Analítica Visual de Datos para Representación de la Interacción en una Red Social Privada y con Restricciones de Privacidad

Jorge Durán-Escudero[1], Francisco J. García-Peñalvo[1] y Roberto Therón Sánchez[1]

Departamento de Informática y Automática, Facultad de Ciencias
Plaza de los Caídos s/n, 37008, Salamanca, España
{jorge.d,fgarcia,theron}@usal.es

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. 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.

Keywords: Analítica visual, redes sociales, arquitectura, generación de datos, interacción de usuarios, interacción en redes sociales.

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 [22]. 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.

El proyecto WYRED [8] 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 ecosistema 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 [9]. En el caso de WYRED, 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.

La propuesta de arquitectura que se presenta en este trabajo está centrada en la comunidad de WYRED, la cual es similar a un foro, ya que los usuarios organizan los debates en comunidades, hilos de discusión y comentario. Sin embargo, la comunidad también está preparada para albergar otras interacciones más complejas como diálogos sociales o proyectos de investigación. Al mismo tiempo, el proyecto tiene una serie de características que lo distinguen del resto [7], como su contexto internacional (varios idiomas, diferentes características socioculturales y muy variados puntos de vista) o la necesidad de salvaguardar la privacidad de los usuarios. Esto es necesario porque muchos de ellos serán menores y para transmitir que es un lugar donde pueden interactuar libremente. Debido a la gran cantidad de datos que el proyecto va a generar, se propone el uso de la analítica visual como una técnica efectiva para representar y extraer conocimiento [16].

El objetivo principal de este trabajo es plantear en estas primeras etapas del proyecto WYRED, una propuesta de la 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 este instante o en un futuro. Para ello debe ayudar a los investigadores en dos tareas principales:

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

Por ello, aunque el principal objeto de estudio del proyecto son 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 influir en las 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.

El artículo se organiza en las siguientes secciones: en primer lugar se trata la generación del conjunto de datos, después se presenta la propuesta de arquitectura, a continuación los resultados obtenidos y un caso de uso donde se demuestra el desempeño de la propuesta planteada y finalmente las principales conclusiones obtenidas y las futuras líneas de investigación.

2. Generación del Conjunto de Datos

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 una propuesta de arquitectura. Es por ello que se ha tomado la decisión de intentar generarlo, para ello las principales maneras para afrontarlo son las siguientes:

- Utilizar los datos de una comunidad similar.
- Usar otras fuentes de datos que tengan características comunes.
- Generar los datos de manera artificial.

Extraer los datos de alguna comunidad similar es el proceso más simple y rápido para obtener un conjunto de datos. Para ello se pueden utilizar las principales redes sociales (Twitter, Facebook, etc.), las cuales han sido analizadas en profundidad por multitud de autores que han publicado sus conjuntos de datos [35] [19]. Pero esta solución no es válida en todos los casos, al tratarse de comunidades demasiado genéricas y de datos anonimizados.

Otros autores [34] han propuesto utilizar las entradas en los ficheros de registro, para generar el conjunto de datos. De 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 artificialmente. Este sistema 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.

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 [30], 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 [29] el cual es un programa desarrollado para generar conjuntos de datos de comunidades para LDBC (*Linked Data Benchmark Council*) [5]. Para asignar a los atributos valores de manera lógica, los autores se basan en S3G2 [26] 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.

En un primer momento, para construir el conjunto de datos se intentó utilizar LDBC-SNB Data Generator, 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. Es por ello que se ha decidido construir desde cero el propio conjunto de datos, para ello se han dado los siguientes pasos:

1. Análisis de las entidades que hay que simular.

2. Identificación de sus principales atributos.
3. Creación del grafo de dependencia entre los mismos según el modelo descrito en S3G2 [26]. En la Fig. 1 se puede ver un ejemplo del mismo.
4. Asignación de valores a cada uno de los atributos según los atributos de los que dependen y los valores que estos presentan de manera usual. Para ello se han utilizado distintos indicadores como la población de un país, las expectativas de uso o distintos estudios sociológicos [18] [10].

