

Minería de Medios Sociales

Máster en Ciencia de Datos e Ingeniería de Computadores



Tema 0: Minería de Medios Sociales

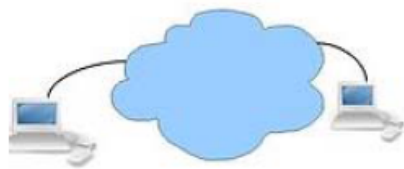
Oscar Cordon García

*Dpto. Ciencias de la Computación e Inteligencia Artificial
ocordon@decsai.ugr.es*

WEB 2.0 Y MEDIOS SOCIALES

EVOLUCIÓN DE LA WEB (1)

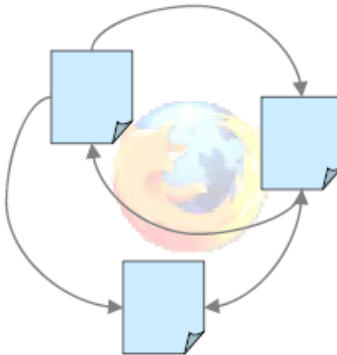
Internet



connecting
computers

1960's

World Wide Web

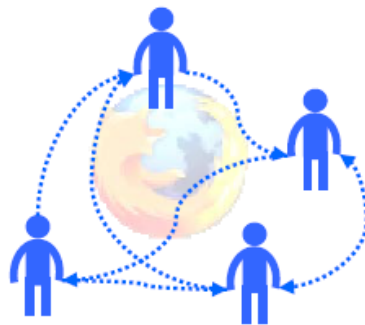


connecting
documents

1990's



Web 2.0,
The Social Web



connecting
people

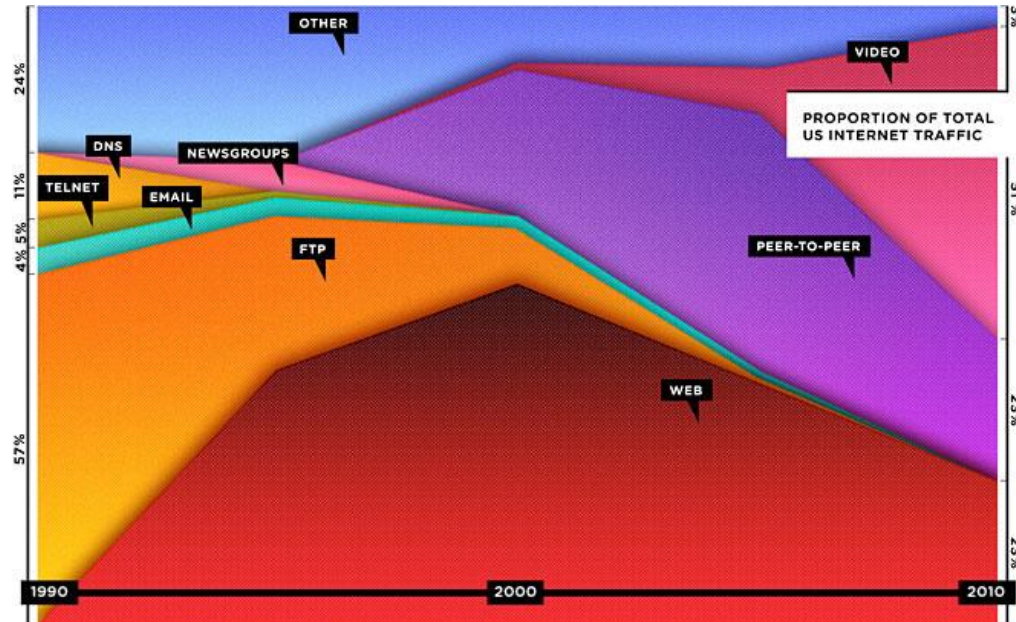
2004-present



EVOLUCIÓN DE LA WEB (2)

Chris Anderson, Wired, “The Web is dead. Long Live the Internet”:

<https://www.wired.com/2010/08/ff-webrip/>



Two decades after its birth, the World Wide Web is in decline, as simpler, sleeker services — think apps — are less about the searching and more about the getting

Los **Medios Sociales** comprenden el uso de herramientas electrónicas y de Internet para compartir y comentar información y experiencias entre seres humanos de forma eficiente

Los **Medios Sociales** son un “conjunto de aplicaciones basadas en Internet que constituyen los fundamentos ideológicos y tecnológicos de la **Web 2.0**, y que permiten la creación y el intercambio de contenidos generados por los usuarios”

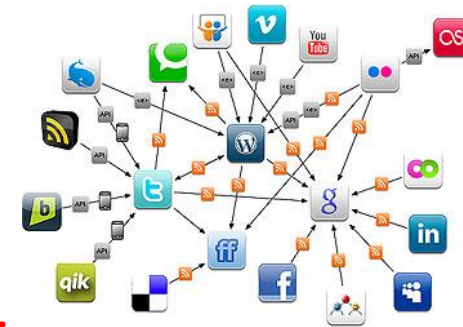
MEDIOS SOCIALES (2)

