Fundamentos de la Minería de Datos Web

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Introducción a la Minería de Datos Web

Introducción a la WWW

UCO ONLINE

World Wide Web

- La **World Wide Web** (la Red) se define como:

> Una iniciativa para la recuperación de información hipermedia de área amplia que pretende dar acceso universal a un gran espectro de documentos interconectados (*páginas web*) cuyos autores son millones de personas diferentes

- Se fundamenta en una estructura de documents **hipertexto** que permiten enlazar unos documentos con otros relacionados siguiendo **hiperenlaces**
  - El hipertexto se inventó en 1965 por Ted Nelson https://xanadu.com
  - Si permite otros elementos (o medios – videos, imágenes, …), se denomina **hipermedia**

World Wide Web

- La Web ofrece mecanismos de comunicación entre usuarios, por los cuales expresan su vision y opinion sobre cualquier aspect, y permite poner en contacto a personas de cualquier parte del mundo, creando una verdadera **Sociedad virtual**

- Para ello, se fundamenta en una gran red de computadoras de escala mundial que permite a los usuarios de uno o más nodos de esa red acceder a información almacenada en los demás nodos. A esta red se le conoce como **Internet**

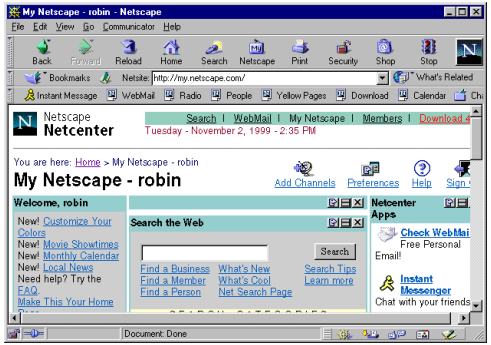
- El desarrollo de la web se implementa sobre un **modelo cliente-servidor**
  - El cliente es un programa que se conecta a una máquina remota (servidor), donde se almacenan los datos

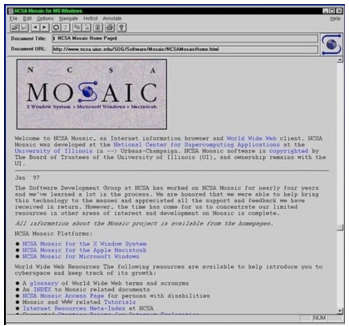
La historia de la WWW

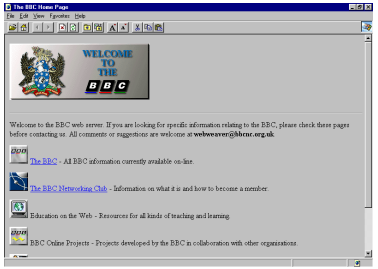
- Tim Berners-Lee (CERN, Suiza) acuñó el **término "World Wide Web"** y escribió el primer servidor, *httpd*, junto al primer cliente (*WorldWideWeb*)
- La **primera propuesta** enviada al CERN en 1989 fue rechazada.
  - En ella se planteaba un protocol sencillo por el que se solicitaba (*request*) información almacenada en un servidor remote a través de redes.
  - La información sigue un esquema de intercambio en un formato común para documentos hiperenlazados a otros documentos, que se devolvía al cliente (*response*)
  - La propuesta marcaba la arquitectura básica de la WWW: un sistema de hipertexto distribuido
  - Introdución **HTTP** (*HyperText Transfer Protocol*), **HTML** (*HyperText Markup Language*) y la **URL** (*Universal Resource Locator*)
- Recirculó la propuesta por el CERN y **en 1990 fue aceptada**

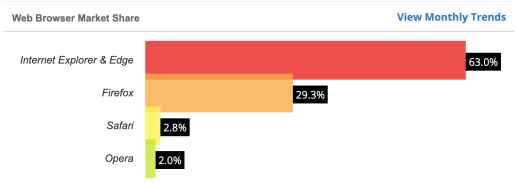
La historia de la WWW

- En 1993 se lanza **Mosaic for X**, el primer navegador con interfaz gráfica y puntero de ratón para pulsar (*click*) sobre los hiperenlaces
- Varios desarrolladores de Mosaic fundaron la empresa **Netscape Communications Corporation**, lanzando **Netscape Communicator** en 1997
- Del proyecto Mosaic surge la empresa Spyglass, propietaria de **Spyglass Mosaic**
- Microsoft intenta comprar fallidamente el navegador de AOL, tras lo cual compra Spyglass ☞ Internet Explorer (1995)

