.*]

[38] R. Theron, "Visual analytics of paleoceanographic conditions," in *2006 IEEE Symposium On Visual Analytics Science And Technology*, Oct 2006, pp. 19–26. [*Citado en pág. 16.*]

[39] R. A. Bridges, J. Collins, E. M. Ferragut, J. Laska, and B. D. Sullivan, "A multi-level anomaly detection algorithm for time-varying graph data with interactive visualization," *Social Network Analysis and Mining*, vol. 6, no. 1, 2016. [*Citado en pág. 16.*]

[40] J. R. Goodall and M. Sowul, "Viassist: Visual analytics for cyber defense," in *Technologies for Homeland Security, 2009. HST'09. IEEE Conference on.* IEEE, 2009, pp. 143–150. [*Citado en pág. 16.*]

[41] D. A. Ellis and H. L. Merdian, "Thinking outside the box: Developing dynamic data visualizations for psychology with shiny," *Frontiers in Psychology*, vol. 6, no. DEC, 2015. [*Citado en pág. 16.*]

[42] D. H. Huson, S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H. J. Ruscheweyh, and R. Tappu, "Megan community edition - interactive exploration and analysis of large-scale microbiome sequencing data," *PLoS Computational Biology*, vol. 12, no. 6, 2016. [*Citado en pág. 16.*]

[43] R. Santamaría and R. Therón, "Treevolution: visual analysis of phylogenetic trees," *Bioinformatics*, vol. 25, no. 15, p. 1970, 2009. [*Citado en pág. 16.*]

[44] R. Santamaría, R. Therón, and L. Quintales, "A visual analytics approach for understanding biclustering results from microarray data," *BMC bioinformatics*, vol. 9, no. 1, p. 247, 2008. [*Citado en pág. 16.*]

[45] N. Harte, V. Silventoinen, E. Quevillon, S. Robinson, K. Kallio, X. Fustero, P. Patel, P. Jokinen, and R. Lopez, "Public web-based services from the european bioinformatics institute," *Nucleic Acids Research*, vol. 32, no. WEB SERVER ISS., pp. W3–W9, 2004. [*Citado en pág. 16.*]

[46] R. Santamaría, R. Therón, and L. Quintales, "Bicoverlapper: a tool for bicluster visualization," *Bioinformatics*, vol. 24, no. 9, pp. 1212–1213, 2008. [*Citado en pág. 16.*]

[47] R. Belford and E. B. Moore, "Confchem conference on interactive visualizations for chemistry teaching and learning: An introduction," *Journal of Chemical Education*, vol. 93, no. 6, pp. 1140–1141, 2016. [*Citado en pág. 16.*]

[48] G. Hesson, A. C. M. Moskal, and K. Shephard, "Using visual analytics to explore community engaged learning and teaching at the university of otago," in *31st Annual Conference of the Australian Society for Computers in Tertiary Education, ASCILITE 2014.* ASCILITE, 2014, Conference Proceedings, pp. 500–504. [*Citado en pág. 16.*]

[49] D.-A. Gómez-Aguilar, F.-J. García-Peñalvo, and R. Therón, "Analítica visual en e-learning," *El profesional de la información*, vol. 23, no. 3, 2014. [*Citado en pág. 16.*]

[50] D. A. G. Aguilar, R. Therón, and F. J. García-Peñalvo, "Semantic spiral timelines used as support for e-learning." *Journal of Universal Computer Science*, 2009. [*Citado en pág. 16.*]

[51] D. A. Gómez-Aguilar, Á. Hernández-García, F. J. García-Peñalvo, and R. Therón, "Tap into visual analysis of customization of grouping of activities in elearning," *Computers in Human Behavior*, vol. 47, pp. 60–67, 2015. [*Citado en pág. 16.*]

[52] J. Yee, R. F. Mills, G. L. Peterson, and S. E. Bartczak, "Automatic generation of social network data from electronic-mail communications," DTIC Document, Report, 2005. [*Citado en pág. 16.*]