Vivimos en la era del big data. Con cientos de millones de personas invirtiendo horas y horas en los medios sociales para compartir, comunicar, conectar, interactuar y crear contenidos generados por los usuarios con un volumen sin precedentes, los medios sociales se han convertido en una fuente única de *big data*. Dicha fuente muestra un gran potencial para la investigación y el desarrollo. Desafortunadamente, tener más datos no necesariamente implica generar mejor información, sólo los datos de buena calidad lo permiten. Así, es necesario disponer de nuevos métodos computacionales para hacer minería sobre esos datos. Los datos de medios sociales son ruidosos, sin formato, de longitud variable y multimedia. Además, las relaciones sociales entre las entidades, llamadas redes sociales, forman una parte inseparable de los datos de medios sociales. Por tanto, es importante que se combine el uso de las teorías sociales y los métodos de investigación con los métodos estadísticos y de minería de datos. Es un tiempo propicio para la minería de medios sociales.



This chapter is from *Social Media Mining: An Introduction*.
By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu.
Cambridge University Press, 2014. Draft version: April 20, 2014.
Complete Draft and Slides Available at: <http://dmml.asu.edu/smm>

- Contenido generado por los usuarios: masivo, dinámico, amplio, instantáneo y ruidoso
- Rico en interacciones entre los usuarios: datos con relaciones/enlaces (*linked data*)
- Entorno colaborativo, sabiduría de las masas
- Muchos grupos pequeños (fenómeno de cola larga)
- La atención es cara
- **Comunicación muchos a muchos: la información llega a los usuarios**
... vía la influencia personal en sus redes sociales
... a través de la transmisión de los medios de masas
- Los medios sociales están diseñados para ser diseminados mediante **interacción social**
 - ¿Cómo interacciona la información transmitida con la influencia personal de las redes sociales?
 - Tensión entre los efectos globales de los medios de masas y los efectos locales provocados por la estructura social





MEDIOS SOCIALES (5):

Categorías

- Redes sociales on-line (Facebook, LinkedIn)
- *Microblogging* (Twitter → X)
- Compartición de fotos (Flickr, Picasa, Instagram, Snapchat)
- Agregadores de noticias (Reddit)
- Compartición de video (YouTube, Tik Tok)
- *Livecasting* (Justin.TV, Twitch)
- Mundos virtuales (Kaneva, Myspace, Decentraland)
- Juegos on-line (Warcraft, ...)
- Búsqueda “social” (Google, Bing)
- Mensajería instantánea (Google Talk, Skype, Whatsapp, Telegram)
- ...

 <p>Online Social Networks</p>	 <p>Blogging</p>	 <p>Microblogging</p>
 <p>Wikis</p>	 <p>Social News</p>	 <p>Social Bookmarking</p>
 <p>Media Sharing</p>	 <p>Opinions and Reviews</p>	 <p>Answers</p>

- Un artículo de una wiki
- Un post en un blog, una discusión en Reddit
- Una opinión/puntuación en una web de restaurantes sobre una hamburguesería famosa de tu ciudad
 - Ejs: Yelp.com, Tripadvisor.es
- Los datos de un usuario en una red social on-line de contactos profesionales
 - Ejs: LinkedIn.com, Facebook.com
- Los datos de una app que indica la localización de las cámaras en una calle
- Los datos de una recomendación de Amazon



WIKIPEDIA
The Free Encyclopedia



MINERÍA DE MEDIOS SOCIALES

*La **Minería de Medios Sociales** es el proceso de representar, analizar y extraer patrones con significado a partir de datos de medios sociales*

La **Minería de Medios Sociales** es un área multidisciplinar formada por:

- **Redes sociales:**
 - interacciones y comunidades
- **Medios sociales:**
 - Gran cantidad de datos (masivos, dinámicos, amplios, instantáneos y ruidosos)
 - Doble naturaleza: información de contenido y relaciones
- **Minería de datos:**
 - Modelado, Aprendizaje y Predicción

- Aparición de nuevos fenómenos observables de las **interacciones** entre personas en los medios sociales
- Oportunidades sin precedentes para **estudiar el comportamiento humano** mediante investigación interdisciplinar y colaborativa entre las ciencias sociales y de la computación:
 - Teorías sociales derivadas de años de investigación
 - Tecnologías computacionales y algoritmos escalables para Minería de Datos/Big Data



MIT Laboratory for Social Machines: <http://socialmachines.media.mit.edu/>

MIT Center for Constructive Communication:

<https://www.media.mit.edu/groups/center-for-constructive-communication/overview/>

Promoting deeper learning and understanding in human networks

Análisis de imagen de marca:

- ¿Qué está opinando la gente de nuestra marca? (**análisis de sentimientos/minería de opiniones, análisis de confianza**)



Campañas de marketing:

