Métodos y técnicas de investigación social/Metodoak eta ikerketa sozialaren teknikak

Leila Chivite Matthews

REFLEXIÓN DEL MÉTODO: Google Analytics como herramienta de investigación social.
/
METODOAREN GOGOETA: Google Analytics ikerketa sozialeko tresna gisa.

TFG/GBL 2014

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Facultad de Ciencias Humanas y Sociales
Giza eta Gizarte Zientzien Fakultatea

Grado en Sociología Aplicada/ Soziologia Aplikatuan Gradua

Grado en Sociología Aplicada

Trabajo Fin de Grado

Gradu Bukaerako Lana

REFLEXIÓN DEL MÉTODO: Google Analytics como herramienta de investigación social.

METODOAREN GOGOETA: Google Analytics ikerketa sozialeko tresna gisa.

Leila Chivite Matthews

FACULTAD DE CIENCIAS HUMANAS Y SOCIALES
GIZA ETA GIZARTE ZIENTZIEN FAKULTATEA

UNIVERSIDAD PÚBLICA DE NAVARRA
NAFARROAKO UNIBERTSITATE PUBLIKOA

Estudiante / Ikaslea
Leila Chivite Matthews

Título / Izenburua
REFLEXIÓN DEL MÉTODO: Google Analytics como herramienta de investigación social./ Google Analytics ikerketa sozialeko tresna gisa.

Grado / Gradu
Grado en Sociología Aplicada/ Soziologia Aplikatuan Gradua

Centro / Ikastegia
Facultad de Ciencias Humanas y Sociales / Giza eta Gizarte Zientzien Fakultatea
Universidad Pública de Navarra / Nafarroako Unibertsitate Publikoa

Director-a / Zuzendaria
Andoni Iso Tinoco

Departamento / Saila
Departamento de Sociología/ Soziologia ko saila

Curso académico / Ikasturte akademikoa
2013/2014

Semestre / Seihilekoa
Primavera / Udaberrik

Resumen

Durante el presente estudio pretendo presentar la herramienta de seguimiento web Google Analytics, como una herramienta útil para la investigación social. Hasta el momento, los Datos Masivos han sido utilizados sobre todo por técnicos pertenecientes al campo de la investigación de mercados. La utilización de Big Data plantea un cambio de perspectiva en la investigación, a causa del volumen de datos que contienen y la continua actualización de los mismos. Google Analytics utiliza un sistema de cookies para obtener sus datos y el enfoque de estadística bayesiana para efectuar sus análisis. Este trabajo destapa el funcionamiento de esta herramienta y explica de qué modo podemos utilizarla y como debemos interpretar sus datos.

Palabras clave: Google Analytics; Investigación Social; Datos Masivos; reflexión del método; Investigación de datos on-line.

Abstract

In this study I try to present the Google Analytics tracking web tool, as a useful tool for social research. Up to the present time, Big Data has been used mainly by scientists from the field of market investigation. The use of Big Data proposes a change of perspective in the research, because Big Data has a large volume of data and updates data continuously. Google Analytics uses a cookie system to obtain data and Bayesian statistical approach to perform their analyzes. This work reveals how this tool works and explains how we can use it and shows some of the limitations in the interpretation of this data.

Keywords**:** Google Analytics; Social Research; Big Data; reflection on method; Data scraping.

ÍNDICE

1. INTRODUCCIÓN

Durante el presente trabajo trataré de situar la herramienta Google Analytics dentro del área de la investigación sociológica. Estudiaré esta tecnología como posible herramienta de investigación social on-line.

Mi trabajo trata ante todo de proporcionar las claves que nos permitan situarnos en el nuevo paradigma y debate en torno a la manipulación de los datos masivos, a través de un estudio de caso sobre la herramienta proporcionada por Google.

Para saber si esta herramienta puede sernos útil en el campo de la investigación social, efectuaré una evaluación de la misma comparándola con una investigación típica de datos secundarios. Se trata de si Google Analytics plasma los requisitos que toda investigación social debe cumplir.

2. FORMULACIÓN DEL PROBLEMA

En la actual era de la Información y las nuevas tecnologías de la comunicación, nos encontramos ante un nuevo paradigma de investigación: la utilización de los datos masivos o Big Data. No solo se trata de que cada día obtengamos más información, sino que además, la información crece cada día más deprisa. Las compañías de Internet han encontrado una oportunidad de investigación en esta nueva era; Google, por ejemplo, procesa aproximadamente 24 petabytes de datos al día. Un solo petabyte son 1.000.000 de gigabytes, en unidades más utilizadas se traduce en 1.000.000.000 de megabytes (Anexo A).

De esta manera la utilización de Big Data en el campo del Marketing se está presentando como una nueva tendencia en el análisis de mercados. La mayoría de los datos que contienen los Big Data provienen del rastro de navegación que dejamos en Internet. Es decir, se trata de la recolección de los datos sobre los hábitos de navegación en Internet, que sitios web visitamos, cuanto los visitamos, que tipo de páginas visitamos, etc.

Es importante remarcar que este proyecto estudia Big data como una fuente de información ya producida, o sea, las posibilidades de "data scraping" o la utilización de datos ya producidos y almacenados; y su posible uso en estudios sociológicos. Se ha de distinguir "data scraping" de herramientas que se pueden utilizar y desarrollar en internet para producir estudios primarios donde se utiliza la colección de datos (se implementa la metodología clásica en Internet, encuestas on-line por ejemplo) o estudios sociológicos utilizando Internet: la "metodología digital". o en otras palabras la colección de data para estudios sociológicos utilizando internet, el internet como vehículo de estudios sociológicos.

Por supuesto, la utilización de datos masivos en investigación no se encuentra libre de controversias. Unos de los grandes problemas de la utilización de Big Data se refiere al tamaño y desorden de sus bases de datos, para poder ejecutar un análisis sobre una base de Big data se necesita emplear una cantidad de tiempo importante

para depurar los datos. Y lo más importante, lo que ha llevado a un debate entre las diferentes disciplinas que se disputan esta cantera de información por investigar: con la utilización de datos masivos no podemos ejercer un control sobre la muestra, por lo que hay quienes dicen que los análisis que podamos obtener de ellos no resultan fiables.

Google Analytics, la herramienta que se estudia en este trabajo, es una de las herramientas más conocidas a nuestra disposición que utiliza datos digitales de origen y que sistematiza su propia base de Big Data. Esta herramienta contabiliza los datos a tiempo real y genera informes de seguimiento.

Presentando así esta herramienta, parece todo solucionado, pero ante todo, hay que ser consciente de que el valor de la información no reside en los datos en sí, sino en la forma de correlacionarlos para descubrir vínculos y patrones entre ellos. Hasta el momento, esta herramienta y otras de seguimiento web han sido utilizadas con fines comerciales, yo intentare indagar en la posibilidad de utilizar Google Analytics como una herramienta de investigación social on-line.

3. ANTECEDENTES

Después de realizar una búsqueda exhaustiva en Internet, no he encontrado antecedentes de investigaciones parecidas a mi estudio.

He recopilado varios estudios sobre la utilización de Google Analytics para optimizar el rendimiento de varios sitios web, desde páginas de comercio electrónico hasta bibliotecas. En todos los casos, salvo en el de las bibliotecas, utilizan esta herramienta para maximizar los beneficios de su empresa. La totalidad de estos estudios concluye que la herramienta les ha proporcionado los informes y datos clave para optimizar su negocio u organización.

Además de este tipo de estudios, he encontrado un único trabajo dedicado a la elaboración de un nuevo tipo de metodología utilizando Google Analytics como herramienta, con el objetivo de avanzar en el campo del análisis de Big Data orientado con fines comerciales y de audiencias. El trabajo plantea la introducción de esta herramienta en el campo del estudio de mercados, no presenta ninguna conclusión interesante, ya que solo apunta que esta tecnología como otras de seguimiento web ofrece informes y datos útiles para la optimización del rendimiento de un sitio web; y que reduciría costes de investigación empresarial porque se trata de una herramienta gratuita.

Estudios donde se ha utilizado Google Analytics como herramienta de seguimiento web con fines comerciales y de audiencias:

- Google Analytics for measuring website performance**.** (Plaza. B. 2011)
- Using Google Analytics for Improving. Library Website Content and Design: A Case Study. (Fang. W. 2007)
- Using Google Analytics to Evaluate the Usability of E-Commerce Sites (Hasan. L; Morris. A, Probets. S. 2009).
- Website Statistics 2.0: Using Google Analytics to Measure Library Website Effectiveness. (Steven J. Turner M. 2010).

- Using Google Analytics as a process evaluation method for Internet-delivered interventions: an example on sexual health (Crutzen.R.; Roosjen. J.; Poelman. J. 2012).

Estudio donde se utiliza Google Analytics en pro del avance científico-social:

- Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series (Plaza. B. 2009).

4. OBJETIVOS

General→ Evaluar y comprobar en qué medida Google Analytics es una herramienta útil para la investigación social.

Específicos→

- Describir la herramienta Google Analytics.
- Averiguar y describir de dónde provienen los datos de Google Analytics.
- Comprobar si Google Analytics cumple como herramienta los requisitos de toda investigación social.
- Comparar un tipo de investigación social con lo que nos ofrece Google Analytics como herramienta de investigación.

5. PREGUNTAS DE INVESTIGACIÓN

- ¿Google Analytics es útil para la investigación social? o ¿solo puede utilizarse en investigación de mercados?
- ¿De dónde provienen los datos de Google Analytics?
- ¿Cumple Google Analytics los requisitos de toda investigación social?

6. MARCO TEÓRICO

Para fundamentar mi estudio, he realizado una revisión bibliográfica acerca de cómo se debe medir en investigación social y qué propiedades hacen una buena metodología de investigación sociológica; dónde nos encontramos, que nuevo paradigma de investigación se abre ante nosotros: El Big Data; y qué es Google Analytics.

6.1 La medida en investigación social

Para poder entender el propósito de mi estudio, tengo que remitirme a la teoría de construcción de la medida en ciencias sociales, y en concreto la medida en la investigación sociológica. Para ello me baso en la teoría presentada por Aaron Cicourel en su libro "Método y medida en Sociología".

El autor apunta que en investigación social, así como en cualquier otra ciencia, las técnicas utilizadas son sujeto de investigación de la Sociología del Conocimiento. Los métodos que utilizamos para investigar están en continua construcción, transformación y perfeccionamiento. Por tanto, el conocimiento actual, depende del estado del método. Este trabajo se respalda en esta evidencia, ya que lo que planteo es la utilización de una nueva herramienta y su ensamblaje en la investigación social para la recogida y proyección de nuevos conocimientos.

