## Grado en Ingeniería Informática

# Explotación de la Información


**Antonio Ferrández Rodríguez**

UNIVERSIDAD DE ALICANTE

*Grupo de Procesamiento del Lenguaje y Sistemas de Información*

# Explotación de la Información

## Big Data y la información no estructurada

Google Jobs — Search all jobs

About Google › Jobs › Search all jobs › Software Engineer

All Jobs
My Applications
Starred

RECENTLY VIEWED View all
Software Engineer, University Gra…
London, UK

**Software Engineer**
Munich, Germany
Software Engineering · Full-time
Know someone who would be interested?

APPLY NOW | Find connections | Sign in to see your connections at Google

Google's software engineers develop the next-generation technologies that change how millions of users connect, explore, and interact with information and one another. Our ambitions reach far beyond just Search. Our products need to handle information at the the scale of the web. We're looking for ideas from every area of computer science, including information retrieval, artificial intelligence, natural language processing, distributed computing, large-scale system design, networking, security, data compression, and user interface design; the list goes on and is growing every day. As a software engineer, you work on a small team and can switch teams and projects as our fast-paced business grows and evolves. We need our engineers to be versatile and passionate to tackle new problems as we continue to push technology forward.

With your technical expertise you manage individual projects priorities, deadlines and del design, develop, test, deploy, maintain, and enhance software solutions.

La paradoja del Big Data viene de que llega un momento en que somos conscientes de que generamos una gran cantidad de datos que pueden sernos útil (una vez convertidos en información) para nuestro negocio; sin embargo, tal y como se almacena, no podemos usarla para nada por tratarse de **información no elaborada o no estructurada**.

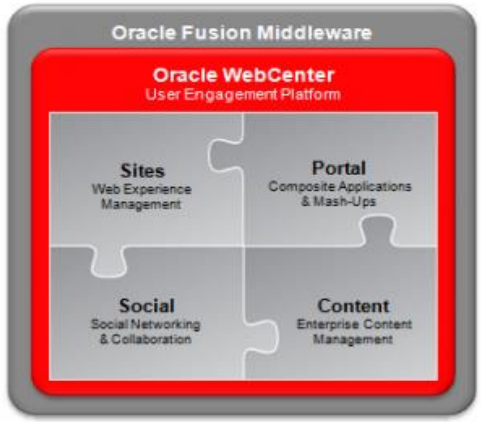

**EL PRODUCTO DE ORACLE PARA MANEJAR LA INFORMACION NO ESTRUCTURADA**

### Minimum qualifications

- BS in Computer Science or equivalent practical experience.
- Programming experience in C, C++ or Java.
- Experience with large systems software design and development, with demonstrable knowledge of UNIX/Linux.
- Skills in AJAX, JavaScript or Python, database design and SQL.

### Preferred qualifications

- MS or PhD in Computer Science or related technical field.
- Strong foundation in computer science, with strong competencies in data structures, algorithms and software design.
- Strong object-oriented skills.
- Skills in AJAX, Python or JavaScript, database design, SQL and/or knowledge of TCP/IP and network programming.



# Programa. Módulos

1. **Introducción a la Explotación de la Información.**
2. **Recuperación de Información (*Information Retrieval*).**
3. **Extracción de Información en Bases de Datos (*Information Extraction*).**
4. **Clasificación y Agrupamiento de Información.**

# Objetivos

- **Que el alumno conozca:**
  - Tipos de información: estructurada (bases de datos) vs. no estructurada (textual, vídeo, imágenes, sonido, etc.)
  - Las diferentes formas de representación, almacenado y acceso de la información
  - Concepto y tipos de minería de datos
  - Las aplicaciones de minería de datos no estructurados más importantes en la actualidad: recuperación, extracción y clasificación de información

# Objetivos

- **Que el alumno conozca (cont.):**
  - La arquitectura tradicional de las aplicaciones de minería de datos no estructurados
  - Los tipos de datos habituales utilizados en las aplicaciones de minería de datos no estructurados, sus representaciones más comunes, su eficiencia y su utilidad
  - Las técnicas de aprendizaje automático computacional utilizadas en las aplicaciones de minería de datos no estructurados
  - Herramientas específicas de programación