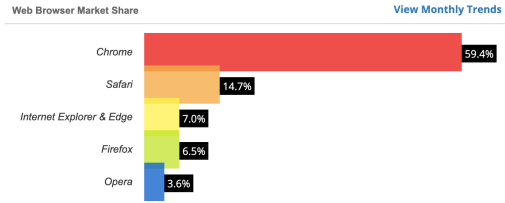




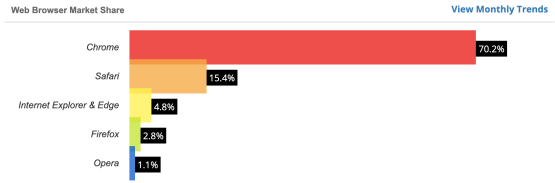




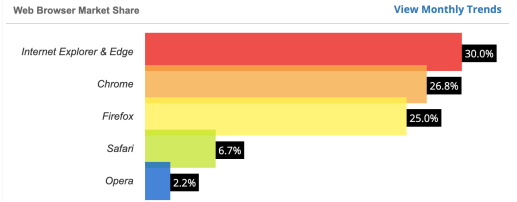
2008



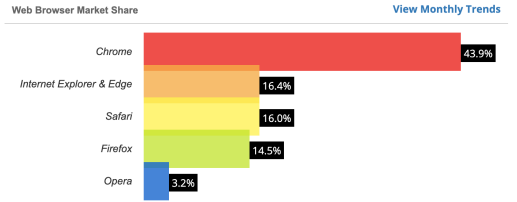
2018



2022



2012



2015

Evolución de los navegadores

La historia de la WWW

- Junto a los navegadores, también evolucionan los buscadores de información
- Los **motores de búsqueda** dan respuesta a la necesidad de los usuarios de encontrar información de forma ordenada y eficiente
- **Excite** es creado en 1993 por seis estudiantes de Stanford
- **Yahoo!** es creado en 1994 por Jerry Yang y David Filo como un listado de sitios favoritos (páginas amarillas), que ofrecía capacidad de búsqueda
- En 1998, Sergey Brin y Larry Page lanzan **Google**, bajado en un proyecto de investigación realizado en Stanford
- En 2004, **Yahoo!** ofrecen capacidad de búsqueda general tras comprar Inktomi
- En 2005, Microsoft lanza su motor de búsqueda **MSN**, en el que trabaja desde 2003

Así hemos llegado hasta hoy…

- En julio de **2020**, había más de **4.800 millones de usuarios de Internet** en el mundo (60% de la población – 90% en Europa y Norteamérica) *(Internet World Stats)*
- Creamos aproximadamente **2,5 quintillones de bytes de datos al día**
- Google, Facebook, Microsoft y Amazon **almacenan al menos 1.200 petabytes de información** *(Science Focus)*
  - 4 PB es la cantidad estimada de nuevos datos que genera Facebook al día *(Facebook Research)*
  - En Twitter, se publican 500 millones de tweets al día *(Statista)*
  - Google procesa más de 20 PB de datos al día
- Se realizan **1,2 billones de búsquedas en Google** al año *(Internet Live Stats)*
- A pesar de procesar sólo el 6,62% de todas las búsquedas en Estados Unidos, Bing obtiene casi **5.000 millones de dólares en ingresos publicitarios** *(StatCounter GlobalStats)*

Así hemos llegado hasta hoy…

- El 80% de los contenidos en línea está disponible en **sólo una décima parte de las lenguas**, lo que supone falta de diversidad. *(World Economic Forum)*
- Se calcula que la cantidad de datos en el mundo será de **175 zettabytes en 2025** *(IDC)*
- Para **2025**, se espera que la cantidad de datos generados alcance a nivel mundial los **463 exabytes cada día**
- En **2025**, habrá **75.000 millones de dispositivos** de Internet de las Cosas (IoT) en el mundo
- En **2030**, 9 de cada 10 personas mayores de seis años (**!**) serán digitalmente activas

Introducción a la Minería de Datos Web

Minería de Datos Web

UCO ONLINE

Características de la Web

- El crecimiento exponencial de datos en la Web la convierten en **el mayor conjunto de datos accesible** del mundo
- **Características únicas** de la Web que la convierten en un reto para la minería de información y extracción de conocimiento:
  - La cantidad de información está en constante crecimiento, así como su diversidad y amplitud
  - Existen todos los tipos de datos en la web, estructurados, semi- y no estructurados
  - La información es heterogénea, procedente de distintos autores, que pueden presenter la misma o similar información → la integración de la información a partir de multiples páginas es un reto
  - La información está enlazada dentro de un mismo o de diferentes sitios
    - En un mismo sitio, los hiperenlaces representan un **mecanismo de organización** de la información
    - Entre sitios distintos, los *hyperlinks* representan una **transmisión implícita de autoridad** a las páginas de destino

