Analítica visual de datos para representación de la interacción en la plataforma WYRED

Jorge Durán-Escudero, Francisco José García-Peñalvo y Roberto Therón

Departamento de Informática y Automática, Facultad de Ciencias
Plaza de los Caídos 1, 37008, Salamanca, España
{jorge.d,fgarcia,theron}@usal.es

Resumen En este trabajo se realiza una propuesta de arquitectura para desarrollar visualizaciones interactivas, con el fin de explorar y extraer conocimiento de los datos que van a ser originados por los usuarios de la comunidad del proyecto WYRED. Además, en el mismo se analiza el impacto de la privacidad y distintos mecanismos para anonimizar los datos de los usuarios, los cuales van a ser generados de manera automática.

Key words: Analítica visual, redes sociales, arquitectura, generación de datos

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 [20]. 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.

El proyecto WYRED [8] es el marco de trabajo bajo el cual se desarrolla este proyecto cuyo objetivo es dar voz a los jóvenes, 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. Sin embargo, el proyecto cuenta con una serie de características propias que lo distinguen del resto, como su uso en un contexto internacional (varias lenguas, distintas características sociológicas y muy variados puntos de vista) o 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 [33]. Para abordar este problema, se platea el uso de la analítica visual como una técnica eficaz para realizar la representación y extracción de conocimiento con grandes volúmenes de datos [14].

2. Objetivos

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.

3. 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 generar un conjunto de datos de prueba lo más similar a los conjuntos reales con los que debería operar esta propuesta. 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 próxima a la que es objeto de estudio, es el proceso más simple y rápido para obtener un conjunto de datos. Para ello se pueden utilizar 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 [35] [17]. 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 algunos datos que son de mas 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. 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, mientras que el resto de los datos pueden servir para añadir contexto a los mismos.

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 [28], 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 [27] el cual es un programa desarrollado para generar conjuntos de datos de comunidades para LDBC (*Linked Data Benchmark Council*) [7]. Para asignar a los atributos valores de manera lógica, los autores se basan en S3G2 [24] 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 [24]. 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 [16] [9].

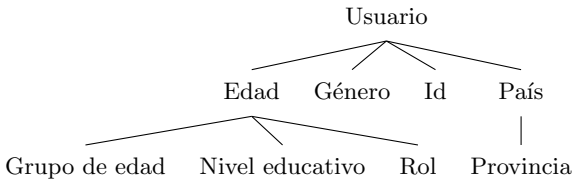

Figura 1. Dependencia entre los atributos de un usuario

4. 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 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. 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 soportar estos requisitos, se ha decidido utilizar una arquitectura denominada de micronúcleo [30]. 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.

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 [19].

Tomando como referencia la arquitectura de Docker [5] 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.

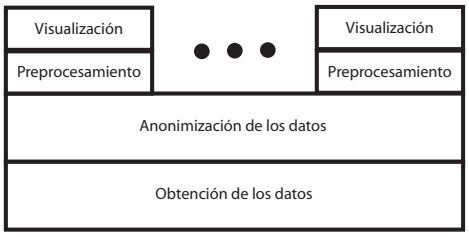

Figura 2. Esquema de la arquitectura propuesta

4.1. 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*) [1], siguiendo la estela de otros estudios que se han enfrentado a este problema [23] [11].

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 [26] [4].

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

4.2. 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 mismos 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 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 [32]. 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.
- 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 como fuente de información en sus investigaciones, tal y como estipula la Unión Europea para los proyectos financiados bajo el paraguas del Horizonte 2020 [22].

4.3. 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 utilizando para ello LDA (*Latent Dirichlet Allocation*) [2]. Uno de los problema 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 [3][13] han propuesto diversos métodos para poder soportarlo. Otro de los hándicap 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.

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.

Teniendo en cuenta lo anterior, había que seleccionar un tipo de gráfico de entre los muchos existentes [15]. 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* [10]. Este sistema ya se había usado de manera efectiva en otros trabajos [6] [18], 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 [29]. Esto permite aprovechar 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 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.

4.4. 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, 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. 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.