- Compañías tratando de posicionar sus productos: publicidad “ganada” en Redes Sociales on-line (*earned media*, boca a boca, procesos de difusión en redes, marketing viral)

Recomendaciones personalizadas:

- **Sistemas de recomendaciones, filtrado colaborativo, marketing viral**



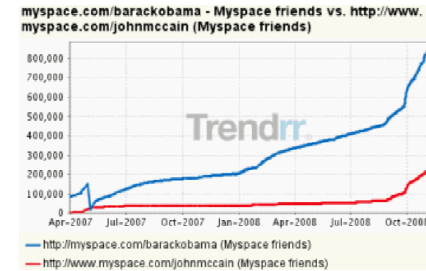
Informes de productos:

- Minería de revisiones de productos en foros para obtener información sobre la percepción de sus características, identificar nuevas necesidades, etc. (**análisis de sentimientos, análisis de confianza, minería de textos**)

Campañas políticas:

- ¿Por qué los ciudadanos apoyan a un candidato? Predicción de/ Influencia en procesos electorales (**boca a boca, procesos de difusión en redes, análisis de influencia, minería de blogs**)

<https://www.technologyreview.com/2012/12/19/114510/how-obamas-team-used-big-data-to-rally-voters/>



Cuerpos de Seguridad del Estado:

- Predicción de actividades tales como disturbios de bandas y manifestaciones no autorizadas (**minería de tuits**)

NYT: Sending the Police Before There's a Crime:

http://www.nytimes.com/2011/08/16/us/16police.html?_r=1

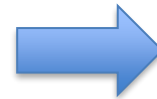
Periodismo ciudadano:

- Mayor valor que los teletipos. Problema de tener que “bucear” en muchos *posts* para localizar información útil
- Creación de resúmenes en tiempo real (**análisis de influencia, minería de textos**)

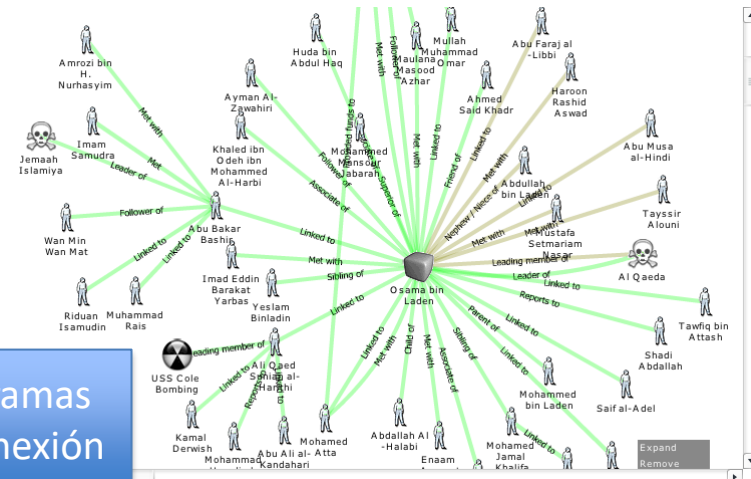


- Procesado de datos de medios sociales para obtener herramientas de análisis para:
 - Identificar redes sociales: miembros, grupos (**descubrimiento de comunidades**)
 - Identificar tópicos y sentimientos (**minería de sentimientos**)
 - Monitorización de tendencias culturales (**minería de *tuits***)
 - Salud 2.0: Predicción de epidemias (**procesos de difusión en redes complejas**)

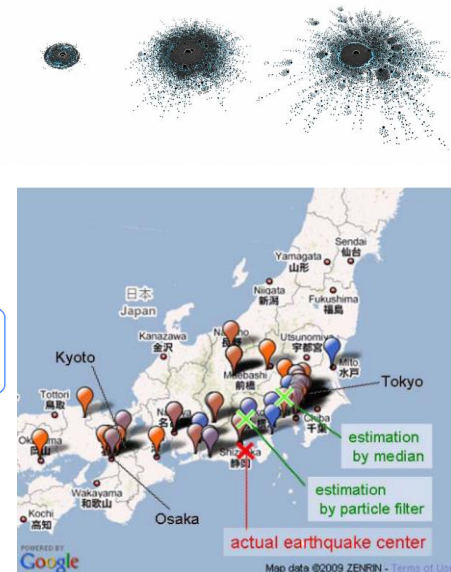
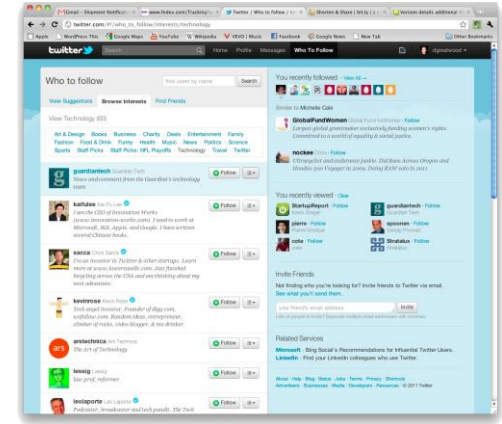
Modelado predictivo



Diagramas de conexión



- Los usuarios tuitean mensajes cortos:
 - Se retuitean mensajes de otros
 - Los tuitean pueden contener URLs de contenido online, noticias
- **Redes sociales:** Los usuarios siguen a los amigos (*friends*) para ver tuits y retuits de amigos
- **Contagio social (cascadas):**
 - ¿se retwiteará un tuit? ¿hasta dónde llegará?
 - La mecánica de los *memes* es similar a la de los virus biológicos e informáticos
- **Detección de eventos en tiempo real:** epidemias, terremotos, movimientos ciudadanos, etc.