[53] O. Noppens and T. Liebig, "Realizing the hidden - interactive visualization and analysis of large volumes of structured data," in *Working Conference on Advanced Visual Interfaces, AVI 08*, 2008, Conference Proceedings, pp. 441–444. [*Citado en pág. 16.*]

[54] M. Lungu, M. Lanza, T. Gîrba, and R. Robbes, "The small project observatory: Visualizing software ecosystems," *Science of Computer Programming*, vol. 75, no. 4, pp. 264–275, 2010. [*Citado en pág. 16.*]

[55] M. Lungu and M. Lanza, "The small project observatory - a tool for reverse engineering software ecosystems," in *32nd ACM/IEEE International Conference on Software Engineering, ICSE 2010*, vol. 2, 2010, Conference Proceedings, pp. 289–292. [*Citado en pág. 16.*]

[56] A. González-Torres, F. J. García-Peñalvo, R. Therón-Sánchez, and R. Colomo-Palacios, "Knowledge discovery in software teams by means of evolutionary visual software analytics," *Science of Computer Programming*, vol. 121, pp. 55 – 74, 2016, special Issue on Knowledge-based Software Engineering. [*Citado en pág. 16.*]

[57] A. González-Torres, F. J. García-Peñalvo, and R. Therón, "Human?computer interaction in evolutionary visual software analytics," *Computers in Human Behavior*, vol. 29, no. 2, pp. 486 – 495, 2013, advanced Human-Computer Interaction. [*Citado en pág. 16.*]

[58] R. Therón, A. González, F. J. García, and P. Santos, *The Use of Information Visualization to Support Software Configuration Management.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 317–331. [*Citado en pág. 16.*]

[59] P. Lamb and H. Croft, "Visualizing rugby game styles using self-organizing maps," *IEEE Computer Graphics and Applications*, vol. 36, no. 6, pp. 11–15, Nov 2016. [*Citado en pág. 16.*]

[60] M. Lage, J. P. Ono, D. Cervone, J. Chiang, C. Dietrich, and C. T. Silva, "Statcast dashboard: Exploration of spatiotemporal baseball data," *IEEE Computer Graphics and Applications*, vol. 36, no. 5, pp. 28–37, Sept 2016. [*Citado en pág. 16.*]

[61] R. Therón and L. Casares, *Visual Analysis of Time-Motion in Basketball Games*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 196–207. [*Citado en pág. 16.*]

[62] A. G. Losada, R. Therón, and A. Benito, "Bkviz: A basketball visual analysis tool," *IEEE Computer Graphics and Applications*, vol. 36, no. 6, pp. 58–68, Nov 2016. [*Citado en pág. 16.*]

[63] N. Hochman and R. Schwartz, "Visualizing instagram: Tracing cultural visual rhythms," in *International AAAI Conference on Web and Social Media*, 2012. [*Citado en pág. 16.*]

[64] A. Benito, A. G. Losada, R. Therón, A. Dorn, M. Seltmann, and E. Wandl-Vogt, "A spatio-temporal visual analysis tool for historical dictionaries," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, ser. TEEM '16. New York, NY, USA: ACM, 2016, pp. 985–990. [*Citado en pág. 16.*]

[65] R. Theron and L. Fontanillo, "Diachronic-information visualization in historical dictionaries," *Information Visualization*, vol. 14, no. 2, pp. 111–136, 2015. [*Citado en pág. 16.*]

[66] Y. Abdelsadek, K. Chelghoum, F. Herrmann, I. Kacem, and B. Otjacques, "Visual interactive approach for mining twitter's networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, 2016, vol. 9714, pp. 342–349. [*Citado en págs. 17 y 68.*]

[67] A. Sallaberry, F. Zaidi, C. Pich, and G. Melançon, "Interactive visualization and navigation of web search results revealing community structures and bridges," in *36th Graphics Interface Conference, GI 2010*, 2010, Conference Proceedings, pp. 105–112. [*Citado en págs. 17, 68, 69 y 70.*]