Según Coombs y Torgerson[1] (Coombs. C 1953), (Torgerson. W 1958) los fenómenos sociológicos interesantes para nuestros colegas, pueden medirse, de igual forma que podemos medir las propiedades físicas de un objeto. Por tanto los acontecimientos sociales pueden codificarse para su posterior análisis. Para evaluar con precisión un proceso social es necesario, en primer lugar, estudiar el significado de la vida cotidiana acerca de dicho proceso. Para ello el investigador debe conocer el

[1] Citado por Cicourel. A. (2011). Método y medida en Sociología.CIS. Pág. 57. Madrid.

lenguaje y manejar con soltura los significados, del lenguaje verbal y no verbal, que se utilizan en la vida cotidiana del fenómeno que se esté estudiando; para obtener unos resultados lo más precisos posibles. La medida en sociología, está sujeta al lenguaje, a la cultura que se estudia.

En la investigación social, comenzamos el razonamiento de nuestros estudios por los significantes y significados culturales que hemos seleccionado y descifrado; algo evidente, que al entrar en la realidad de Internet se vuelve difuso (Hine. C. 2004). Nuestro manejo del lenguaje, de los significados sociales es complicado, ya que el lenguaje se encuentra en continua transformación, y varía según la cultura que se esté estudiando.

Este razonamiento, fundamenta el hecho de nuestro desconocimiento sobre el lenguaje utilizado en Internet, tanto el lenguaje social como el conjunto de señales y códigos utilizados por las máquinas. En el caso de los lenguajes sociales compartidos en Internet, tenemos el famoso ejemplo del "XD", utilizado por la mayoría de comunidades virtuales para expresar la risa en una conversación, simboliza una cara riéndose. El lenguaje cibernético de Internet también impedimenta la construcción de una buena metodología de investigación social on-line, el sistema de cookies es uno de los muchos ejemplos que podemos encontrar en la red, un ejemplo que expongo más adelante en el apartado de análisis.

Para poder estudiar cualquier fenómeno sociológico debemos manejar su lenguaje, y tenemos que estudiarlo desde el interior. Con los fenómenos ocurridos en Internet ocurre de igual manera.

> *"Los sociólogos deben actuar desde dentro de la sociedad, empleando su lenguaje nativo y sus muchos significados culturales indefinidos. Adquirir el punto de vista de dentro significa aprender o asumir la cultura común nativa".*(Cicourel. A. 2011. 67)

Por otro lado, Cicourel señala las principales propiedades que toda medida en investigación social debe cumplir; que no son otras que las características recogidas

por Lazarsfel y Barton.[2] (Lazarsfel. P y Barton. A 1951). Propiedades que corresponden con las leyes de la equivalencia lógica (Cicourel.A. 2011. 69):

1- Toda relación social a estudiar, debe estar sujeta a la equivalencia lógica sin que se falsee su significado original.
2- Hay que tener en cuenta que, la categorización utilizada para clasificar las propiedades empíricas de los actores sociales generan valores limitados asumibles. Las propiedades de la equivalencia lógica se imponen a los conceptos y definen el valor falso o verdadero de los datos obtenidos.
3- Las tres leyes de equivalencia deben cumplirse para asegurarse de que podemos hacer cálculos acerca de los fenómenos sociales (reflexividad, transitividad, simetría).
4- Teniendo en cuenta las tres leyes de equivalencia no debemos olvidar que el contexto espacio-temporal afecta a los fenómenos sociológicos. Un hecho social X probablemente no sea igual si cambiamos la escena, el medio y el tiempo social.
5- Las categorías que obtenemos a través de los datos siempre deben adquirir su importancia y significado a través de fundamentos teóricos y metodológicos. Los datos en sí mismos carecen de significado.
6- No podemos perder de vista la terminología utilizada para medir y transmitir datos, el lenguaje utilizado para medir puede llegar a cosificar y falsear caprichosamente los datos.
7- Debemos cuidarnos de no crear escalas de medida que supongan relaciones lógicas que no se corresponden con la realidad social.
8- Nunca debemos crear o trasformar la medida para que cumpla las leyes de equivalencia lógica, y obtener los resultados que presuponemos de antes de ejecutar la recogida de información.

Por último, el autor apunta que el progreso científico depende del cuestionamiento de nuestras teorías, instrumentos y técnicas de medida. Si comenzamos a tomar la metodología en investigación, como algo puro y perfecto

[2] Citado por Cicourel. A. (2011). Método y medida en Sociología.CIS. Pág. 69. Madrid.

acabaremos creando ideología, en vez de ciencia. Para poder avanzar en el nuevo paradigma de investigación social en Internet deberemos cuestionar, transformar, adaptar y tal vez crear, nuevas técnicas y herramientas para la medida de datos sobre los fenómenos acontecidos en la World Wide Web.

6.2 El auge de los datos masivos: la nueva realidad analítica de los Big Data

Una vez que los ordenadores llegaron a la mayoría de la sociedad occidental y se generalizó la utilización de Internet, los datos comenzaron a acumularse hasta llegar al nuevo paradigma actual de los datos masivos en Internet. Aunque no existe ninguna definición concreta acerca de los datos masivos o Big Data, entendemos que los datos masivos provienen del siguiente contexto:

> *"En un principio, la idea era que el volumen de información había aumentado tanto que la que se examinaba ya no cabía en la memoria que los ordenadores emplean para procesarla, por lo que los ingenieros necesitaban modernizar las herramientas para poder analizarla. Ese es el origen de las nuevas tecnologías de procesamiento, como MapReduce, de Google, y su equivalente de código abierto, Hadoop, que surgió de Yahoo. Con ellos se pueden manejar cantidades de datos mucho mayores que antes, y esos datos no precisan ser dispuestos en filas ordenadas ni en las clásicas tabulaciones de una base de datos"*.(Mayer. V. y Niel. K. 2013. 17).

Los Big data, están referidos a los análisis a nivel macro, no a investigaciones micro. Estos nuevos tipos de conocimiento y perspectiva proporcionan nuevas percepciones sobre los datos, crean tipos de valor hasta ahora desconocidos y generan transformaciones en los mercados, organizaciones, gobiernos y relaciones sociales

A lo largo de estos años almacenando datos, ha habido muchos intentos de determinar la cantidad exacta de datos digitales que disponemos. Uno de los estudios

que más se ha acercado a dicha cifra es el elaborado por Martin Hilbert, de la Annenberg School de comunicación y periodismo de la universidad del Sur de California. Hilbert apunta que ya en 2007 existían más de 300 exabytes de datos digitales (un exabyte acumula mil millones de gigabytes), lo más curioso es que sólo el 7% de los datos eran analógicos (Mayer. V. y Niel. K. 2013).

Ya en los años sesenta se utilizaban los términos "revolución de la información" y "era digital", pero es ahora cuando podemos ver la autentica revolución digital. Los datos digitales se propagan rápido, tanto que la última cifra de almacenamiento de datos digitales en 2013 es de 1200 exabytes, una cantidad que apenas llegábamos a imaginar antes de la aparición de Internet.

Según Victor Mayer autor de *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* la sociedad deberá adaptarse a este nuevo paradigma y comenzar a analizar las correlaciones entre los datos en vez de elaborar investigaciones buscando la causalidad: centrarse en el qué y no el por qué.

El autor explica que para poder explotar esta nueva realidad digital debemos cambiar de perspectiva en tres aspectos: el primero es aceptar el hecho de que el muestreo fue producto de una realidad en la que los datos escaseaban y que ahora pierde parte de su sentido, el segundo es que debido a que prácticamente disponemos de toda la población tenemos que desprendernos de nuestro anhelo por la exactitud y asumir más errores en la medición de los datos, y por último tenemos que adaptarnos a las nuevas y desordenadas bases de datos hasta el momento desconocidas.

> *"Lo que perdemos en exactitud en el nivel micro, lo ganamos en percepción en el nivel macro".*(Mayer. V. y Niel. K. 2013. 26).

Mayer insiste en que el análisis a través de correlaciones es el futuro del análisis de los datos masivos, y que tenemos en el caso de los Big Data, olvidarnos de averiguar el porqué y centrarnos en qué nos dicen los datos. Además apunta que el muestreo aleatorio, uno de los métodos más utilizados hasta el momento, no se encuentra libre de errores y inexactitudes en su ejecución:

"De forma aún más preocupante, el muestreo aleatorio no resulta sencillo de extrapolar para incluir subcategorías, por lo que al parcelar los resultados en subgrupos cada vez menores aumenta la posibilidad de llegar a predicciones erróneas".(Mayer. V. y Niel. K. 2013. 39).

Este autor también apunta que al emplear la totalidad de los datos podemos señalar conexiones entre los mismos y detalles que con el muestreo no logramos.

"Igualmente, ya que los datos masivos se basan en toda la información, o por lo menos en toda la posible, nos permiten examinar detalles o explorar nuevos análisis sin correr el riesgo de que se vuelvan borrosos. Podemos someter a prueba nuevas hipótesis a muchos niveles distintos de granularidad".(Mayer. V. y Niel. K. 2013. 46).

Por supuesto con ello no quiere decir que el muestreo quede obsoleto en la actualidad, y apunta que ese tipo de metodología sigue siendo muy útil en otro tipo de escenarios.

Otra de las ventajas que apunta Víctor Mayer es el valor añadido que nos ofrecen los datos masivos, un valor que se encuentra presente en la reutilización de los mismos. Poder reutilizar los datos nos ofrece una series de posibilidades de investigación y además abaratamos costes al no tener que realizar encuestas. La mayoría de las grandes compañías se están uniendo al entusiasmo por el análisis de datos online y están dejando de hacer análisis offline. Además están comenzando a adquirir los conocimientos y especialistas necesarios para analizar lo que ahora conocemos como "los desechos de datos". Los desechos de datos se producen a través de cada "click", representan el rastro de navegación que dejamos los usuarios de Internet. Estos datos pueden indicarnos las preferencias o rechazos y todo tipo de patrones de conducta producidos en Internet.

Así pues, no nos encontramos ante una disputa entre disciplinas, no se trata de que el análisis de Big Data desbanque a la investigación estadística tradicional, sino que, simplemente, se está vislumbrando un avance en las técnicas del macro análisis estadístico. Disponemos de bases de datos tan inmensas que no podemos utilizar

antiguas técnicas, debemos modificar nuestra perspectiva, abrir nuestra mirada y aprovechar la oportunidad que se nos brinda.