Características de la Web

- En la Web, **la información es ruidosa**:
  - El contenido de las página: secciones principales, enlaces de navegación, anuncios, notas, políticas de privacidad, etc. Solo una parte de esta información es útil, el resto es ruido
  - Un análisis de grano fino o la minería de datos requiren datos limpios (sin ruido)
  - La Web no tiene control de calidad, por lo que hay información de baja calidad, sesgada, errónea o falsa
- La Web permite **acceso a través de servicios**: compra de productos, envío de tweets, pago de facturas, etc.
- La Web **es dinámica**, lo que obliga a guardar los cambios y monitorizarlos
- La Web **es una Sociedad virtual**: no solo son los datos, también las relaciones y las interacciones entre personas, organizaciones o sistemas automatizados

¿Qué es la Minería de datos Web?

- La Minería de datos Web (Minería Web) **permite la extracción y descubrimiento de información y conocimiento a partir de la Web**
- Requiere pericia con los métodos clásicos de Minería de Datos para aplicarlos en distintas tareas de la Minería Web
  - La Minería Web no es realmente una aplicación de la Minería de Datos
  - Las características únicas de la Web (riqueza y diversidad de la información) han motivado que se creen métodos y algoritmos específicos
- Mientras que **la Minería de Datos utiliza datos estructurados**, generalmente de fuentes relacionales, hojas de cálculos, ficheros planos o conjuntos tabulares, la Minería Web utiliza otras fuentes diversas: **estructura de los hiperenlaces**, **contenido de las páginas**, **datos de uso**

Minería de la Estructura Web

- Descubre conocimiento útil, significativo y novedoso a partir de la estructura de los documentos Web:
  - Permite extraer patrones a partir de los hiperenlaces
  - Extraer información de la estructura de un documento HTML, JSON, XML, etc. (árboles)
- Fundamentada en la representación de la Web como un grafo dirigido, en el que las páginas son nodos y los hiperenlaces son las aristas
  - Es la base de los buscadores actuales
- Permite también el descubrimiento de comunidades online de usuarios que comparten intereses comunes
- La minería de datos tradicional no realiza estas tareas: no suele considerarse una estructura de enlaces en conjuntos de datos relacionales

Minería de Contenido Web

- Extrae información a partir del contenido de las páginas Web haciendo uso de técnicas similares a las empleadas por la Minería de Datos tradicional
- Los datos pueden estar estructurados, semi-estructurados o no estructurados
- La **recuperación de la información** (IR, *Information Retrieval*) permite realizar búsquedas en información estructurada y no estructurada, si bien no categoriza, filtra o interpreta documentos
- Algunas aplicaciones habituales:
  - Uso de técnicas de clasificación y agrupamiento para la categorización de páginas según sus tópicos de interés
  - Descubrimiento de patrones para extraer datos tales como la descripción de productos o los posts en foros, etc.
  - Extraer revisiones de clientes y sus opiniones para descubrir los sentimientos del consumidor

Minería de Uso Web

- Se refiere al descubrimiento de patrones de acceso de usuarios a partir de los datos contenidos en los *logs* de uso de los servidores Web (accesos, clicks, navegación, etc.) – **Útiles para descubrir las necesidades de una aplicación Web**
- Utiliza habitualmente algoritmos clásicos de minería de datos
- Las entradas de datos se producen a distintos niveles:
  - **Datos del servidor Web**: Trazas de usuarios (IPs, referencias a páginas, tiempo de acceso, etc.)
  - **Datos del servidor de aplicaciones**: Sobre todo los servidores de aplicaciones comerciales y tecnologías enfocadas a e-commerce permiten trazar un amplio espectro de eventos de negocio
  - **Datos a nivel de aplicación**: Eventos de la propia aplicación a partir de la interacción de sus usuarios.
- Un **factor clave** en minería de uso Web es el preprocesado de los flujos continuos de entrada (*clicks*) a partir de los logs para producer el conjunto de datos de entrada al algoritmo de minería