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 o lejanía 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 entre el número de enlaces que comparten:

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

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

Respecto a las características de interacción implementadas, para ayudar a resolver de manera sencilla y efectiva la tarea de la detección de subcomunidades, se ha 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.

4.5. 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 esta razón se eligió utilizar las coordenadas paralelas [12], ya que permiten representar en un contexto bidimensional n dimensiones o atributos.

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.

4.6. 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.
- 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.

5. 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 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* " [31].

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.

Una vez implementada la arquitectura, se han analizado los siguientes casos de uso:

1. ¿Cuáles son las principales comunidades sobre educación y empleo y qué características tienen?[2]
2. ¿Quiénes son los usuarios más activos de Turquía hablando sobre privacidad?[3]
3. ¿Cómo influye el género a la hora de hablar sobre la tolerancia y la inmigración? [4]
4. ¿Cuál es la evolución temporal de las discusiones sobre acoso, según los países participantes? [5]

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

[2] Disponible en `https://jorge-duran.com/research/tfm/videos/mainCommunities.mp4`

[3] Disponible en `https://jorge-duran.com/research/tfm/videos/privacyTurkey.mp4`

[4] Disponible en `https://jorge-duran.com/research/tfm/videos/gender.mp4`

[5] Disponible en `https://jorge-duran.com/research/tfm/videos/themesEvolution.mp4`

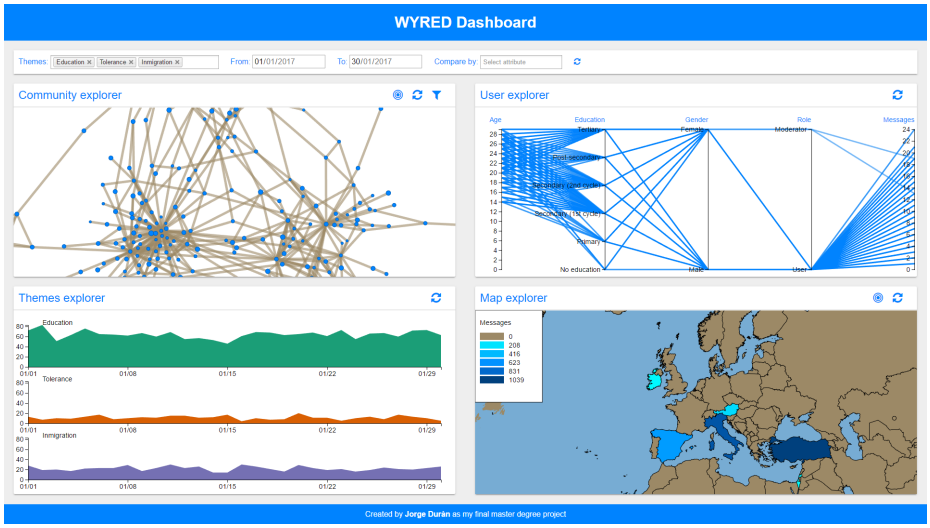

Figura 3. Panel de monitorización del proyecto[1]

6. Conclusiones y futuras líneas de investigación

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.

En este trabajo también se ha presentado la propuesta de arquitectura para elaborar 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. Por ello, se puede afirmar que este trabajo ha cumplido los objetivos planteados en un principio.

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 con usuarios de la usabilidad del sistema propuesto. Para ello habría que seleccionar a dichos usuarios, que podrían limitarse a 5 según el estudio de Nielsen [21].
- 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 [25].
- Abordar la integración del sistema propuesto con otros sistemas, para favorecer la labor investigadora [25].

7. Agradecimientos

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.