[68] W. Ribarsky, D. Xiaoyu Wang, and W. Dou, "Social media analytics for competitive advantage," *Computers and Graphics (Pergamon)*, vol. 38, no. 1, pp. 328–331, 2014. [*Citado en págs. 17 y 59.*]

[69] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, 2002. [*Citado en págs. 17, 30, 68 y 69.*]

[70] A. Inselberg and B. Dimsdale, *Parallel Coordinates for Visualizing Multi-Dimensional Geometry*. Tokyo: Springer Japan, 1987, pp. 25–44. [*Citado en págs. 17, 36 y 66.*]

[71] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *2nd IEEE Conference on Visual Analytics Science and Technology 2011, VAST 2011*, 2011, Conference Proceedings, pp. 231–240. [*Citado en págs. 17, 30, 36, 68 y 69.*]

[72] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," *Queue*, vol. 10, no. 2, pp. 30:30–30:55, Feb. 2012. [*Citado en pág. 17.*]

[73] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Information Visualization, 2006. IV 2006. Tenth International Conference on.* IEEE, 2006, pp. 9–16. [*Citado en pág. 17.*]

[74] Z. Liu and J. Heer, "The effects of interactive latency on exploratory visual analysis," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2122–2131, 2014. [*Citado en pág. 17.*]

[75] T. R. G. Green and M. Petre, "Usability analysis of visual programming environments: a cognitive dimensions framework," *Journal of Visual Languages & Computing*, vol. 7, no. 2, pp. 131–174, 1996. [*Citado en pág. 17.*]

[76] E. Hoque, S. Joty, L. Màrquez, and G. Carenini, "CQAVis: Visual text analytics for community question answering," in *22nd International Conference on Intelligent User Interfaces, IUI 2017*, vol. Part F126745. Association for Computing Machinery, 2017, Conference Proceedings, pp. 161–172. [*Citado en pág. 17.*]

[77] R. Zafarani and H. Liu. Social computing data repository at ASU. [Online]. Available: http://socialcomputing.asu.edu [*Citado en pág. 18.*]

[78] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014. [*Citado en pág. 18.*]

[79] J. Yee, R. F. Mills, G. L. Peterson, and S. E. Bartczak, "Automatic generation of social network data from electronic-mail communications," DTIC Document, Report, 2005. [*Citado en pág. 18.*]

[80] H. Pérez-Rosés and F. Sebé, "Synthetic generation of social network data with endorsements," *Journal of Simulation*, vol. 9, no. 4, pp. 279–286, 2015. [*Citado en pág. 18.*]

[81] A. Prat and X. Sanchez. Ldbc-snb data generator. [Online]. Available: https://github.com/ldbc/ldbc_snb_datagen [*Citado en pág. 18.*]

[82] O. Erling, A. Averbuch, J. Larriba-Pey, H. Chafi, A. Gubichev, A. Prat, M.-D. Pham, and P. Boncz, "The ldbc social network benchmark: Interactive workload," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 619–630. [*Citado en pág. 18.*]

[83] M.-D. Pham, P. Boncz, and O. Erling, "S3g2: A scalable structure-correlated social graph generator," in *Technology Conference on Performance Evaluation and Benchmarking*. Springer, 2012, Conference Proceedings, pp. 156–172. [*Citado en págs. 18 y 20.*]

[84] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 01, pp. 57–87, 1997. [*Citado en pág. 19.*]

[85] E. Reiter, R. Dale, and Z. Feng, *Building natural language generation systems*. MIT Press, 2000, vol. 33. [*Citado en pág. 19.*]

[86] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. [*Citado en pág. 19.*]

[87] F. J. García-Peñalvo and J. Durán-Escudero, *Interaction Design Principles in WYRED Platform*. Springer International Publishing, 2017, pp. 371–381. [*Citado en pág. 20.*]

[88] A. Lenhart. Teens, social media & technology overview 2015. [Online]. Available: http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/ [*Citado en pág. 21.*]