6.3 Estadística bayesiana y estadística clásica

A lo largo de la última década se ha generado gran interés acerca de la estadística bayesiana, basada en el Teorema de Bayes, como alternativa a la clásica para el análisis de enormes cantidades de datos. Razón que me llevó a interesarme por este tipo de estadística a causa del volumen de datos que contienen los Big Data.

Este tipo de estadística presenta la idea de que todo tipo de incertidumbre puede describirse a través de la probabilidad, y de que la probabilidad es el lenguaje que debe utilizarse para describir la lógica que se encuentra tras la incertidumbre.

La estadística bayesiana utiliza una distribución inicial o a priori para su primer análisis, y después va añadiendo todos los datos acerca del fenómeno a estudiar, generando así una distribución a posteriori, ésta última, la "final", es la que se utiliza para ejecutar las inferencias y de donde se obtienen los estimadores óptimos de los parámetros estudiados (O'Hagan y Forster, 2004). Se trata entonces, de bases de datos que se actualizan constantemente, utilizan toda la información disponible para medir y analizar los parámetros. Un método que encaja a la perfección con el nuevo paradigma de Big Data.

La estadística clásica o inferencial, trabaja con muestras de población, un muestreo de la población completamente al azar que será estadísticamente representativo de la población entera bajo estudio. Este tipo de estadística, utilizada para el contraste de hipótesis, primero asigna un valor "P" (poblacional) y seguidamente, asigna un valor mínimo de significancia para poder aceptar o rechazar la hipótesis nula. Normalmente el nivel de significación es de 0,05, si la inferencia supera dicho valor rechazamos la hipótesis nula y si no se acepta. El caso es que, rechazar o aceptar dicha hipótesis en base al valor de "p" , tiene una relación estrecha

con el tamaño de la muestra, por lo que si un investigador estudia una muestra pequeña es posible que no pueda obtener conclusiones válidas porque supere el nivel de significación, y sugiera que tal vez con una muestra mayor hubiera obtenido conclusiones empíricas.

La estadística tradicional se ha basado en muestreos porque raramente en un estudio s tiene la capacidad de obtener la información completa de toda la población de interés o en otras palabras un Censo de la población de interés. El nuevo paradigma de Big data plantea que ahora en muchas circunstancias y para poblaciones bien definidas, si se tiene la información de toda la población por lo cual no se necesita estadística, los números que ofrecen describen la realidad tal y como es. Esta propiedad en Big Data solo se puede asumir en casos donde la población bajo estudio está bien definida pues aquellos que utilizan el Internet representan una sub-sección de la población.

También trabajar estadística clásica y con muestras poblacionales grandes presentarían siempre resultados significativos. El nivel de significación pierde relevancia cuando el tamaño maestral es muy grande (cuando disponemos de todos los datos, como con el caso de Big Data), ya que cualquier diferencia que se detecte en la muestra nos lleva a rechazar la hipótesis nula (Ayçaguera y Benavides, 2003). De esta forma, la estadística bayesiana encaja como técnica, puesto que utiliza todos los datos a su disposición, y el enfoque es de fácil implementación informática.

Con estadística estimamos los parámetros de las poblaciones que estudiamos, con Big data en muchos casos observamos los parámetros directamente. Tanto la estadística tradicional como las nuevas metodologías que se están desarrollando para estudiar Big data sufren del las mismas preguntas y limitaciones cuando se trata de estudiar relaciones causales. Estamos pues en el momento del desarrollo de metodologías de análisis de Big data, la mayor parte del análisis que encontramos es muy básico, limitándose a la descripción y correlación de datos, por lo que es necesario complementar un tipo de estadística con la otra: utilizar el software con enfoque bayesiano para describir el fenómeno y mirar relaciones, y responder a las

preguntas e hipótesis de nuestra investigación con el enfoque de estadística clásica, por ejemplo con experimentación o utilizando metodología cualitativa.

6.4 Google Analytics una poderosa herramienta de seguimiento web

La herramienta de seguimiento web que se presenta en este trabajo, Google Analytics, es un servicio gratuito de la empresa Google que ofrece estadísticas de páginas web. Hasta el momento una herramienta utilizada por especialistas pertenecientes al campo del análisis de audiencias y mercados.

Este producto comenzó su desarrollo en 2005 y su impulsado gracias a la compra, por parte de Google, de Urchin, la mayor compañía de seguimiento web estadístico hasta ese momento.

Su tecnología funciona de la siguiente manera: Google añade un código JavaScript[3] a la página de la que se desea obtener los datos. Este código procede a la carga de archivos a los servidores de Google y monitoriza la web enviando y almacenando los datos a un servidor Google. Una vez almacenados los datos, la herramienta Google Analytics genera una series de informes analíticos descriptivos. Además de ofrecer una extensa analítica descriptiva sobre la web seleccionada, Google Analytics cuenta con una interfaz sencilla con gráficos implementados por Adobe Flash[4], lo que convierte esta herramienta en un producto que puede utilizar casi cualquier hardware y software sin problemas, ya que no requiere de una gran potencia tecnológica.

Los programas de analítica web se dividen en dos categorías que muestro a continuación:

[3] Lenguaje desarrollado por Sun Microsystems en conjunto con Netscape; aunque es parecido a Java se diferencia de él en que los programas están incorporados en el archivo HTML. (http://www.internetglosario.com/letra-j.html).

[4] Adobe Systems Incorporated es una empresa de software, fundada en 1982 por John Warnock y Charles Geschke cuando salieron de Xerox Parc. Son los creadores de PDF, y de programas como Photoshop, Illustrator, Acrobat, entre otros. (http://www.internetglosario.com/letra-s.html).

- Basados en el seguimiento de etiquetas (Este tipo de seguimiento es el que utiliza Google Analytics)

  Este tipo de análisis se efectúa insertando una etiqueta dentro de un código fuente[5] de la página que queremos analizar. Este código se encarga de comunicar los datos al servidor donde se almacenan y posteriormente se generan análisis descriptivos sobre los mismos.

- Basados en el servidor

  Este tipo de seguimiento web ofrece el mismo tipo de resultado pero para ello es necesario instalar un software y adquirir nuestro propio servidor. Normalmente cuando se contrata un servicio de hosting[6] suele incluir su propio software de analítica web.

Google Analytics nos ofrece una gran cantidad de datos, he podido clasificarlos de la siguiente manera:

- Tráfico Global versus Tráfico por Página
- Términos de Búsqueda utilizados
- Origen del tráfico (ver si procede de buscadores, de entrada directa, de enlaces compartidos, de blogs, etc)
- Segmentación Geográfica del Visitante
- Tiempo de Visitas
- Acciones por Visita
- Rebotes (cuanto menos porcentaje mejor)
- Tráfico por Navegador
- Tráfico por S.O (sistema operativo)
- Control de Conversiones
- Informe Actual versus Informe Histórico

[5] Conjunto de instrucciones que componen un programa, escrito en cualquier lenguaje. En inglés se dice "source code". (http://www.internetglosario.com/letra-s.html).

[6] El servicio de Web Hosting consiste en el almacenamiento de datos, aplicaciones o información dentro de servidores diseñados para llevar a cabo esta tarea. (http://www.internetglosario.com/letra-s.html

De esta manera, Google Analytics se ha convertido en el producto preferido de la mayoría de responsables de campaña de posicionamiento SEO[7] (que tu página aparezca de las primeras en un motor de búsqueda). ¿Puede convertirse esta herramienta en una de las favoritas de análisis de datos secundarios on-line en el campo de la sociología?

[7] En inglés Search Engine Optimization. Optimización para Motores de Búsqueda. El término se usa para describir la técnica de mercadeo de preparar un website para mejorar sus oportunidades de colocarse en las primeras posiciones de un motor de búsqueda (search engine) cuando se busca información relevante. (http://www.internetglosario.com/letra-s.html).

7. METODOLOGÍA

El objetivo de este estudio es evaluar y comprobar en qué medida Google Analytics puede servir como herramienta para la investigación social. Para ello he seguido una serie de pasos para realizar una puesta a punto de la herramienta:

1. He realizado una descripción de la herramienta: sobre qué datos proporciona y para qué ha sido creada.
2. He explorado y expuesto en mi estudio el origen de los datos que ofrece Google Analytics.
3. He comprobado si las medidas que ofrece esta herramienta cumplen las propiedades de equivalencia lógica.
4. He indagado sobre los posibles usos de Google Analytics más allá de la investigación de mercados y la posibilidad de realizar análisis estadístico.

La siguiente tabla muestra como he planteado mi investigación y que técnicas he utilizado:

OBJETIVO	TÉCNICA	DESCRIPCIÓN
Descripción de Google Analytics	• Observación de la herramienta	Análisis y seguimiento de la recogida de datos e informes que genera Google Analytics, durante un período de 30 días. (*1*)
	• Documentación	Lectura y emisión de un informe sobre la descripción de la herramienta publicada por la empresa Google.
Desvelar el origen de los datos	• Observación de la herramienta	(*1*)
	• Reuniones con un Ingeniero Informático	Para poder familiarizarme con el lenguaje informático que utiliza la herramienta, y el origen de los datos, he utilizado uno de mis contactos: Un Ingeniero Informático, experto en programación familiarizado con esta herramienta y su

		funcionamiento me ha ayudado lo largo de mi estudio siempre que se me presentaba alguna dificultad con el lenguaje que utiliza Google Analytics.
Comprobación de la validez de las medidas utilizadas por Google Analytics	• Observación de la herramienta	(*1*)
	• Comparación de las medidas utilizadas con las medidas utilizadas habitualmente en investigación social.	He descrito cada medida utilizada en Google Analytics y he comprobado que cumple con las propiedades de la equivalencia lógica.
Exploración sobre los posibles usos de Google Analytics en investigación social	• Observación de la herramienta	(*1*) He expuesto un pequeño ejemplo para ver alguna de las posibilidades de análisis que ofrece Google Analytics.

Cronograma

EJE TEMPORAL						
TAREA	**ESPECIFICACIÓN**	FEBRERO	MARZO	ABRIL	MAYO	JUNIO
Planteamiento del problema	Revisión bibliográfica para la construcción del marco teórico.					
Diseño de la metodología	Planteamiento y organización de mi estudio.					
Trabajo de campo	Observación de la herramienta Google Analytics. Reuniones con un Ingeniero Informático					
Análisis de datos	Revisión y descripción de las medidas utilizadas por la herramienta y el origen de los datos. Comprobación de la validez de las medidas de Google Analytics.					
Redacción del informe	Elaboración y revisión del informe final.					
Preparación para la defensa ante tribunal	Preparación de un discurso para la presentación del estudio ante Tribunal.					
Nº TOTAL DE HORAS repartidas en 5 meses = 150						

8. ANALISIS DE LA HERRAMIENTA DE SEGUIMIENTO WEB GOOGLE ANALYTICS

El análisis que presento a continuación consta de varios apartados que responden a las siguientes preguntas: ¿Qué es Google Analytics?, ¿Para qué sirve?, ¿Qué datos e informes proporciona?, ¿De dónde proviene la información que recoge esta herramienta? y ¿Cumplen los análisis de esta herramienta con el principio de equivalencia lógica?