Referencias

1. Apereo: About cas, `https://www.apereo.org/projects/cas/about-cas`
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
3. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 75–82. AUAI Press (2009)
4. Cattell, R.: Scalable sql and nosql data stores. SIGMOD Rec. 39(4), 12–27 (May 2011)
5. Docker: What is docker, `https://www.docker.com/what-docker`
6. Dou, W., Wang, X., Chang, R., Ribarsky, W.: Paralleltopics: A probabilistic approach to exploring document collections. In: 2nd IEEE Conference on Visual Analytics Science and Technology 2011, VAST 2011. pp. 231–240 (2011)
7. Erling, O., Averbuch, A., Larriba-Pey, J., Chafi, H., Gubichev, A., Prat, A., Pham, M.D., Boncz, P.: The ldbc social network benchmark: Interactive workload. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 619–630. SIGMOD '15, ACM, New York, NY, USA (2015)
8. García-Peñalvo, F.J., Kearney, N.A.: 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). pp. 3–9. ACM (2016)
9. Greenwood, S., Perrin, A., Duggan, M.: Social media update 2016, `http://www.pewinternet.org/2016/11/11/social-media-update-2016/`
10. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: Themeriver: visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics 8(1), 9–20 (2002)
11. Huang, F., x. Wang, C., Long, J.: 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. pp. 732–737 (Nov 2011)
12. Inselberg, A., Dimsdale, B.: Parallel Coordinates for Visualizing Multi-Dimensional Geometry, pp. 25–44. Springer Japan, Tokyo (1987)
13. Jagarlamudi, J., Daumé III, H.: Extracting multilingual topics from unaligned comparable corpora. In: European Conference on Information Retrieval. pp. 444–456. Springer (2010)

14. Keim, E.D., Kohlhammer, J., Ellis, G.: Mastering the information age: Solving problems with visual analytics, eurographics association (2010)
15. Kucher, K., Kerren, A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In: 2015 IEEE Pacific Visualization Symposium (PacificVis). pp. 117–121 (2015)
16. Lenhart, A.: Teens, social media & technology overview 2015, http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/
17. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (Jun 2014)
18. Liu, S., Zhou, M.X., Pan, S., Qian, W., Cai, W., Lian, X.: Interactive, topic-based visual text summarization and analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 543–552. CIKM '09, ACM, New York, NY, USA (2009)
19. Matkovic, K., Freiler, W., Gracanin, D., Hauser, H.: Comvis: A coordinated multiple views system for prototyping new visualization technology. In: 12th International Conference Information Visualisation, IV08. pp. 215–220 (2008)
20. Micro Focus: How much data is created on the internet each day?, https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
21. Nielsen, J., Landauer, T.K.: 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. pp. 206–213. CHI '93, ACM, New York, NY, USA (1993)
22. OpenAire: What is the open research data pilot?, https://www.openaire.eu/opendatapilot
23. Pardos, Z.A., Kao, K.: Moocrp: An open-source analytics platform. In: 2nd ACM Conference on Learning at Scale, L@S 2015. pp. 103–110. Association for Computing Machinery, Inc (2015)
24. Pham, M.D., Boncz, P., Erling, O.: S3g2: A scalable structure-correlated social graph generator. In: Technology Conference on Performance Evaluation and Benchmarking. pp. 156–172. Springer (2012)
25. Pike, W.A., Stasko, J., Chang, R., O'connell, T.A.: The science of interaction. Information Visualization 8(4), 263–274 (2009)
26. Pokorny, J.: Nosql databases: a step to database scalability in web environment. International Journal of Web Information Systems 9(1), 69–82 (2013)
27. Prat, A., Sanchez, X.: Ldbc-snb data generator, https://github.com/ldbc/ldbc_snb_datagen
28. Pérez-Rosés, H., Sebé, F.: Synthetic generation of social network data with endorsements. Journal of Simulation 9(4), 279–286 (2015)
29. Ribarsky, W., Xiaoyu Wang, D., Dou, W.: Social media analytics for competitive advantage. Computers and Graphics (Pergamon) 38(1), 328–331 (2014)
30. Richards, M.: Software architecture patterns. O'Reilly Media (2015)
31. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages. pp. 336–. VL '96, IEEE Computer Society, Washington, DC, USA (1996)
32. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
33. WYRED: Requirements document wp3_d3.1 (2017), https://doi.org/10.5281/zenodo.292978

34. Yee, J., Mills, R.F., Peterson, G.L., Bartczak, S.E.: Automatic generation of social network data from electronic-mail communications. Report, DTIC Document (2005)
35. Zafarani, R., Liu, H.: Social computing data repository at ASU, `http://socialcomputing.asu.edu`