- Las redes sociales on-line y los servicios de microblogging se configuran en una **estructura de red** (compleja) **social**
- La estructura de la red determina los procesos dinámicos de transmisión de información que se producen sobre ella (**análisis de redes sociales**)
- Es posible determinar la **influencia de un usuario** a partir de la red social (p.ej. a quién seguir en Twitter en una crisis)
- Igualmente, es posible determinar la **estructura de comunidades** subyacente

Que es Klout

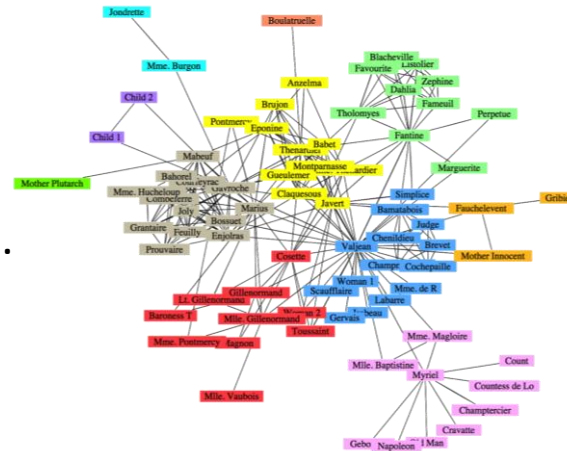
Klout es un Servicio Web que mediante un índice llamado Klout Score mide el grado de influencia de una persona o una marca en las Redes Sociales. Para determinar el Klout Score de una persona el Servicio Web analiza más de 400 parámetros distintos de las 7 Redes Sociales más importantes y se asigna una puntuación entre 1 y 100 a los usuarios.

El promedio de los usuarios de Klout es de 40 y se considera como un **influencer** a aquellas personas con un índice alto por encima de la media, por ejemplo, los que pertenecen al grupo del 5% de usuarios con un valor superior a 60 están considerados como los más influyentes.

No es un índice cuantitativo que mida el volumen de contenidos generados o de seguidores de una persona, si no la capacidad de influir que tiene al transmitir sus contenidos. Usuarios con miles de fans en Facebook o Twitter pero cuyos contenidos no sean compartidos ni interaccionen con ellos (Me gusta, RT, Respuestas etc.) no tendrán un Klout Score alto y en cambios otros con menos contactos pero cuyos contenidos se expandan por la Red al ser compartidos y comentados pueden tener índices muy altos.