Minería de Uso Web

- Múltiples aplicaciones de interés:
  - Márketing personalizado para e-commerce
  - Lucha contra el terrorismo y actividades criminales
  - Mejora en la retención y relaciones con los clientes entendiendo mejor sus pautas y costumbres, perfilado de clientes y usuarios
  - Predicción del comportamiento de los usuarios
- Hay que observar algunos problemas asociados a la naturaleza de los datos que se manejan:
  - Violación de la **privacidad** de los usuarios
  - Puede llegarse a una **categorización engañosa**, ya que los usuarios se desindividualizan en términos de los patrones de clicks de ratón en lugar de por información personal, méritos o características
  - **Sesgo** en los algoritmos por motivos contrarios a la legislación (orientación sexual, religion, raza, etc.)
  - **Uso indebido** de los datos personales: comercialización de los datos

Comparación de tipos

	Minería de contenidos		Minería de estructura	Minería de uso
	Para búsqueda de información	Como conjunto de datos		
Tipos de datos	Estructurado y no estructurado	Semi-estructurados, contenido como base de datos	Estructura de enlaces	Interactividad
Fuente de datos	Documentos texto e hipertexto	Documentos hipertexto	Estructura de enlaces	Logs del servidor y cliente
Representación	• Bolsas de palabras y n-gramas • Conceptos y ontologías • Relacional	• Grafo etiquetado • Relacional	• Grafo	• Grafo • Tabla relacional
Método	Aprendizaje automático, estadística (incluyendo NLP)	Algoritmos propietarios, reglas de asociación	Algoritmos propietarios	Aprendizaje automático, estadística, reglas de asociación
Aplicaciones	Categorización, agrupamiento, reglas, patrones en texto	Subestructuras frecuentes, descubrimiento del esquema del sitio	Categorización, agrupamiento	Construcción del sitio, patrones de acceso, gestión del sitio web

Introducción a la Minería de Datos Web

Minería de Medios Sociales

UCO ONLINE

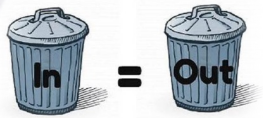
¿Qué es la Minería de Medios Sociales?

- Los medios sociales tratan de romper la frontera entre el mundo real y el mundo virtual
- Su análisis permite integrar teorías sociales con métodos computacionales para estudiar individuos (átomos sociales) que interactúan y forman comunidades (moléculas sociales)
- La **minería de medios sociales** es el proceso de representar, analizar y extraer patrones procesables a partir de datos de los medios sociales
  - La disciplina utiliza técnicas y metodologías de las ciencias computacionales, sociología, etnografía, estadística, optimización y matemáticas
  - Desarrolla las herramientas para representar formalmente, medir, modelar y extraer patrones significativos y útiles a partir de datos a gran escala de los medios sociales

Retos de la Minería de Medios Sociales

- Realizar minería sobre contenido generado por usuarios con relaciones sociales tiene aparejados una serie de **retos**:

  - **La paradoja del Big Data**. Los medios sociales tienen una ingente cantidad de datos. Sin embargo, al hacer zoom en individuos concretos para los que queremos hacer recomendaciones específicas, encontraremos que no hay tantos datos disponibles de un individuo particular
  - **La falacia de la eliminación de ruido**. En la minería de datos clásica se habla del preprocesado y limpieza de datos bajo la premisa de "*garbage in, garbage out*". En medios sociales, cada pequeña pieza de datos puede contener una gran cantidad de ruido:



    - Eliminar el ruido puede empeorar el problema, ya que la información valiosa suele estar escondida entre datos ruidosos
    - La eliminación se vuelve complicada en tanto en cuanto depende de la tarea a realizar

Retos de la Minería de Medios Sociales

- Realizar minería sobre contenido generado por usuarios con relaciones sociales tiene aparejados una serie de **retos**:
  - **Recopilación de suficientes ejemplos**. Las APIs son un mecanismo habitual para recopilar datos pero estas APIs suelen estar limitadas en tráfico/uso (p.ej. en cantidad de datos/día)
    - Es difícil conocer la distribución de la población para todos los datos existentes como para saber bien dónde "samplear"
  - **El dilema de la evaluación**. Un procedimiento estándar para evaluar es contar con una base de verdad (*ground-truth*), p.ej. dividiendo entre *train* y *test* el conjunto de datos: aprendemos con el *train* y tomamos el de *test* como base de verdad para evaluar el modelo. ¿Cuál es la base de verdad en un medio social? Y por otra parte: ¿cómo garantizamos la validez de los patrones sin una adecuada evaluación?

¡Gracias!

UCO ONLINE