Además he añadido una propuesta de cómo podríamos utilizar Google Analytics en el campo de la investigación social.

8.1 ¿QUÉ ES GOOGLE ANALYTICS Y PARA QUÉ SIRVE?

Google Analytics es una herramienta de seguimiento web gratuita que ofrece informes estadísticos sobre el sitio web que seleccionemos. Normalmente, el seguimiento web ha sido utilizado por técnicos y especialistas en la investigación de mercados y audiencias.

Habitualmente se utilizan los datos e informes proporcionados por Google para maximizar el rendimiento de un sitio web, y poder así maximizar a su vez los beneficios de la empresa.

8.2 ¿QUÉ DATOS E INFORMES NOS PROPORCIONA GOOGLE ANALYTICS?

Para comenzar ordenaré las métricas que ofrece Google Analytics. En primer lugar, la herramienta divide sus métricas principales en tres apartados de datos: *Público, Adquisición y Comportamiento*. Los datos sobre *Público* nos proporcionan la información sobre los usuarios que visitan la página web, sus datos demográficos y sus intereses. El apartado de *Adquisición* contiene los datos acerca de dónde provienen las visitas y el apartado de *Comportamiento* reúne los datos sobre el contenido de la página web que ha visitado el usuario.

Todos los informes y métricas que proporciona Google Analytics se ajustan a tres periodos temporales predeterminados: el último día, semana o mes. Pero también podemos ajustar nosotros mismos un periodo de tiempo y pedir todas las métricas e informes del periodo que propongamos.

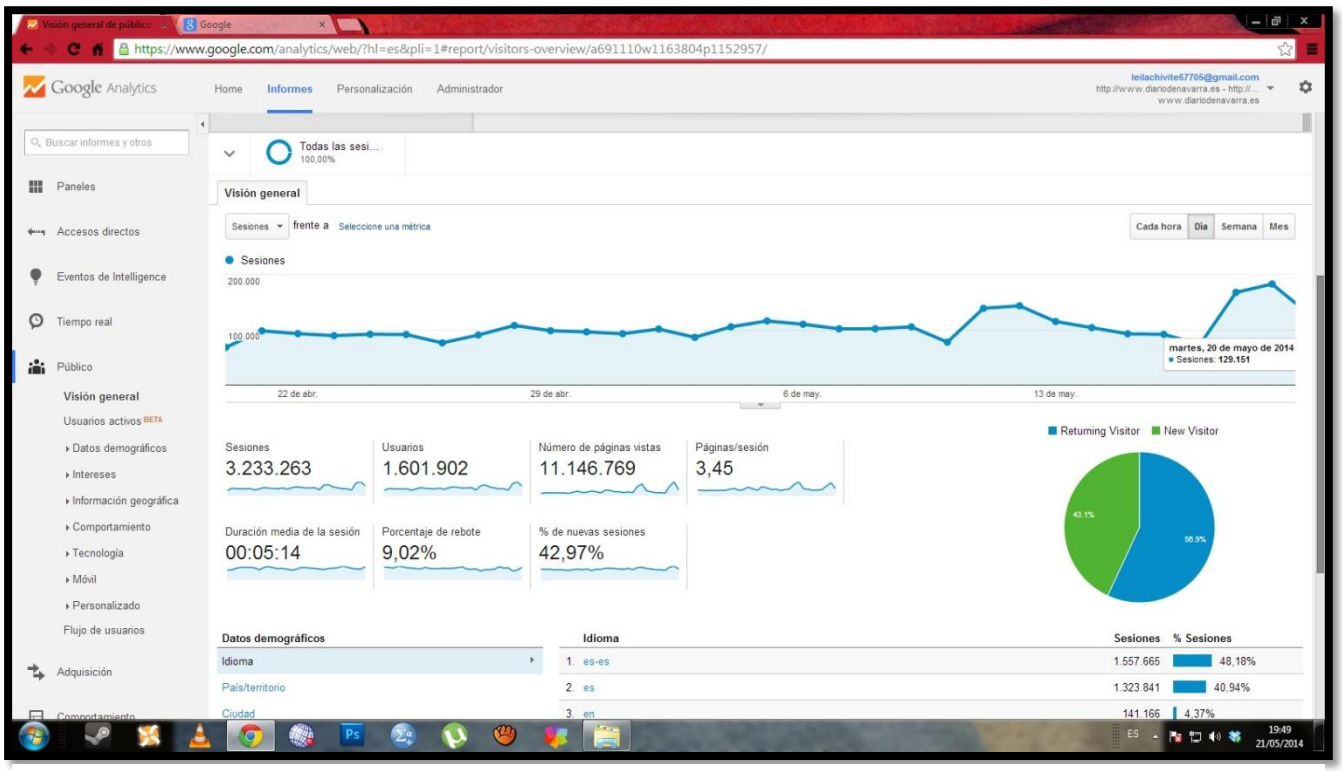

Figura1. Visión general de Google Analytics y sus métricas principales

En el presente análisis he repasado cada apartado para poder ofrecer una descripción de qué tipo de datos ofrecen y que significan; pero para organizar mejor las métricas, en primer lugar he decidido describir las métricas trasversales a los tres apartados:

- **Número de usuarios:** A lo largo de la observación de la herramienta y gracias al experto en informática que me ha asesorado, sé que el número de visitantes se refiere a personas y no ha máquinas. Es decir, se trata de el número de visitas individuales real, no se trata de ordenadores que envían señales, es necesario que un ser humano esté navegando por Internet para que Google lo detecte a través del sistema de Cookies y contabilice el dato.

- **Duración media de visita:** Esta métrica se refiere a cuánto tiempo ha estado navegando una persona en el sitio web que analiza la herramienta. Este dato

no es del todo fiable, porque el usuario del sitio web ha podido estar activo en los primero segundos dentro de la página web, y luego irse y dejar el ordenador encendido sin cerrar la página. Este tipo de comportamiento altera las estadísticas. Según el experto en informática, si queremos utilizar esta métrica en una investigación, lo preferible sería atender solo a los datos que se encuentran entre 0 y 5 minutos. Esto depende, claro, del tipo de página web que se esté analizando, no se da el mismo comportamiento en una web de consumo de aparatos electrónicos, que en una web de un medio de comunicación.

- **Porcentaje de rebote:** Este tipo de dato representa la pérdida de público. Cuanto mayor es el porcentaje, mas personas han dejado de consultar la web. Representa la cantidad de usuarios del sitio web que han entrado e inmediatamente después han abandonado la página, es decir, han visto la página y no han interactuado en ella.

- **Visitantes nuevos y visitantes recurrentes:** Esta métrica nos ayuda a conocer la cantidad de población que llega por primera vez a nuestra web, y cuanto público vuelve a visitar la página. Google Analytics recuerda a cada usuario de la web que analiza, lo que significa que cuando un usuario ya reconocido por el programa vuelve a entrar, también recopila datos nuevos de este usuario. Se trata de una base de datos que está en continua actualización. La métrica de visitantes recurrentes siempre debería representar un valor numérico mayor que la de visitantes nuevos.

MÉTRICAS E INFORMES DEL APARTADO *PÚBLICO*

Sobre este apartado he descrito varios parámetros, pero en mi análisis sobre el mismo me he centrado en los *datos demográficos*, ya que para la investigación social, podrían representar el parámetro más útil en este caso:

- **País de origen, idioma y ciudad**: Google a través de los dominios y del sistema cookies, puede determinar de dónde proviene la visita. Se trata de los datos geográficos. Estos datos pueden sernos útiles para reconocer la procedencia de la población a estudiar.

- **Flujo de usuarios:** se trata de un gráfico de sistema de flujos que nos ayuda a comprender el comportamiento de los usuarios de la página web. Podemos en qué momento han dejado la web, en que apartado de la página, cuantos han abandonado la web, y por ejemplo, cuantos usuarios han visitado la página completa en todos sus apartados. Para la investigación de mercados es muy útil, ya que pueden rediseñar la web, detectando a través de este informe qué áreas del sitio web son menos atractivas o interesantes para su clientela. Para nosotros, los investigadores sociales, no puede proporcionar otro tipo de información. Supongamos que estamos analizando el comportamiento web dentro de un blog sobre política nacional, podríamos saber que secciones del blog interesan más a la población y que temáticas son menos visitadas, lo que nos podría ayudar a describir un escenario de opinión pública.

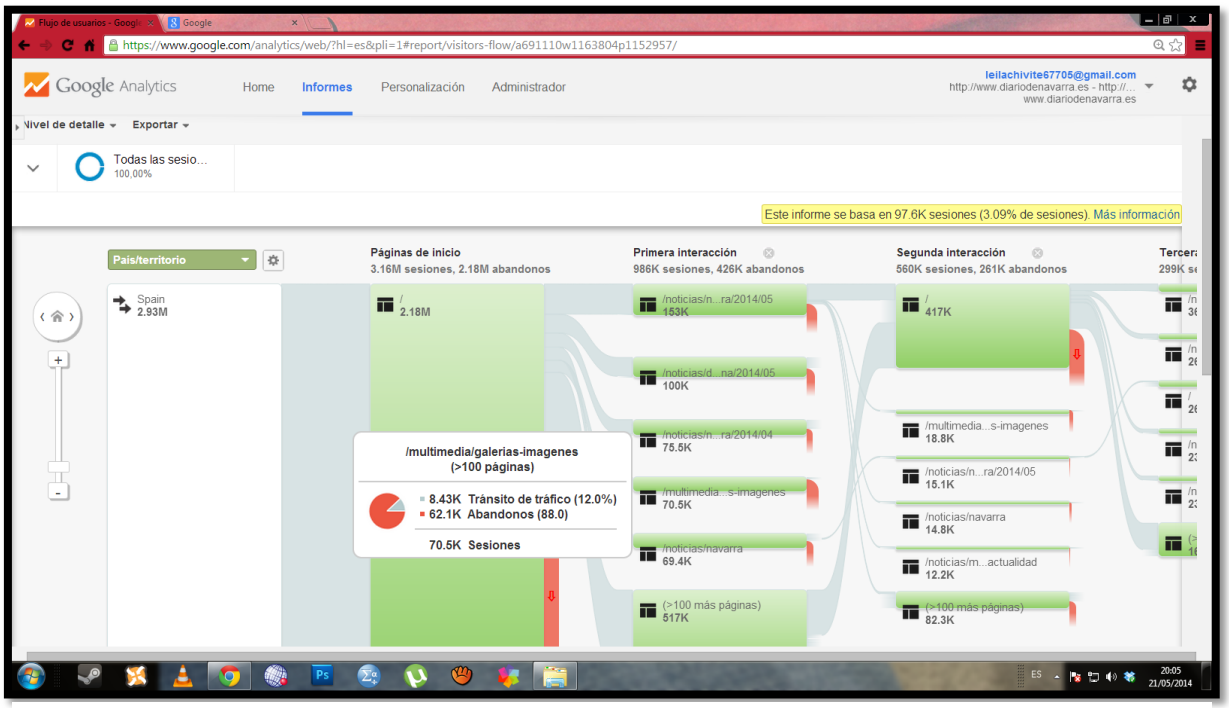

Figura 2. Gráfico de flujo de usuarios

- **Datos sobre el navegador y el dispositivo:** Google Analytics también nos ofrece información sobre qué sistema de navegación (Chrome, Internet Explorer, Safari) está utilizando el usuario y que tecnología (Movil, Tablet, PC), incluyendo la marca del dispositivo.