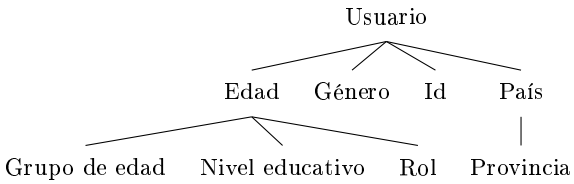


Figura 1: Dependencia entre los atributos de un usuario

3. Propuesta de Arquitectura

Una propuesta de arquitectura 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 el caso 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 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 soportarlos se ha decidido utilizar una arquitectura denominada de micronúcleo [32], que 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.

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 [21].

Tomando como referencia la arquitectura de Docker (https://goo.gl/LGk7vj) se ha diseñado la propuesta reflejada en la Fig. 2, 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.

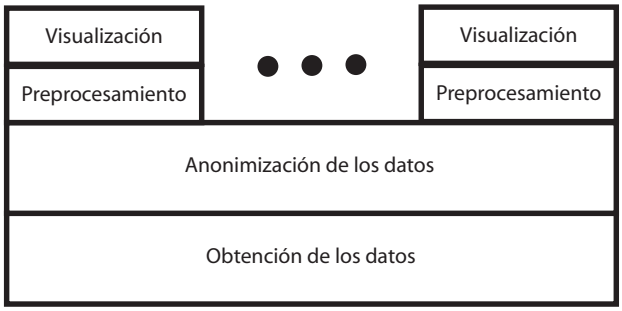


Figura 2: Esquema de la arquitectura propuesta

3.1. 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.

La información privada de los usuarios se almacena en un CAS, *Central Authentication Service* (https://goo.gl/xD4Jkg), siguiendo la estela de otros estudios que se han enfrentado a este problema [25] [12].

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 [28] [3].

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

3.2. 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.

La manera de trabajar con parte de estos datos es sencilla, ya que cuestiones 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 están anonimizados, ya que por la combinación de

los datos restantes puede ser posible identificar al usuario inicial [33]. 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.
- 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. La utilización de este valor permite asegurar la anonimicidad de los datos, que serán publicados de manera abierta, para que otros investigadores puedan utilizarlos, tal y como estipula la Unión Europea para los proyectos financiados bajo el paraguas del Horizonte 2020 [24].

3.3. 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, 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.).

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 usando LDA (*Latent Dirichlet Allocation*) [1]. Aunque esto platea algunos problemas como la gestión de varios idiomas, en cuyo soporte han trabajado algunos autores [2][14] o la obtención de un nombre representativo para cada tema.

Para 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 medir la influencia de los atributos de los usuarios.

Teniendo en cuenta lo anterior, había que seleccionar un tipo de gráfico de entre los muchos existentes [17]. 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.

En un principio se pensó utilizar una visualización basada en el concepto de *Theme River* [11]. Este sistema ya se había usado de manera efectiva en otros trabajos [4] [20], 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. Para reducir estas desventajas, se ha combinado esta representación con otra basada en representar cada tema de manera individual, sobre líneas temporales paralelas [31]. 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 según los atributos de los usuarios.
- Soporte para reordenar los temas, ya que es más sencillo comparar aquellos que están más próximos entre sí.
- Posibilidad de conocer el nivel de relevancia de ese tema en un instante.
- Capacidad de hacer *zoom* de manera automática.
- Posibilidad de restringir la selección a un periodo temporal.

3.4. Módulo para la Detección de Comunidades

Otro de los aspectos más relevantes a la hora de explorar una comunidad, es detectar las comunidades que implícitamente crean los usuarios. Para ello se han utilizado multitud de técnicas como el *clustering* jerárquico, la detección de nodos centrales o el cáclulo de la centralidad [6], pero esto requiere de la ejecución de complejos algoritmos. Por ello se propone abordar esta tarea mediante la visualización interactiva.

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. En el caso concreto de la visualización que se propone, las relaciones hacen referencia al número de comentarios que intercambian.