<https://theproductmanager.com/general/what-happened-klout/>



One Per Cent

Taking the sweat out of technology

How to find the right Twitter user in a crisis

12:46:20 March 2013

Twitter social media

Hal Hodson, technology reporter



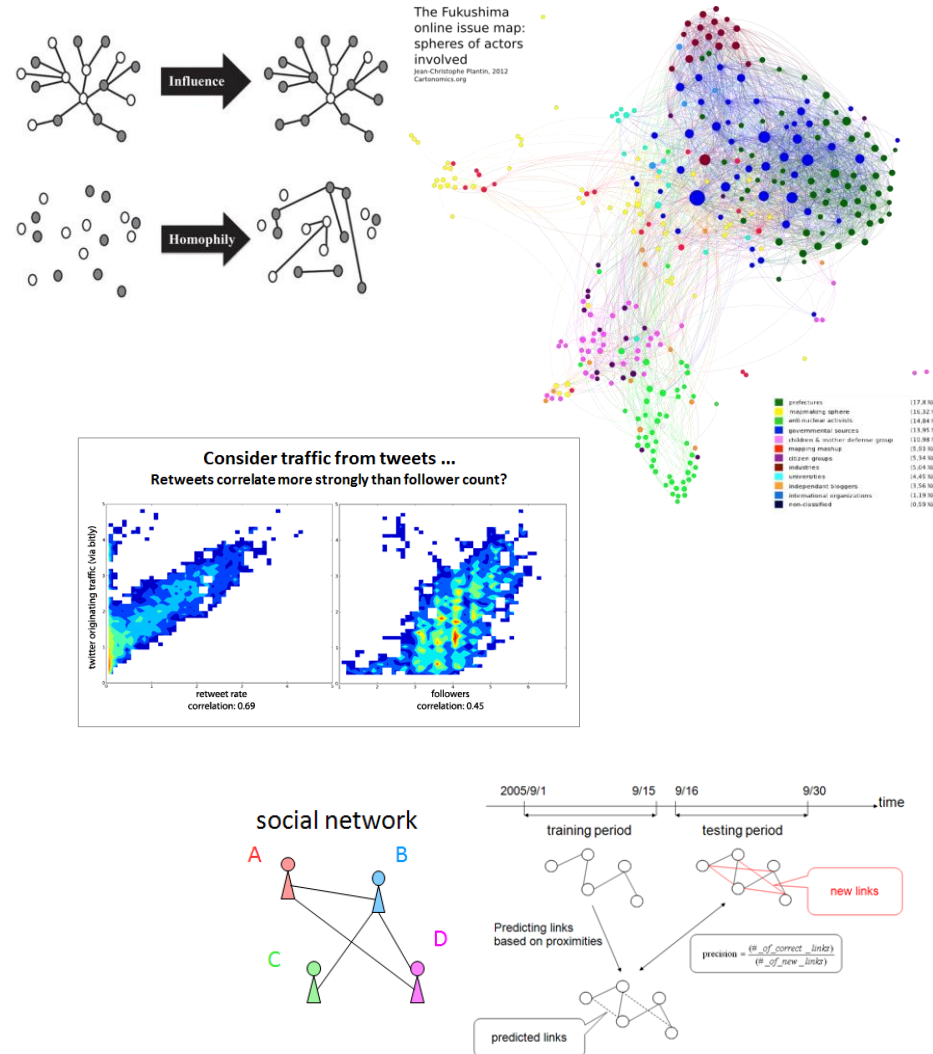
(Image: AFP/Getty)

Honing your Twitter feed can be a chore at the best of times, but when you want the latest information during a natural disaster or national uprising, things get a lot harder.

APLICACIONES (3):

Influencia y Predicción de Enlaces

- Las teorías de relación social (**asortatividad**) son trasladables a los medios sociales:
- Influencia**: personas relacionadas tienden a tener intereses similares. **Homofilia**: personas con intereses similares tienden a relacionarse
- En Twitter se modelan con cuatro relaciones: *co-post*, *co-following*, *co-followed*, *following*
- Se usan medidas como el grado de entrada, el número de menciones o el de retuits para medir **influencia de usuarios**
- En redes sociales on-line, se pueden usar para **recomendar amigos/contactos** (listas ordenadas de pares de nodos proclives a conectarse)
- Son también muy útiles en sistemas de recomendaciones colaborativos



- El papel de la confianza es crítico en algunas comunidades on-line tales como los **sitios web de revisión de productos y de comercio electrónico**
- Los usuarios de sitios de este tipo (ej. el antiguo *Epinions*) proporcionan tanto revisiones de productos como **redes de confianza**
- Estas redes se explotan para predecir la calidad de la revisión y para mejorar la precisión de los sistemas de recomendaciones

Epinions Unbiased Reviews by Real People

Categories: Cars, Books, Movies, Music, Computers & Software, Electronics, Gifts, Home & Garden, Kids & Family, Office Supply, Sports

Home > Member Center > mrroland

Web of Trust

Who trusts:

- luquillo
- k1rdm
- java73
- sam-pro
- becky2259

View all 26 members whom mrroland trusts

Who is trusted by:

- luquillo
- k1rdm
- java73
- sam-pro
- rubez08

View all 26 members who trust mrroland

Web of Trust

Trust mrroland

Block mrroland

Whom should I trust?

's Profile

About: [Redacted]

Member: [Redacted]

Epinions.com ID: [Redacted]

Location: Atlanta, GA

Member Since: Jun 14 '08

Activity Summary

Reviews Written: 65

Member Visits: 1,893

Total Visits: 5,755

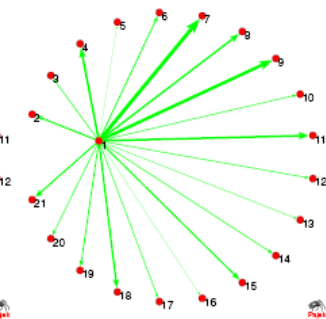
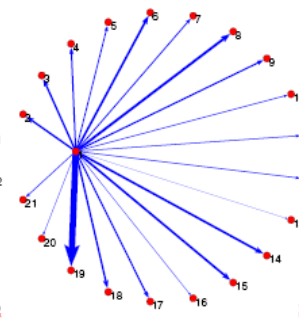
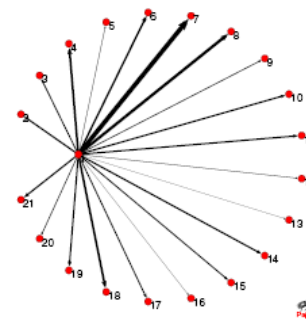
Favorite Websites: Atlanta Braves