[89] Agencia Española de Protección de Datos. Facebook adecua a la legislación española la edad mínima de sus usuarios. [Online]. Available: https://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2010/notas_prensa/common/febrero/180210_Facebook_adecua_legislacion_es_edad_min_usuarios.pdf [*Citado en pág. 21.*]

[90] S. Greenwood, A. Perrin, and M. Duggan. Social media update 2016. [Online]. Available: http://www.pewinternet.org/2016/11/11/social-media-update-2016/ [*Citado en pág. 21.*]

[91] OCDE. Adult education level. [Online]. Available: https://data.oecd.org/eduatt/adult-education-level.htm#indicator-chart [*Citado en pág. 21.*]

[92] Ministerio de Educación, Cultura y Deporte. Datos y cifras. curso escolar 2015-2016. [Online]. Available: https://www.mecd.gob.es/servicios-al-ciudadano-mecd/dms/mecd/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/datos-cifras/Datosycifras1516esp.pdf [*Citado en pág. 21.*]

[93] INE. Población de 16 y más años por nivel de formación alcanzado, sexo y grupo de edad. [Online]. Available: http://www.ine.es/jaxiT3/Datos.htm?t=6347 [*Citado en pág. 21.*]

[94] WYRED. Wyred stakeholder questionnaire. english version. [Online]. Available: https://repositorio.grial.eu/handle/grial/800 [*Citado en pág. 22.*]

[95] M. Richards, "Software architecture patterns," *O'Reilly Media*, 2015. [*Citado en pág. 25.*]

[96] N. Günnemann and M. P. D. Jarke, "D-vita: A visual interactive text analysis system using dynamic topic mining," in *Workshopband Datenbanksysteme fur Business, Technologie und Web, BTW 2013 - Workshop on Database Systems for Business, Technology and Web, BTW 2013*, V. Koppen, T. Neumann, A. Henrich, G. Saake, and W. Lehner, Eds., vol. P-216. Gesellschaft fur Informatik (GI), 2013, Conference Proceedings, pp. 237–246. [*Citado en pág. 25.*]

[97] Microsoft. Model-view-controller. [Online]. Available: https://msdn.microsoft.com/en-us/library/ff649643.aspx [*Citado en pág. 25.*]

[98] Microsoft. The mvvm pattern. [Online]. Available: https://msdn.microsoft.com/en-us/library/hh848246.aspx [*Citado en pág. 25.*]

[99] D. Malandrino, I. Manno, G. Palmieri, A. Petta, D. Pirozzi, V. Scarano, L. Serra, C. Spagnuolo, L. Vicidomini, and G. Cordasco, "An architecture for social sharing and collaboration around open data visualisations," in *19th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2016*, vol. 26-February-2016. Association for Computing Machinery, 2016, Conference Proceedings, pp. 357–360. [*Citado en pág. 25.*]

[100] Docker. What is docker. [Online]. Available: https://www.docker.com/what-docker [*Citado en pág. 25.*]

[101] K. Matkovic, W. Freiler, D. Gracanin, and H. Hauser, "Comvis: A coordinated multiple views system for prototyping new visualization technology," in *12th International Conference Information Visualisation, IV08*, 2008, Conference Proceedings, pp. 215–220. [*Citado en pág. 25.*]

[102] Apereo. About cas. [Online]. Available: https://www.apereo.org/projects/cas/about-cas [*Citado en pág. 27.*]

[103] Z. A. Pardos and K. Kao, "Moocrp: An open-source analytics platform," in *2nd ACM Conference on Learning at Scale, L@S 2015*. Association for Computing Machinery, Inc, 2015, Conference Proceedings, pp. 103–110. [*Citado en pág. 27.*]

[104] F. Huang, C. x. Wang, and J. Long, "Design and implementation of single sign on system with cluster cas for public service platform of science and technology evaluation," in *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, Nov 2011, pp. 732–737. [*Citado en pág. 27.*]