Respecto a la visualización se plantea codificar cada nodo con un tamaño relativo al número de mensajes que ha publicado en la plataforma. Además, la longitud de los enlaces debe representar la cercanía de un nodo respecto a otro, atendiendo a las interacciones que han tenido. Para ello se introduce el concepto de distancia relativa d_r entre dos nodos A y B, como un valor proporcional al número de enlaces totales de cada nodo entre el número de enlaces que comparten:

$$d_r(A, B) = k * \frac{g(A) + g(B)}{E(A, B)} \tag{1}$$

Respecto a las características de interacción se ha establecido las siguientes:

- Posibilidad de conocer las características de un usuario, al visitar su nodo.
- Capacidad de hacer *zoom* para analizar en mayor profundidad el grafo y ayudar a que esta visualización pueda escalar bien.

- Soporte para seleccionar un conjunto de nodos y explorarlos.
- Posibilidad de mover y analizar en detalle cada una de las comunidades que se formen.

3.5. Módulo para la Exploración de los Usuarios

El problema de representar los atributos de los usuarios de una plataforma es bastante complejo, debido al gran número de usuarios y características a mostrar. Por estas razones, se hace necesario el uso de una visualización que escale bajo demanda, compacta y sencilla a la hora de ser interpretada. Por ello se eligió utilizar las coordenadas paralelas [13], ya que permiten representar en un contexto bidimensional n dimensiones o atributos. Además se propone dotarlas de las siguientes características de interacción:

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

3.6. 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 como afecta esta dimensión al análisis de los datos de la plataforma, siendo la visualización que mejor representa este concepto el mapa. Sin embargo, hay multitud de tipos de mapas, tanto atendiendo a las características que representan como a la proyección que utilizan. En la propuesta, se ha elegido utilizar la proyección Mercator, por ser la más familiar, para representar los países y las regiones del mundo. Además, se va a utilizar el color para representar el número de mensajes que han sido generados por los usuarios de cada uno de los territorios. Respecto a las características de interacción, se proponen las siguientes:

- Posibilidad de moverse por el mapa y volver al punto inicial.
- *Zoom* semántico, mostrando los países o las provincias según corresponda.
- Capacidad de conocer el número de mensajes exacto en cada país.
- Posibilidad de filtrar los datos por país.

4. Resultados

Para desarrollar la arquitectura propuesta, se ha recurrido a utilizar tecnologías y lenguajes de programación web. Esta decisión permite principalmente dos cosas: conseguir que los desarrollos sean accesibles a un mayor público y dotarlos de un mayor grado de interactividad.

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. 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 analítica visual enunciado por Keim *Analyze first, show the important, zoom, filter and analyze further, details on demand* [15].

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, para constituir un panel de monitorización que permita explorar todas las facetas del proyecto, al mismo tiempo, como se puede ver en la Fig. 3.

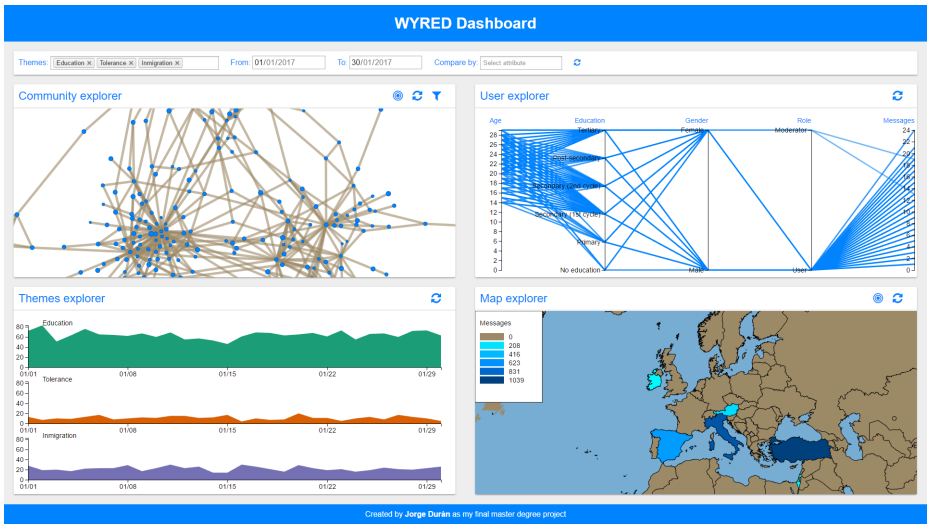


Figura 3: Panel de monitorización del proyecto[1]