- **Datos demográficos:** Gracias a la ayuda prestada por el ingeniero informático con el que me he reunido he llegado a entender como Google puede saber el sexo o la franja de edad a la que pertenece un usuario de una página web. Se trata de algo muy sencillo, interesante para la investigación on-line, y al mismo tiempo aterrador. El sistema de cookies que utiliza Google (y otras empresas), recoge toda la información sobré que visitamos en internet, Google a través de las cookies conoce el contenido de lo que visitamos en internet y lo utiliza para reconocernos como hombres o mujeres, adultos o niños, etc. Esto significa que si por ejemplo yo navego por páginas "determinadas como femeninas", que pueden ir desde blogs sobre la compra de zapatos de tacón, hasta páginas referidas a la fisiología femenina, Google determinará que soy una mujer. La lógica de este sistema es acumulativa, si de repente comenzara a navegar por "páginas masculinas", Google rectificaría y computaría mi dato como masculino. Esta información es clave. No podemos interpretar estos datos demográficos como válidos al cien por cien, porque un dato obtenido en función de la navegación del usuario no siempre va a situar al individuo con una edad y sexo correctos. Pero sí podemos aceptarlos y utilizarlos, ya que el margen de error es mínimo y contamos con todos los datos. Esta métrica también nos ofrece los intereses generales de la población, o sea, la descripción de la navegación de los usuarios que visitan nuestra página. Estos datos presentan una peculiaridad muy interesante. A través de la observación que he realizado de la herramienta, encontré una serie de datos en este apartado etiquetados como "not provided". Tenía X hombres, X mujeres y X "not provided". Al consultar con el informático, este me explicó que Google anunció que protegería los datos obtenidos a través de las cookies de los usuarios de su navegador para proteger así la privacidad de sus "clientes". Una razón más para

tener cuidado con la interpretación de esta métrica. Teniendo en cuenta que a lo largo de mi análisis he podido ver que la mayoría de los visitantes de la web que utilizaban un buscador, utilizaban el buscador de Google, significaría que los datos demográficos se presentan un tanto escasos.

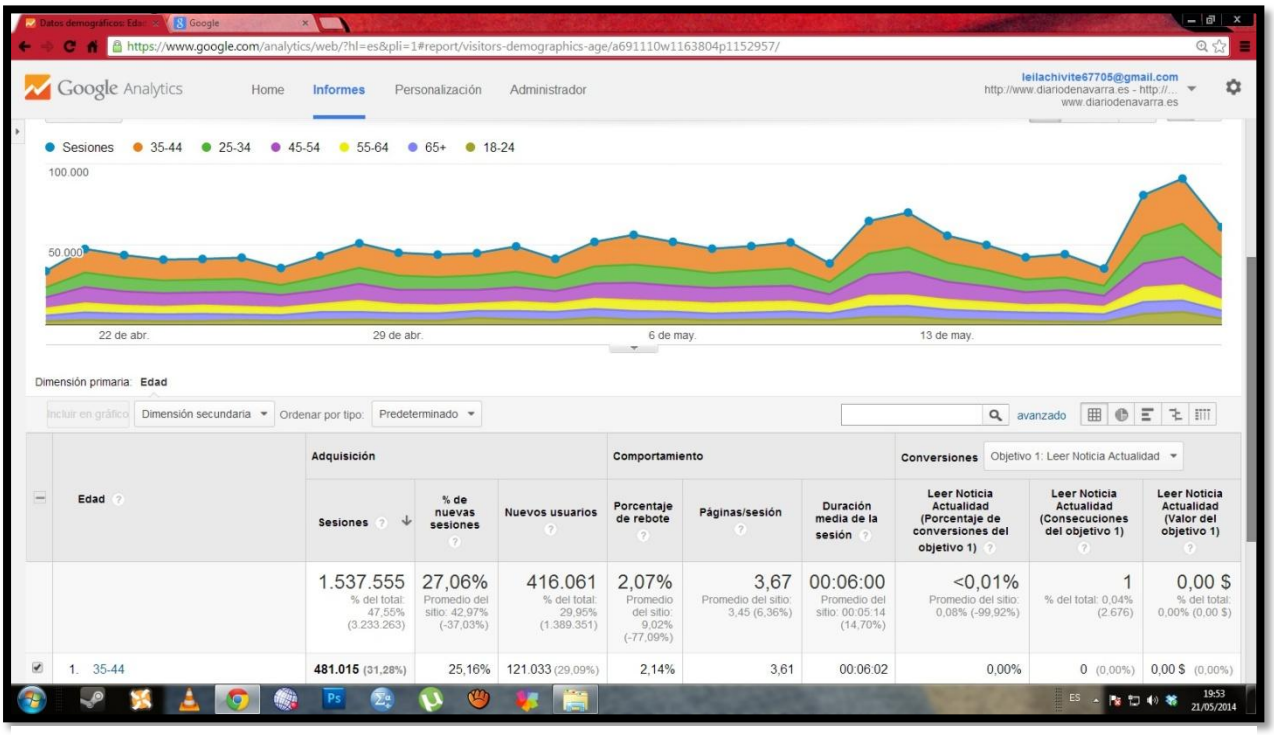

Figura 3. Gráfico de trafico de usuarios por edad

MÉTRICAS RELACIONADAS CON ADQUISICIÓN

Dentro de este apartado he centrado mi análisis en la métrica de **Canales**. Este apartado describe el carácter de procedencia de las visitas a nuestra página web, que puede ser de cuatro tipos: Tráfico directo, tráfico de búsqueda, Redes Sociales y tráfico de referencia. El tráfico directo se refiere a los usuarios que ya conocen la página web y han introducido en el navegador la "url" (dirección web) completa. El tráfico de búsqueda reúne a los usuarios que han dado con la página a través de un buscador. Redes sociales, se refiere a la procedencia de usuarios desde wikis, blogs, y todo tipo

de Social Media. Y por último el tráfico de referencia se refiere a los usuarios que han accedido a la página a través del correo electrónico, un link de otro sitio web, etc.

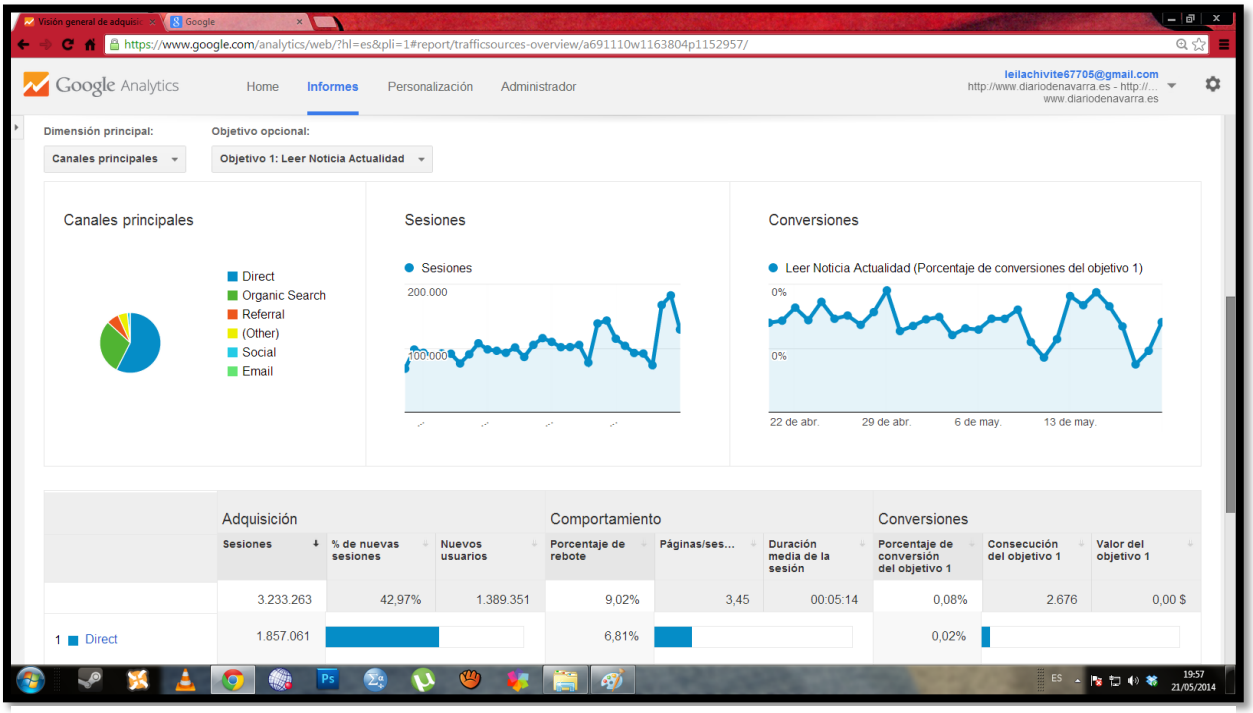

Figura 4. Gráficos generales de adquisición

MÉTRICAS RELACIONADAS CON COMPORTAMIENTO

Las métricas de comportamiento describen qué contenido de la página ha sido más visto, o sea, valorado por la población. He seleccionado tres métricas:

- **Número de páginas vistas:** Este es un índice que muestra cuantas visitas únicas (o sea la primera visita a la sección de un usuario concreto) ha recibido una sección de nuestra página.
- **Páginas de salida:** refleja el dato de que ha sido lo último que ha visitado el usuario en la página web.
- **Analítica de página:** Se trata de una función de Google Analytics que sirve para visualizar los datos de visita por contenido de la página. Es una forma más sencilla de ver los focos de atención de la población en nuestra página. Cuando

activamos este informe, aparece la página web que se está analizando, nos permite mover el ratón por encima de los apartados de la misma, y nos da los datos acerca del contenido que estamos apuntando. He de añadir, que es un informe muy valioso, ya que ejecuta una visualización de datos espléndida, pero tiene un problema básico: no todos los ordenadores pueden activar esta función. Se necesita cierta potencia tecnológica para poder ver este informe.