POPULAR AUTHOR: TOP 1,000,000,000 more

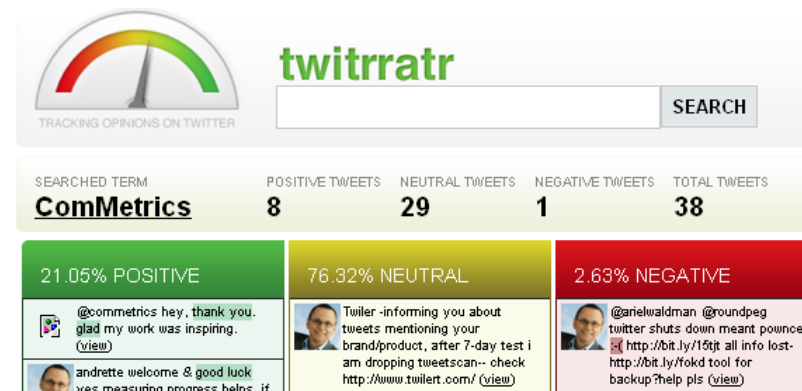
mrroland's Recent Opinions

Date Written	Review Title	Product / Topic	Product Rating	Review Rating
Oct 13 '12	Nancy Drew Mysteries Volume 31-The Ringmaster's Secret	Product info temporarily unavailable. Category info temporarily unavailable.	★★★★☆	Very Helpful
Oct 07 '12	Stole this "Bunny-licious thing" from luquillo...	Writer's Corner: General Non-Fiction in Member Center	n/a	Very Helpful
Oct 03 '12	Marvel's The Avengers (2012) Four Disc Combo Pack Blu-Ray/DVD/3D/Digital Copy	in Videos & DVDs	★★★★★	Very Helpful
Sep 27 '12	2011 Topps Allen & Ginter Baseball Card 53-Chipper Jones	2011 Topps Allen & Ginter Baseball Card 53-Chipper Jones - Atlanta Braves - In a Protective Displa. in Sport and Outdoor	★★★★★	Very Helpful
Sep 25 '12	The Hardy Boys Volume 2-The House on the Cliff	Product info temporarily unavailable. Category info temporarily unavailable.	★★★★★	Very Helpful

View more opinions by mrroland



- El objetivo del análisis de sentimientos es clasificar un fragmento de texto con una **valoración emocional positiva o negativa**
- La agregación de los sentimientos puede dar una idea de la percepción de la gente sobre una empresa, un producto o un tema
- Es un área de aplicación de la **minería de textos** y la **clasificación automática**. Se puede considerar también la influencia de las relaciones sociales
- Se realiza sobre tuits, entradas en un blog, opiniones en un portal, informes de productos en una web, etc.



- Los usuarios de foros on-line promocionan las marcas, un 79.2% ayuda a amigos en sus decisiones de compra (frente a un 47.6% de no usuarios). Un 65% aconseja sobre lo leído on-line (frente a un 35%)

<http://www.socialmediaexaminer.com/new-studies-show-value-of-social-media>

postrelease®

In thinking about the following activities, please indicate which, if any, you're involved in.

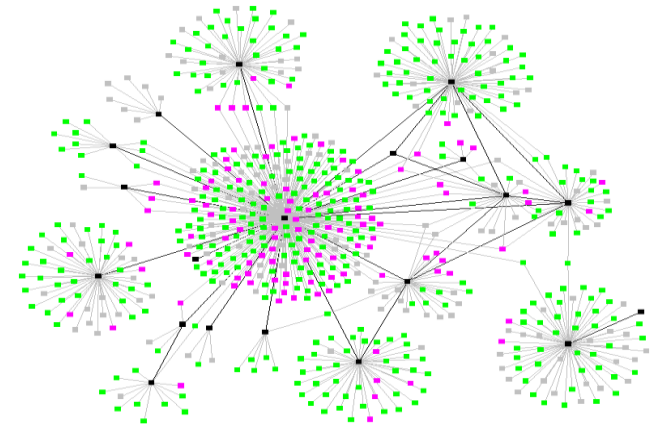
Responses (in order of frequency among overall population)	Those who contribute to online forums	Those who do NOT contribute to forums
Helping a friend or family member make a decision about a product purchase	79.2%	47.6%
Sharing advice (offline and in person) based on information that I've read online	65.0%	35.0%
Posting online ratings/reviews of products/services	66.0%	16.8%
Proactively recommending that someone make a particular purchase	57.7%	16.9%
Sharing links to articles about new products or with reviews of products	43.6%	12.0%
Attending an offline event or meet up where people with similar interests or who share the same hobby connect	35.9%	13.8%
Publishing a blog	29.6%	2.1%
Taking an active role in organizing an offline event or meet up for a group that met originally online	18.8%	2.4%
None of the above	0.0%	39.4%