Para demostrar cómo el sistema propuesto podría ser usado para resolver algunas preguntas de investigación típicas, a continuación se describe un caso de uso del sistema, cuya pregunta de investigación es *¿Cuáles son las principales comunidades sobre educación y empleo, y qué características tienen?*.

Lo primero de debe realizar un investigador es identificar los principales temas que están presentes en la pregunta de investigación, en este caso, educación y empleo. Por esta razón, tendrá que seleccionar ambos en el selector de temas, como se puede apreciar en el extremo izquierdo de la Fig. 4.

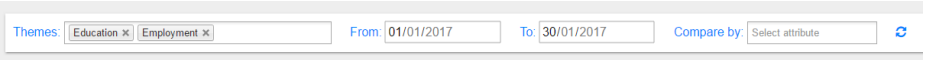
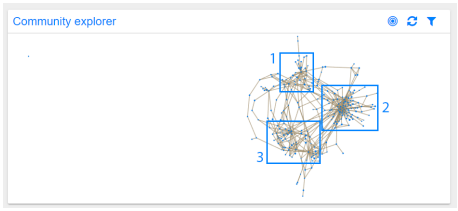


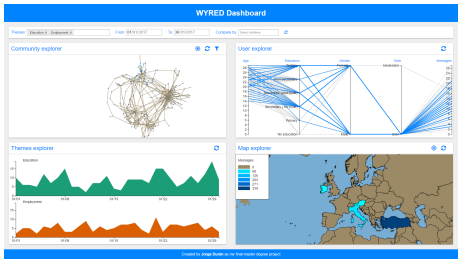
Figura 4: Selección de los temas para la pregunta de investigación

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

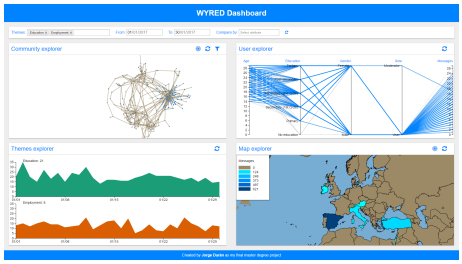
Después, con la visualización creada por el explorador de comunidades podrá identificar las principales comunidades que se han formado, teniendo en cuenta los grupos de usuarios más compactos. En la Fig. 5a se puede apreciar que, fundamentalmente, los usuarios se agrupan en torno a 3 comunidades, las cuales han sido marcadas para facilitar la demostración. Para explorar una comunidad, el investigador deberá incluir en el rectángulo de selección, que aparece al hacer clic en la visualización, todos los puntos que, a su juicio, forman parte de esa comunidad, los cuales mantendrán su color azul para favorecer su identificación.



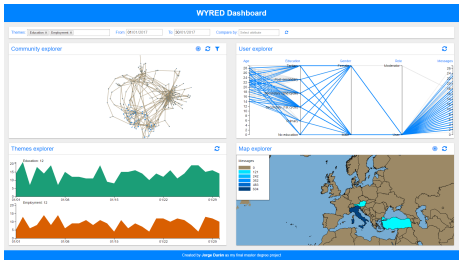
(a) Identificación de las comunidades



(b) Datos de la primera comunidad



(c) Datos de la segunda comunidad



(d) Datos de la tercera comunidad

Figura 5: Principales comunidades y sus datos[2]

Al analizar los datos de la primera comunidad (Fig. 5b), se puede ver cómo los principales usuarios son de Turquía, debido a que este país es el que presenta un tono más oscuro en el mapa. Respecto a los temas, los mimos hablan más sobre educación que sobre empleo, como muestra el explorador de temas.