INFORMES EN TIEMPO REAL

Cuando me encontraba a la mitad de mi análisis, Google Analytics, terminó de implementar una función muy interesante. Los informes sobre público y comportamiento en tiempo real. Podemos saber cuántas personas están utilizando la web, donde se están concentrando esas visitas y a qué hora. El informe se actualiza cada pocos segundos, se trata de un avance muy interesante. Supongamos que queremos medir el impacto o el interés en un determinado evento social, pero que dicho evento no es el centro de nuestra página, gracias al informe en tiempo real podemos saber cuándo hay más movimiento en nuestra web y si se debe o no a la sección que habla del evento.

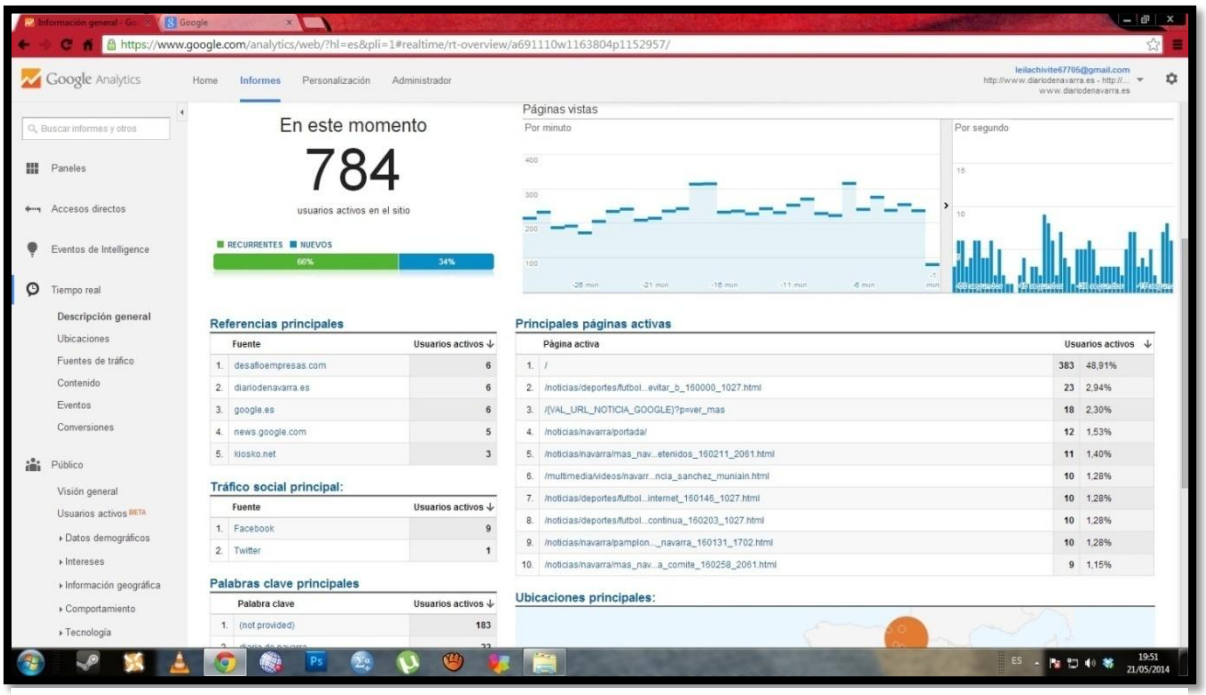

Figura 5. Gráficos de la función tiempo real

8.3 ¿DE DÓNDE PROVIENEN LOS DATOS QUE RECOGE GOOGLE ANALYTICS?

En un principio, cuando comencé mi análisis intuía que Google Analytics obtenía los datos utilizando algún tipo de sistema de etiquetas y de almacenamiento de hipertextos, y que para obtener los datos demográficos utilizaba la información de los registros de Gmail. A lo largo de la investigación encontré algunos problemas, sobre todo con los datos demográficos que no alcanzaba a comprender. Cuando di con la categoría de datos "not provided" en el apartado de datos demográficos, pensé que tal vez era algún error del sistema, así que tuve en indagar un poco más y para comprender de dónde provienen los datos que analiza Google Analytics tuve que consultar a un experto en informática, porque a pesar de saber que esta herramienta, y la propia empresa Google, utilizan el sistema de cookies, no llegaba a entender el proceso en su totalidad.

Una cookie, en el lenguaje informático, es una información enviada por la web y que se almacena en el navegador del usuario de Internet. En un inicio, este sistema sirve para que el sitio web pueda recordar al usuario, lo que facilita la navegación por Internet a las personas.

Las cookies cumplen dos funciones básicas: ejercen un control sobre la navegación de los usuarios para facilitar el proceso de navegación en Internet, y almacenan información acerca de los hábitos de los usuarios en Internet. Cuando hablo del control de los usuarios no me refiero a las personas, sino a las máquinas. La cookie identifica la señal del computador al navegador. Pero la función que me ha interesado para entender el funcionamiento de Google Analytics es la segunda, la recopilación de la información sobre los hábitos de navegación de la población en Internet. Esta función no solo crea dilemas morales y de privacidad, también resulta ser un potente sistema de creación de Big Data. Nos encontramos ante un potente arma de doble filo, que nos proporciona un nuevo y fascinante campo de estudio; y al

mismo tiempo abre las puertas de nuestros ordenadores (de nuestras vidas) a crackers[8] e indeseables que pueden robar nuestra identidad.

Los archivos que conforman una web viajan a través de la red en forma de "paquete de información" (paquetes de datos) usando el protocolo HTTP (Protocolo de Transferencia de Hipertexto). Cuando navegamos por las páginas de Internet recibimos y enviamos estos paquetes de información. Antes de la existencia de las cookies esta comunicación entre ordenadores y servidores web era más costosa, se debía a que el protocolo HTTP no guarda "estados", no guarda información. Es decir, cada vez que escribíamos la url de la web de la universidad, el ordenador tenía que enviar todos los datos al servidor, tenía que empezar desde cero cada vez que queríamos acceder a esa página web: las cookies nacieron para hacer ese proceso más sencillo. El sistema de cookies permite al ordenador recordar los "paquetes de información", por ejemplo, recuerda que hemos visitado la página web de la universidad. Además facilita el cambio de estado de una url, es decir, cuando abrimos la página de la universidad digamos que se encuentra en un estado de "inicio", cuando pinchamos en un link o una pestaña de la página la url cambia de estado. Esto se traduce en que si dejamos que nuestro ordenador utilice el sistema de cookies navegaremos de una forma más cómoda y más rápida. De este avance informático derivaron dos intereses: la utilización de la información de las cookies para la investigación de mercados y el robo de identidad en Internet (mucho más fácil con la información almacenada de las cookies).

[8] Se utiliza para referirse a las personas que rompen algún sistema de seguridad. Los crackers pueden estar motivados por una multitud de razones, incluyendo fines de lucro, protesta, o por el desafío.(Colleman. G. 2013).

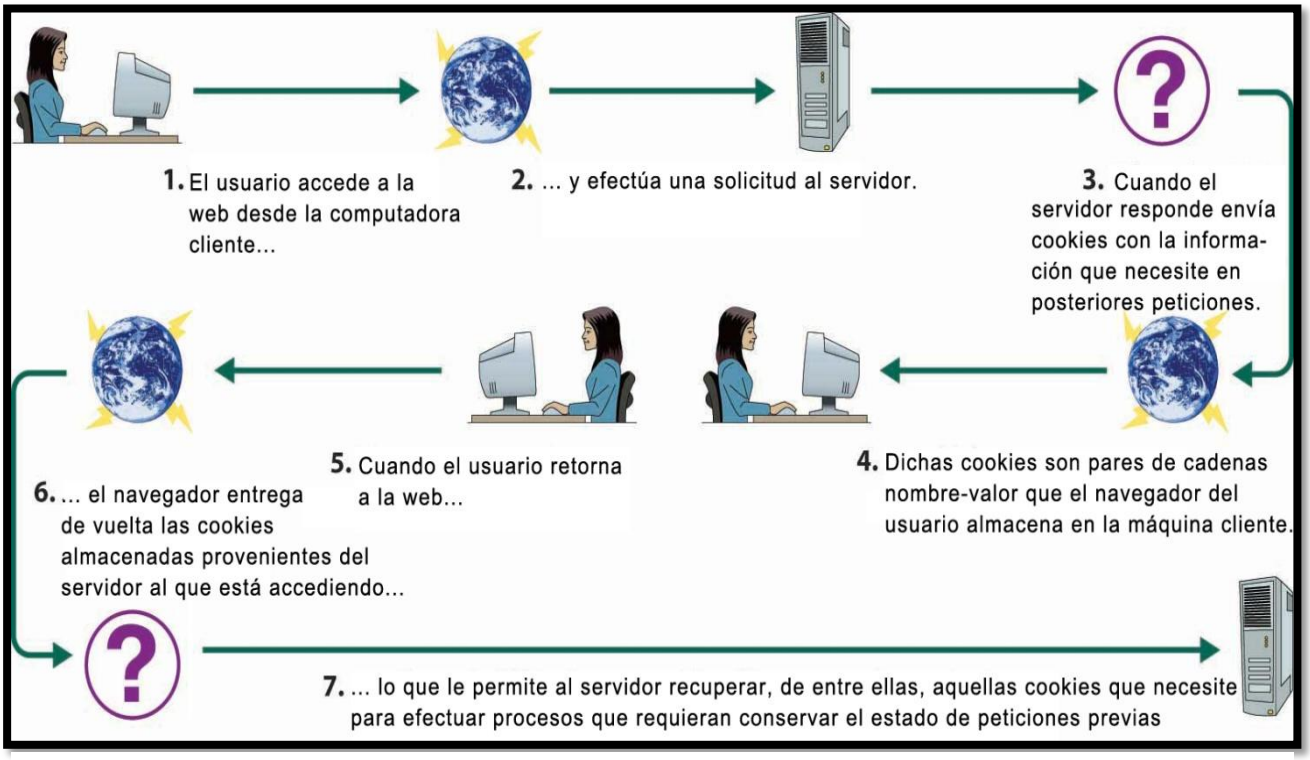

Figura 6. Esquema de funcionamiento de las cookies

(Esta foto puede encontrarse en la quinta página web citada en el capítulo de referencias)

Google Analytics utiliza nuestros hábitos de navegación y la información que "depositamos" en las cookies y las envía al propietario de la web en lenguaje JavaScript. Este lenguaje es leído y contabilizado por la herramienta para crear los informes estadísticos. Google utiliza las cookies HTTP, siempre están formadas por un nombre (utilizan un identificador anónimo para diferenciar a los usuarios) y un valor numérico. Las cookies que utiliza Google Analytics pertenecen a Google, no las utiliza otra empresa. La biblioteca de JavaScript que utiliza Google almacena los siguientes tipos de cookies (tabla1 en la siguiente página).