Source: PostRelease.com/Synovate



- El contagio social permite modelar procesos de publicidad no convencional para campañas de marketing: **boca a boca**
- Analizando la influencia de los usuarios de la red se puede determinar las **semillas** a las que es necesario recompensar (cupones, promociones, etc.) para que hablen bien de un producto
- Este proceso se hace para provocar una **cascada** en la red con el menor presupuesto de campaña posible

negro: líderes de opinión; morado: influenciados
verde: no influenciados; gris: indecisos



“APLICACIONES” (7):

Condicionamiento del voto (Caso Cambridge Analytica)



Mayor empresa de datos del mundo (500M de personas)



Robert Mercer, mecenas de la nueva derecha estadounidense, invirtió 15M\$ en Cambridge Analytica (CA)



Steve Bannon estratega jefe de Donald Trump, contrata servicios de CA

Nigel Farage (UKIP)



Se ganó por menos del 2% de los votos



Christopher Wylie

"Explotamos Facebook para acceder a millones de perfiles de usuarios. Y construimos modelos para explotar lo que sabíamos de ellos y apuntar a sus demonios internos. Esa era la base sobre la cual la compañía [CA] se fundó"



17/03/2018

The New York Times

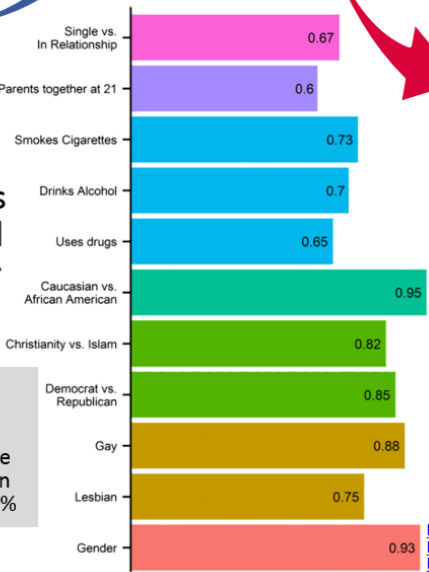


¡Qué escándalo, qué escándalo! He descubierto que aquí se juega

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell, Thore Graepel
PNAS April 9, 2013. 110 (15) 5802-5805
<https://doi.org/10.1073/pnas.1218772110>

A través de los “me gusta” de Facebook predicen atributos personales altamente sensibles. Datos de más de 58.000 voluntarios. Discrimina correctamente entre hombres homosexuales y heterosexuales en el 88%, demócratas y republicanos en el 85%



- A partir de los datos se segmenta al usuario por perfiles psicológicos
- Te diriges a quien es susceptible de entrar en teorías conspiratorias
- Fabricas webs y blogs con noticias falsas y se las muestras todo el tiempo
- Después ven la CNN y no hay nada de lo que ven todo el tiempo en internet, y piensan que la CNN esconde algo

En otro experimento en Facebook se consiguió aumentar en 340.000 el número de votantes en las elecciones de 2010 en EE UU



Ganó por 80.000 votos en estados clave

CRÓNICA

EL  MUNDO

La gurú de San Francisco que hizo ganar las elecciones a Mariano Rajoy



Isabelle wright @iamisabellew · 27 jun.

Celebrating tonight's victory in Spain with Prime Minister Rajoy #26j 💡 #afavor  #PartidoPopular 



← ↻ ❤️ 3 ⋮

- El PP contrató a la agencia de las campañas de Obama, Macri, Renzi y Cameron para preparar las elecciones
- La victoria 'le costó' al partido 290.000 euros

MARTIN MUCHA

 @Mart1nMucha

JAVIER NEGRE

 @javiernegre10

03/07/2016 19:02

En la fiesta de la calle Génova, **Isabelle** celebra. Sonrisa congelada. Es la confirmación de su talento y su esfuerzo. Llegó a España a finales de mayo. Sin saber español. Trabajando más de 14 horas al día en la zona noble del [Partido Popular](#). Que incluso la había mandado al hospital. Con los ojos rojísimos de tanto calcular y mirar estadísticas, el día de las elecciones es su Día D. El presidente [Rajoy](#) levanta los brazos en señal de victoria. [Jorge Moragas](#) - su jefe de gabinete en La Moncloa y responsable de la campaña de los *populares*, quien convenció a los líderes del partido de traer a España a la empresa para la que trabaja Isabelle: **The Messina Group (TMG)**- aprieta el puño. "Es un secreto. Deben mantenerlo en secreto", les dijo el consejero delegado de TMG, **Jim Messina**, a los altos cargos del PP. Y así fue. La llegada de Isabelle fue un secreto. Pero en el fondo es quien revolucionó la campaña electoral de Rajoy por dentro. La que permitió un éxito que sus rivales aún no se explican. **Crónica desvela cómo trabajaron y cómo consiguieron arrebatarnos los votos decisivos a los partidos de la competencia.** Cómo el PP venció con la asesoría del equipo que antes ayudó a ganar a Barack Obama y David Cameron. Isabelle escribe el 27 de junio, 1.00 de la mañana. "Increíble resultado para el PP en España esta noche. Los sondeos públicos predijeron 117 escaños para el PP, TMG predijo 135, resultado final 137". Euforia.