[105] J. Pokorny, "Nosql databases: a step to database scalability in web environment," *International Journal of Web Information Systems*, vol. 9, no. 1, pp. 69–82, 2013. [*Citado en pág. 27.*]

[106] R. Cattell, "Scalable sql and nosql data stores," *SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, May 2011. [*Citado en pág. 27.*]

[107] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002. [*Citado en pág. 27.*]

[108] OpenAire. What is the open research data pilot? [Online]. Available: https://www.openaire.eu/opendatapilot [*Citado en pág. 28.*]

[109] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003. [*Citado en págs. 29 y 69.*]

[110] E. Chen. What is a good explanation of latent dirichlet allocation? [Online]. Available: https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation [*Citado en pág. 29.*]

[111] J. Boyd-Graber and D. M. Blei, "Multilingual topic models for unaligned text," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 75–82. [*Citado en pág. 29.*]

[112] J. Jagarlamudi and H. Daumé III, "Extracting multilingual topics from unaligned comparable corpora," in *European Conference on Information Retrieval*. Springer, 2010, pp. 444–456. [*Citado en pág. 29.*]

[113] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 543–552. [*Citado en pág. 30.*]

[114] W. Ribarsky, D. Xiaoyu Wang, and W. Dou, "Social media analytics for competitive advantage," *Computers and Graphics (Pergamon)*, vol. 38, no. 1, pp. 328–331, 2014. [*Citado en págs. 30, 68 y 69.*]

[115] M. Stone, "In color perception, size matters," *IEEE Computer Graphics and Applications*, vol. 32, no. 2, pp. 8–13, March 2012. [*Citado en pág. 31.*]

[116] C. G. Healey, "Choosing effective colours for data visualization," in *Visualization'96. Proceedings.* IEEE, 1996, pp. 263–270. [*Citado en pág. 31.*]

[117] C. Brewer. (2017) Color brewer 2.0. [Online]. Available: http://colorbrewer2.org [*Citado en pág. 31.*]

[118] T. Murata and K. Takeichi, "Discovering and visualizing network communities," in *2007 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT Workshops 2007*, 2007, Conference Proceedings, pp. 217–220. [*Citado en pág. 33.*]

[119] M. Jacomy. I want hue. [Online]. Available: http://tools.medialab.sciences-po.fr/iwanthue/ [*Citado en pág. 35.*]

[120] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in *Proceedings of the 1st conference on Visualization'90*. IEEE Computer Society Press, 1990, Conference Proceedings, pp. 361–378. [*Citado en pág. 36.*]

[121] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates," in *Eurographics (STARs)*, 2013, Conference Proceedings, pp. 95–116. [*Citado en pág. 37.*]

[122] H. To, G. Ghinita, and C. Shahabi, "Privgeocrowd: A toolbox for studying private spatial crowdsourcing," in *2015 31st IEEE International Conference on Data Engineering, ICDE 2015*, vol. 2015-May. IEEE Computer Society, 2015, Conference Proceedings, pp. 1404–1407. [*Citado en pág. 38.*]

[123] D. Guo, "Flow mapping and multivariate visualization of large spatial interaction data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1041–1048, 2009. [*Citado en págs. 38, 68 y 69.*]

[124] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, ser. VL '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 336–. [*Citado en pág. 41.*]

[125] D. A. Keim, F. Mansmann, and J. Thomas, "Visual analytics: How much visualization and how much analytics?" *SIGKDD Explor. Newsl.*, vol. 11, no. 2, pp. 5–8, May 2010. [*Citado en pág. 41.*]

[126] J. C. Roberts, "State of the art: Coordinated multiple views in exploratory visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, July 2007, pp. 61–71. [*Citado en pág. 43.*]

[127] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems," in *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, ser. CHI '93. New York, NY, USA: ACM, 1993, pp. 206–213. [*Citado en pág. 53.*]