Tabla 1. Cookies utilizadas por Google Analytics

Nombre de la cookie-HTTP	DESCRIPCIÓN
__utma	Se utiliza para distinguir a los usuarios de la página web de las sesiones, distingue digamos entre humanos y máquinas.
__utmb	Se utiliza para distinguir entre una sesión y una visita nueva en nuestra página web. Así como el comienzo y fin de la visita o sesión.
__utmc	Se utiliza para recordar el número y el tiempo de visitas anteriores en nuestra web.
__utmz	Esta cookie nos proporciona los datos del origen de la visita (las fuentes de tráfico web) a nuestra web: si hemos accedido desde un buscador o desde un enlace por ejemplo.
__utmv	Se utiliza para almacenar las preferencias de los usuarios, los contenidos más visitados, los intereses de la población.

8.4 ¿CUMPLE ESTA HERRAMIENTA LAS PROPIEDADES DE EQUIVALENCIA LÓGICA?

A lo largo de mi análisis he descrito cada métrica de Google Analytics, y gracias a la observación que he realizado durante un mes, he podido ver que en un principio, toda métrica utilizada por esta herramienta cumple con las propiedades de equivalencia lógica.

La equivalencia lógica respalda que nuestros estudios estén midiendo "la realidad" de la población de interés. La equivalencia lógica utilizando Google Analytics

puede ser posible siempre y cuando la población esté bien definida. Es decir, si lo que queremos saber es cuál es el grado de interés que puede tener la población de nuestra universidad en las noticias de la Facultad, podríamos utilizar Google Analytics y sus datos sobre el número de visitas y pérdida de visitas nos darían una buena descripción de cuanto interesan las noticias de la facultad y cuales interesan más que otras. Pero si lo que queremos saber, por ejemplo, es las diferentes edades que visitan esta página entonces la equivalencia lógica pierde fuerza pues depende de que Google tenga bien definidas las franjas de edad, una tarea que no siempre se puede asumir.

Google Analytics puede ciertamente crear ilusiones de realidad, por ejemplo, cuando a través de Google Flu Trends (otra herramienta gratuita de Google que ofrece estadísticas basadas en Big Data) "detectaron" una posible epidemia de gripe utilizando el Big data generado por la búsqueda de síntomas gripales en la red. Los datos que se obtuvieron sobre síntomas sin duda aumentaron a causa de la mediatización de esta enfermedad: que exista un elevado número de visitas y consultas a cerca de la gripe no tiene por qué indicar que nos aceche una epidemia. Por esta razón, la equivalencia lógica tiene que ser mirada muy detenidamente en cada estudio que use Google Analytics pues será más o menos robusta dependiendo del problema que estemos estudiando.

Los modelos bayesianos utilizan bases de datos que se actualizan constantemente, y utilizan toda la información disponible para medir y analizar los parámetros. Un método que encaja a la perfección con el nuevo paradigma de Big Data, seguramente debido a la inmensidad de sus bases de datos.

Por supuesto, que esta herramienta no cumpla con las propiedades de la equivalencia lógica, no significa que sus datos nos sean válidos y fiables para realizar una investigación. Simplemente tenemos que cambiar de perspectiva. Cuando disponemos de todos los datos, como en este caso, el margen de error es ínfimo. En el caso del Big Data, y claro, en el caso de los informes que genera Google Analytics, sustituimos el control que ejercemos sobre la muestra poblacional y el proceso en la recogida de los datos, por aceptar que nuestra base de datos contenga errores, pero sabemos que serán mínimos y que no repercutirán en exceso en las conclusiones que

obtengamos en la investigación. También deberemos tener en cuenta que este tipo de herramienta solo nos será útil para realizar investigaciones de carácter descriptivo, es decir, apuntará a los "qué", o sea, nunca podremos obtener conclusiones de causalidad con esta herramienta.

El caso más curioso que he encontrado a lo largo de la investigación y que me sirve para ilustrar la nueva perspectiva del Big Data, es el de los datos demográficos. A simple vista podemos pensar que los datos demográficos que genera Google Analytics no son válidos, a causa de que determina el sexo y la edad a través de las cookies. Pero lo cierto es que tenemos que reconocer que es una idea brillante. Primero debemos abrir la mirada y pensar, que no se basa en cientos de datos sobre nuestros hábitos de navegación para saber si somos hombres o mujeres, sino en cientos de miles, por lo que es complicado que se equivoquen en su deducción. Lo que intento apuntar es que el margen de error con el que juega Google Analytics, no será mucho mayor que el que podemos controlar con una muestra, y que además queda compensado cuando utilizamos toda la población para ejecutar nuestro análisis.

Es evidente que no solíamos disponer de los recursos para realizar investigaciones utilizando a la mayoría de la población, pero eso ha cambiado con la aparición de los Big Data. Tenemos a nuestra disposición bases de datos inmensas y podríamos utilizar la estadística clásica para analizarlas por supuesto, sobre todo si queremos conocer el "porqué" de un fenómeno, pero sería un error desperdiciar la tecnología que se nos ofrece para poder ejecutar análisis de tipo descriptivo sobre poblaciones enormes de datos. Teniendo a nuestro alrededor semejante cantidad de datos digitalizados es lógico utilizar este sistema y por supuesto, utilizar la herramienta gratuita Google Analytics.

El verdadero problema que se plantea aquí no es el tipo de metodología que se emplea sobre las bases de Big Data. La problemática se encuentra en que cualquier base de Big Data solo recoge los datos de la población que navega por Internet. Dejando a un lado minorías étnicas, los colectivos frenados por la brecha digital, personas que no utilizan esta tecnología, o el caso concreto de los niños y niñas que navegan en internet. Google Analytics utiliza en sus datos demográficos unas franjas

de edad que comienzan desde los 18 años, sería un error pensar que los hábitos de navegación infantil no estén computados dentro de la franja de edad más joven. Este dato, debemos tenerlo siempre presente a la hora de realizar cualquier investigación con Google Analytics.

Toda investigación social que queramos hacer debe quedar limitada a la población que utiliza Internet, y más concretamente a la población que utiliza la página web que está analizando Google Analytics. En ningún caso podremos extrapolar los datos que obtengamos de Google Analytics a toda la población fuera de la página que estamos analizando. En pocas palabras, Google Analytics nos ofrece un censo de la población que visita la web, y nos da unos informes descriptivos acerca de la utilización de la web, los intereses de la población y sus datos demográficos. Esto significa que podemos hacer análisis demográfico de esta población concreta, pero no podemos extrapolar nuestra descripción al resto de la comunidad. Para poder contestar a nuestras hipótesis y preguntas de investigación siempre necesitaremos metodología que complemente ha esta herramienta.

8.5 POSIBLES USOS DE GOOGLE ANALYTICS MÁS ALLÁ DE LA INVESTIGACIÓN DE MERCADOS

Aunque Google Analytics este diseñada para que podamos optimizar el rendimiento de un sitio web, esta herramienta puede reinterpretarse. Nos puede ayudar a acercarnos a una comunidad virtual, como Facebook, y ayudarnos a describir los mecanismos de comunicación social que se establecen en ella y la población que la visita.

Gracias a esta herramienta podemos realizar pequeños experimentos, así como las empresas experimentan con los contenidos de publicidad en sus páginas web para medir el impacto en la población y aumentar las ventas, nosotros podemos

centrarnos en observar su comportamiento en función de los diferentes estímulos que podemos introducir en el sitio web.

Imaginemos que queremos realizar una investigación social sobre el comportamiento que se da en las redes sociales y qué tipo de población atraen, y en concreto pensemos en Facebook. Considero que Facebook presenta un ejemplo claro de cómo podemos implementar la herramienta de Google Analytics en una investigación sociológica sin fines meramente lucrativos.

Podríamos crear una página en Facebook que tratara sobre el grado de sociología, que informara sobre las ultimas normativas de la facultad, fiestas universitarias, descripciones de las asignaturas, recomendaciones para obtener bibliografía útil, eventos para compartir apuntes o ayudar a alumnos/as de otros cursos, etc. Sé que este ejemplo no corresponde a la perfección con la idea de Big Data porque nuestro grado cuenta con poco alumnado, pero en cualquier caso, es un ejemplo útil para ilustrar las facilidades que puede aportar la herramienta Google Analytics.

Ahora imaginemos que queremos saber qué tipo de alumnado participa más en esta página y cuales fomentan más la compartición de apuntes y la ayuda entre compañeros/as. Gracias a los datos que ofrece Google Analytics sabríamos si el perfil más "cooperativo" en nuestro grado es femenino o masculino, la sección de *datos demográficos* ofrece una descripción bastante fiable de los perfiles que visitan nuestra web, por lo que podríamos mejorar la cooperación entre los alumnos.

También pudiera ser interesante saber que asignaturas les resultan más complicadas para el alumnado, dato que podríamos obtener gracias a esta herramienta sin que un profesor/a plantee ese delicado tema en su clase. En primer lugar podríamos acudir a la *métrica de número de visitas* en la sección de comportamiento, fijándonos en que asignatura ha generado más visitas. Después podríamos ver el *gráfico de flujo de usuarios* y ver si las visitas que ha generado esa asignatura han llevado a los usuarios de la página a la sección de "ayuda y compartición de apuntes". De esta forma sabríamos qué asignaturas resultan más complicadas para el alumnado. Una vez conocido este dato podríamos pensar en

poner más trabajos en grupo en la asignatura, dedicar más horas del plan de estudios, o lo que se considerara oportuno.

En resumen, no porque los pioneros en investigar las bases de Big Data pertenezcan al campo de la investigación de mercados, significa que la oportunidad Big Data esté vetada para la sociología. La utilidad que puedan tener los Big Data está a la espera de nuestra curiosidad e imaginación.