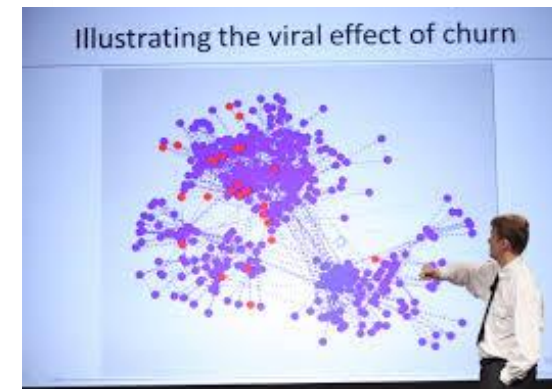
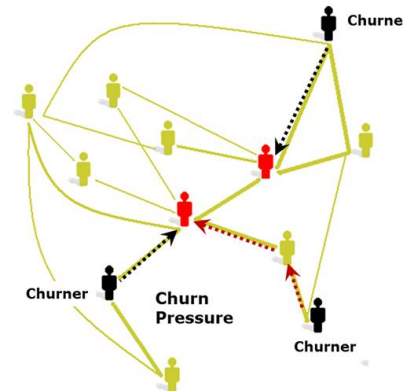
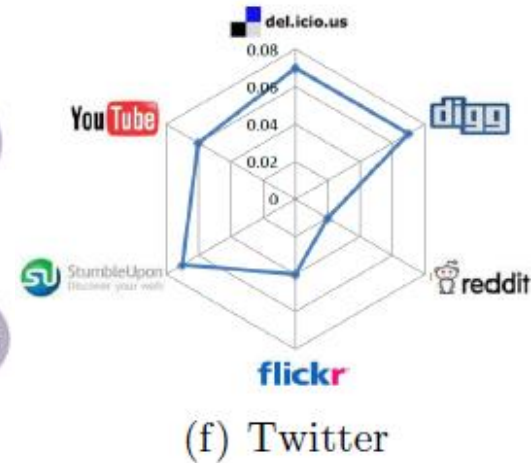
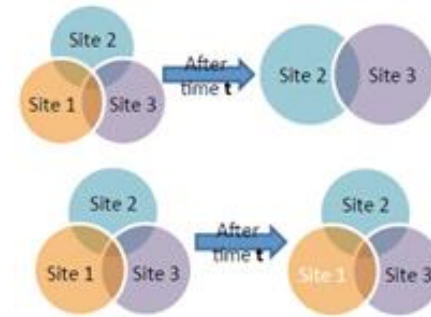
NETFLIX

‘El Gran hackeo’ un documental de Netflix sobre el polémico escándalo de Cambridge Analytica y Facebook



"El dilema de las redes sociales" de Netflix: 5 secretos de los dueños de las redes para engancharnos y manipularnos, según el documental

- Toda empresa (ej. Telecom.) sufre de la **migración voluntaria de consumidores** a la competencia
- La prevención de este abandono es un factor clave para su supervivencia. Se aplican técnicas de **clasificación automática**
- Las relaciones sociales son muy influyentes. Considerar la **estructura de red social/comunidades** y la **influencia** de los usuarios mejora la detección
- Se modela el *churn* como **procesos de contagio complejo en redes sociales**



1. Paradoja del Big Data:

- Los datos de medios sociales son “Big Data” pero no están distribuidos uniformemente
- Es habitual que cuando nos centramos en un individuo concreto nos falten datos y haya que obtenerlos por agregación

2. Obtención de muestras representativas:

- La información se obtiene de las APIs de las aplicaciones. ¿Cómo podemos saber si nuestra muestra de datos es representativa del conjunto completo?

3. Falacia de la eliminación de ruido:

- El preprocesamiento habitual en minería de datos no es adecuado: eliminar muchos datos ruidosos provoca la pérdida de información útil
- La eliminación del ruido es compleja y dependiente del problema

4. Dilema de la evaluación:

- Evaluar los resultados es complejo al no tener un *ground truth* en muchos casos

Referencias y Agradecimientos

Para diseñar los materiales de este tema, he hecho uso de material desarrollado por expertos en el área disponible en Internet:



- P. Gundecha, H. Liu. "Mining Social Media: A Brief Introduction".
Tutorial INFORMS 2012: <https://pubsonline.informs.org/doi/10.1287/educ.1120.0105>
- H. Liu. "Some Computational Challenges in Mining Social Media".
Tutorial ASONAM 2013: <http://www.public.asu.edu/~huanliu/papers/ASONAM13.pdf>
- K. Lerman. "Social Media. A Responsible User's Guide"
University of South California: <http://www.isi.edu/integration/people/lerman/talks.html>
- J. Leskovec. "Social Media Analytics". Tutorial ACM SIGKDD 2011
Stanford University: <http://snap.stanford.edu/proj/socmedia-kdd>