En el caso de la segunda comunidad (Fig. 5c), se puede ver como los usuarios son principalmente españoles, que en momentos puntuales hablan mucho sobre empleo, pero que de manera más continua, hablan más de educación, como se aprecia en el explorador de temas.

Si se observan los datos de la tercera comunidad (Fig. 5d), se puede apreciar como está constituida principalmente por italianos cuyo comportamiento es

[2] Accesible en `https://jorge-duran.com/research/tfm/Articulos/JorgeDuran/img/`

similar a los usuarios de la segunda comunidad. Sin embargo, la comunidad presenta un número inusualmente bajo de usuarios con educación postsecundaria, lo cual se puede ver al comprobar en el explorador de usuarios, que la mayor parte de las líneas que atraviesan la marca de postsecundaria están en gris.

Para mostrar las características interactivas de este desarrollo y como un investigador usaría estas visualizaciones para resolver la pregunta de este caso de uso, se ha grabado un vídeo (https://goo.gl/js3hkp) donde se puede comprobar todo el proceso.

5. Conclusiones y Futuras Líneas de Investigación

En este trabajo se ha presentado la propuesta de arquitectura para elaborar un conjunto de visualizaciones interactivas que permitan explorar los datos del proyecto WYRED. La cual se basa en la arquitectura de micronúcleo que consta de 2 capas básicas (adquisición y anonimización de datos) y de 4 módulos (exploración de temas, comunidades, usuarios y geográfica), que han permitido cumplir los objetivos planteados en un principio. Aunque el mismo podría ser ampliado con la puesta en marcha de un estudio de usabilidad con 5 usuarios [23], el soporte al trabajo colaborativo con las visualizaciones o permitiendo integrar el mismo con otros sistemas [27].

6. Agradecimientos

Este proyecto ha recibido fondos del programa de la Unión Europea de investigación e innovación Horizonte 2020 en virtud del acuerdo de subvención número 727066.

Referencias

1. 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.
2. 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.
3. R. Cattell, "Scalable sql and nosql data stores," *SIGMOD Rec.*, vol. 39, no. 4, pp. 12–27, May 2011.
4. 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.
5. 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.
6. S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

7. F. J. García-Peñalvo and J. Durán-Escudero, *Interaction Design Principles in WYRED Platform.* Springer International Publishing, 2017, pp. 371–381.
8. 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.
9. 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
10. S. Greenwood, A. Perrin, and M. Duggan. Social media update 2016. [Online]. Available: http://www.pewinternet.org/2016/11/11/social-media-update-2016/
11. 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.
12. 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.
13. A. Inselberg and B. Dimsdale, *Parallel Coordinates for Visualizing Multi-Dimensional Geometry.* Tokyo: Springer Japan, 1987, pp. 25–44.
14. J. Jagarlamudi and H. Daumé III, "Extracting multilingual topics from unaligned comparable corpora," in *European Conference on Information Retrieval.* Springer, 2010, pp. 444–456.
15. 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.
16. E. D. Keim, J. Kohlhammer, and G. Ellis, "Mastering the information age: Solving problems with visual analytics, eurographics association," 2010.
17. 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.
18. 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/
19. J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.
20. 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.
21. 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.
22. 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/

23. 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.
24. OpenAire. What is the open research data pilot? [Online]. Available: https://www.openaire.eu/opendatapilot
25. 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.
26. 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.
27. 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.
28. 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.
29. A. Prat and X. Sanchez. Ldbc-snb data generator. [Online]. Available: https://github.com/ldbc/ldbc_snb_datagen
30. 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.
31. 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.
32. M. Richards, "Software architecture patterns," *O'Reilly Media*, 2015.
33. 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.
34. 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.
35. R. Zafarani and H. Liu. Social computing data repository at ASU. [Online]. Available: http://socialcomputing.asu.edu