[128] W. A. Pike, J. Stasko, R. Chang, and T. A. O'connell, "The science of interaction," *Information Visualization*, vol. 8, no. 4, pp. 263–274, 2009. [*Citado en págs. 54 y 68.*]

[129] A. García-Holgado and F. J. García-Peñalvo, "The evolution of the technological ecosystems: An architectural proposal to enhancing learning processes," in *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, ser. TEEM '13. New York, NY, USA: ACM, 2013, pp. 565–571. [Online]. Available: http://doi.acm.org/10.1145/2536536.2536623 [*Citado en pág. 55.*]

[130] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014. [*Citado en pág. 59.*]

[131] D. Huang, B. Sherman, and R. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009. [*Citado en pág. 59.*]

[132] Z. A. Pardos and K. Kao, "Moocrp: An open-source analytics platform," in *2nd ACM Conference on Learning at Scale, L@S 2015*. Association for Computing Machinery, Inc, 2015, Conference Proceedings, pp. 103–110. [*Citado en pág. 59.*]

[133] N. A. Granitz and J. C. Ward, "Virtual community: A sociocognitive analysis," *NA-Advances in Consumer Research Volume 23*, 1996. [*Citado en pág. 59.*]

[134] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004. [*Citado en págs. 60 y 63.*]

[135] F. W. Neiva, J. M. N. David, R. Braga, and F. Campos, "Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature," *Information and Software Technology*, vol. 72, pp. 137 – 150, 2016. [*Citado en pág. 61.*]

[136] The Center for Evidence-Based Management. What is a picoc? [Online]. Available: https://www.cebma.org/faq/what-is-a-picoc/ [*Citado en pág. 61.*]

[137] A. Harzing and S. Alakangas, "Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, pp. 787–804, 2016. [*Citado en pág. 62.*]

[138] K. Reda, C. Tantipathananandh, A. Johnson, J. Leigh, and T. Berger-Wolf, "Visualizing the evolution of community structures in dynamic social networks," *Computer Graphics Forum*, vol. 30, no. 3, pp. 1061–1070, 2011. [*Citado en págs. 68, 69 y 70.*]

[139] N. K. Ahmed and R. A. Rossi, "Interactive visual graph analytics on the web," in *9th International Conference on Web and Social Media, ICWSM 2015*. AAAI Press, 2015, Conference Proceedings, pp. 566–569. [*Citado en págs. 68 y 70.*]

[140] S. Conglei, F. Siwei, C. Qing, and Q. Huamin, "Vismooc: Visualizing video clickstream data from massive open online courses," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 2015, Conference Proceedings, pp. 159–166. [*Citado en págs. 68 y 69.*]

[141] E. Meeks, *D3.js in action data visualization with JavaScript*. Manning, 2017. [*Citado en pág. 73.*]

[142] M. Heydt, *D3.js by example*. Packt Publishing, 2015. [*Citado en pág. 73.*]

[143] Google. Google charts. [Online]. Available: https://developers.google.com/chart/ [*Citado en pág. 74.*]

[144] M. Bostock. D3 gallery. [Online]. Available: https://github.com/d3/d3/wiki/Gallery [*Citado en pág. 74.*]

[145] J. Ferraiolo, F. Jun, and D. Jackson, *Scalable vector graphics (SVG) 1.0 specification*. iuniverse, 2000. [*Citado en pág. 74.*]

[146] A. Deveria. Can i use svg? [Online]. Available: http://caniuse.com/#search=svg [*Citado en pág. 74.*]

With the support of the EU Horizon 2020 Programme in its "Europe in a changing world – inclusive, innovative and reflective Societies (HORIZON 2020: REV-INEQUAL-10-2016: Multi-stakeholder platform for enhancing youth digital opportunities)" Call. Project WYRED (netWorked Youth Research for Empowerment in the Digital society) (Grant agreement No 727066). The sole responsibility for the content of this document lies with the authors. It does not necessarily reflect the opinion of the European Union. The European Commission is not responsible for any use that may be made of the information contained therein.