9. CONCLUSIONES Y CUESTIONES ABIERTAS

Este estudio presenta evidencias claras. En primer lugar, nos encontramos ante una nueva realidad social, y una nueva realidad de datos: El Big Data. La cantidad de datos que obtenemos hoy de la población y que digitalizamos genera un nuevo paradigma de investigación, que podemos abordar desde la investigación social.

Queda claro que el uso de Big Data ha proporcionado ayuda y da éxito a muchas empresas. La velocidad de la recogida de datos, la cantidad de datos de la que disponemos aumenta las posibilidades de investigación, y ahora con la tecnología informática de seguimiento web y de análisis estadístico, facilita nuestro trabajo.

Google Analytics no solo nos ofrece un análisis válido, sino que además lo hace gratis. Por supuesto, todo tiene su lado negativo. Google nos proporciona los datos de los usuarios gracias al sistema de cookies que utiliza, un sistema que nos facilita la recogida de datos sobre la población, pero que al mismo tiempo pone en riesgo nuestra seguridad informática y nuestra privacidad en Internet.

El lado positivo, y más interesante de mi análisis es la información que implementa Google Analytics como datos demográficos. Determinar si un usuario de Internet es hombre o mujer, y qué edad tiene, supone toda una revolución en el campo de la informática y en el campo de la recogida de datos.

Para comprender hasta que punto estos datos pueden ser valiosos y fiables, debemos abrir nuestra mirada. Resulta evidente que los datos demográficos obtenidos por Google Analytics no son cien por cien fiables, pero hay que tener en cuenta que Google no utiliza una pequeña cantidad de cookies, o sea de hábitos de navegación, para determinar si un usuario es hombre o mujer: utiliza cantidades inimaginables de datos para ello, cientos de miles. Esto apunta a que ya desde un inicio los datos demográficos de esta herramienta contarán con un margen de error muy bajo. Pero lo más importante es que, al utilizar el enfoque bayesiano, al trabajar con toda la población, con todos los datos, no importa que los datos contengan algunos errores.

Por supuesto, este nuevo estilo de análisis no significa que la estadística tradicional, y los métodos cuantitativos de investigación social queden obsoletos. Las estadísticas que nos ofrece Google Analytics solo nos ofrecen una descripción de un fenómeno o de la población, pero no pueden responder al "por qué". Para responder a la causalidad de un fenómeno ya sea ocurrido en Internet o no, es necesario utilizar otra metodología. El caso es que la estadística clásica fue concebida en un momento en el que no podíamos ejecutar un análisis sobre una población entera, por lo que se perfecciono el método para acercarnos lo más posible a la población a través de una muestra. Ahora con los datos digitalizados, y la tecnología adecuada, podemos realizar análisis descriptivos de la población sin hacer uso de una muestra: tenemos el censo (Mayer. V. y Niel. K. 2013).

El auténtico problema de las bases de Google Analytics o de cualquier Big Data es que los censos que genera solo computan datos sobre la población que navega en Internet (discriminando minorías étnicas y personas que no utilizan Internet a causa de la Brecha Digital por ejemplo) lo que me lleva a insistir en que no podemos extrapolar nuestras conclusiones al resto de la población. Las descripciones que obtenemos a través de Google Analytics responden a la población que navega por la página que analiza, no nos dice nada a cerca de su vida off-line, solo sobre sus hábitos en dicha página.

La capacidad de generar información valiosa de Google Analytics despierta un interés para nosotros y un deber de contribuir con el avance científico. La lógica Big Data nos puede ayudar a detectar problemas sociales y a responder a los mismos, además, la filosofía que encierra esta lógica no solo es la de la acumulación y actualización de la información, sino que también aprende sobre sus propios errores; lo que puede ser de gran utilidad para gobiernos, empresas, organizaciones no lucrativas, etc. Pero en ningún caso podemos perder de vista el hecho de que las bases de Big Data, ya sean generadas por Google Analytics, o por otro sistemas, solo nos ofrecen DESCRIPCIONES y solo sobre la población que navega en Internet.

Utilizar Big Data para establecer correlaciones entre los hábitos de navegación para detectar una pandemia, por ejemplo, es un acto irresponsable que puede costar

muy caro a investigadores y gobiernos. Big Data facilita nuestro trabajo, en tanto que ahora podemos describir poblaciones en la red fácilmente, pero no representa una solución a los problemas sociales, y no es un oráculo que desbancará al científico por el adivino.

Este estudio representa el posible principio de otras investigaciones basadas en la evaluación de herramientas de medición y utilización de Big Data. Una herramienta que podría ser sujeto de este tipo de investigación es Google Flu Trends, esta herramienta ha levantado cierta polémica en los últimos meses a causa de la imprudencia que cometen algunos utilizando los datos que proporciona: cómo es el famoso caso de la epidemia de gripe pronosticada por Google. La empresa Google llevaba varios años pronosticando los brotes de gripe, hasta ahora. El Centro de Control de Enfermedades ha descubierto que las previsiones de Google estaban siendo exageradas. Como he explicado anteriormente, los datos de navegación a cerca de la gripe y sus síntomas aumentaban con la mediatización de la posible epidemia, lo que provocaba la exageración de los pronósticos de Google Flu Trends. Podría resultar muy interesante evaluar esta herramienta.

10. REFERENCIAS

LIBROS

- **Mayer. V.; Niel. K.** (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* John Murray Publishers. London.
- **Jones. S**. (1999) *Doing Internet Research: Critical Issues and Methods for Examining the Net*. SAGE.
- **Clifton. B.** (2008) *Advanced Web Metrics with Google Analytics*. Sybex
- **Cicourel. A.** (2011). *Método y medida en Sociología*.CIS. Madrid.
- **Rivadulla, A.** (1991). *Probabilidad e inferencia científica.* Barcelona: Anthropos.
- **Colleman. G.** (2013). *Coding Freedom: The ethics and aesthetics of hacking.* New Jersey . Published by Princeton University Press.
- **Hine.C.** (2004) *Etnografía virtual. Barcelona*. UOC.
- **Zikopolous, P.; Deroos D.; Deutsch T.; Lapis G**. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. IBM Developements.
- **Foster. K.; Nathan. S.; Rajan. D.; Ballard. C.** (2011). *IBM InfoSphere Streams: Assembling Continuous Insight in the Information Revolution*. IBM RedBooks.
- **O'Hagan, A.; Forster, J**. (2004). *Bayesian Inference: Kendall's Advanced Theory of Statistics*. London: Arnold.

ARTÍCULOS

- **Fang. W.** (2007). Using Google Analytics for Improving. Library Website Content and Design: A Case Study. *Library Philosophy and Practice (e-journal)*. Vol.6. (1). [Disponible en (07/03/2014):http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.5924&rep=rep1&type=pdf]

- **Plaza. B.** (2011). Google Analytics for measuring website performance. *Tourism Management*. 477-481. Vol. 32,.(3).[Disponible en (10/03/2014):http://www.sciencedirect.com/science/article/pii/S0261517710000622]
- **Hasan. L; Morris. A, Probets. S.** (2009). Using Google Analytics to Evaluate the Usability of E-Commerce Sites. Human Centered Design. *Lecture Notes in Computer Science*. 697-706. Vol.5619. [Disponible en (14/02/2014):http://link.springer.com/chapter/10.1007/978-3-642-02806-9_81]
- **Plaza. B.** (2009). Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. *Aslib Proceedings.* 474 - 482. vol 61.(5). [Disponible en (15/02/2014):http://www.emeraldinsight.com/journals.htm?articleid=1811878&show=abstract]
- **Steven J. Turner M.** (2010). Website Statistics 2.0: Using Google Analytics to Measure Library Website Effectiveness. *Technical Services Quarterly*. 261-278 Vol. 27. (3). [Disponible en (17/02/2014):http://www.tandfonline.com/doi/abs/10.1080/07317131003765910#.U1pTFvl_uQA]
- **Crutzen.R.; Roosjen. J.; Poelman. J.** (2012). Using Google Analytics as a process evaluation method for Internet-delivered interventions: an example on sexual health. *Department of Health Promotion, Maastricht University. The Netherlands. Published by Oxford University Press.* 36-42. Vol.28. (1):. [Disponible en (15/02/2014):http://heapro.oxfordjournals.org/content/28/1/36.short]
- **Alamilla,N. Jiménez,J**. (2010) Contraste de Hipótesis: Clásico vs Bayesiano. *Revista digital Matemática, Educación e Internet*. Vol. 11 (1). [Disponible en (06/03/2014):[http://www.tecdigital.itcr.ac.cr/revistamatematica/ARTICULOS_V11_N1_2010/NAlamilla_ConstrastedeHipotesis/1_NAlamilla_JJimenez_Constraste%20de%20hipotesis.pdf]

- **Ayçaguera, L y Benavides, A.** (2003). Apuntes sobre subjetividad y estadística en la investigación en salud. *Revista Cubana de Salud Pública*. 170-173.Vol.29.(2)[Disponible en (14/02/2014):http://scielo.sld.cu/scielo.php?pid=S0864-34662003000 200012& script=sci_arttext & tlng=es].

PÁGINAS WEB

(Glosarios on-line y blogs)

- http://www.efectosjavascript.com/javascript.html [Última consulta: 17/03/2014]
- http://www.enriquedans.com/tag/big-data [Última consulta: 11/03/2014]
- http://blogs.sap.com/latinamerica/big-data-una-oportunidad-para-empresas-de-todos-los-tamanos/ [Última consulta: 17/03/2014]
- http://planetbigdata.com/ [Última consulta: 11/03/2014]
- http://cursojavaee.blogspot.com.es/2013/05/esquema-explicativo-del-funcionamiento.html [Última consulta: 16/03/2014]
- http://www.internetglosario.com/letra-j.html [Última consulta: 17/05/2014]

11. ANEXOS

A. Tabla de equivalencias de datos informáticos (elaboración propia).

UNIDAD DE MEDIDA DE LA INFORMACIÓN	EQUIVALENCIA
Dígito binario	**1 Bit**
1 Byte	**8 Bits**
1 Kilobyte	**1024 bytes**
1 Megabyte	**1024 Kilobyte**
1 Gigabyte	**1024 Megabyte**
1 Terabyte	**1024 Gigabyte**
1 Petabyte	**1024 Terabyte**
1 Exabyte	**1024 Petabyte**
1 Zettabyte	**1024 Exabyte**
1 Yottabyte	**1024 Zettabyte**
1 Brontobyte	**1024 Yottabyte**
1 Geopbyte	**1024 Brontobyte**