# Objetivos

- **Que el alumno comprenda:**
  - La necesidad actual de las empresas de analizar la información no estructurada (actualmente alrededor del 70% de la información de una empresa)
  - La necesidad de adaptar la representación interna de la información a los requerimientos de la aplicación a resolver

# Objetivos

- **Que el alumno sea capaz de:**
  - Distinguir entre representaciones alternativas de la información y razonar sobre la solución escogida en cuanto a coste computacional
  - Organizar un determinado volumen de información de la forma más racional posible para optimizar su posterior acceso
  - Analizar, diseñar y construir aplicaciones de minería de datos no estructurados

# Metodología de trabajo

- **Clases teóricas. Materiales:**
  - Transparencias (campus virtual).
  - Material adicional (artículos, monografías, demos, etc.) en la ficha de la asignatura:
    - http://cv1.cpd.ua.es/ConsPlanesEstudio/cvFichaAsiEEES.asp?wCodEst=C203&wcodasi=34035&wLengua=C&scaca=2015-16

# Metodología de trabajo

- **Clases prácticas:**
  - Lenguaje de programación: C++
  - Enunciados y materiales de prácticas en el apartado de materiales del campus virtual
  - Las prácticas se realizarán de forma individual
  - Laboratorio L01 de la EPS

# Metodología de trabajo

- **Evaluación:**
  - Continua
  - No habrá examen final
  - En caso que alguien no opte por la evaluación continua se le valorará en función de las memorias de trabajos y prácticas con ordenador, en su defensa presencial
  - La asistencia a las clases prácticas y teóricas será obligatoria con un máximo de 3 faltas permitidas sin justificar

# Metodología de trabajo

- **Evaluación:**
  - Se guardan partes aprobadas para la convocatoria de julio. No se pide nota mínima para promediar:
    - 50% Teoría
      - 10% Asistencia participativa a clase
      - 15% Revisión presencial práctica 1
      - 15% Revisión presencial práctica 2
      - 15% Revisión práctica 3
      - 45% Controles tipo test y ejercicios de clase
    - 50% Práctica
      - 30% práctica 1
      - 50% práctica 2
      - 20% práctica 3
    - 2 puntos adicionales para los que hagan los trabajos optativos

# Tipos de información

- **La época actual de la sociedad de la información:**
  - La ventaja:
    - La disponibilidad de dicha información
  - El problema:
    - La localización de la información que interesa a cada persona
    - La explotación de esa información:
      - Extracción “automática” de la información requerida
      - Almacenamiento de esa información para posteriores procesos
      - Generación de nueva información

# Tipos de información

- **Información:**
  - Conjunto organizado de datos que tienen un significado
  - Dato:
    - Representación simbólica (numérica, alfabética o algorítmica) de un atributo o característica de una entidad
    - Para ser útiles, los datos deben convertirse en información para ofrecer un significado, conocimiento, ideas o conclusiones
- **Tipos de información:**
  - Estructurada: base de datos
  - Desestructurada o no estructurada: texto, vídeo, sonido, etc.

# Tipos de información

- **Importancia de la información no estructurada:**
  - Existe un vasto crecimiento de la información disponible:
    - En forma de datos estructurados (bases de datos)
    - Como datos no estructurados (p. ej. emails o documentos) y fuentes externas (la Web):
      - Alrededor del 70% de la información de una empresa
      - Las empresas consideran critica la gestión de ese tipo de información (gustos y preferencias de los consumidores)
      - Se necesitan herramientas que sean capaces de analizar ese tipo de información (**minería de sentimientos** u **opiniones**, también conocido como **análisis de sentimientos/opiniones**)

# Tipos de información

- **Minería de sentimientos u opiniones:**

EL PAIS

SOCIEDAD

## Cómo mantener la reputación 'on line'

El 70% de las empresas con presencia en Internet monitoriza todo lo que se dice de ellos
Los directivos deciden estrategias en función de lo que se habla en las redes sociales

# Tipos de información

## Monitorizan sus marcas el 70% de las empresas que están en Internet

En 2012 las empresas entendieron que lo importante no era tanto conseguir el máximo de seguidores sino prestar atención a sus comentarios. Vieron que podían obtener ventajas haciéndolo. Hacer un seguimiento de los textos y comentarios sobre sus marcas que se publican en los periódicos *online*, en los blogs y en las redes sociales les permite intentar poner freno a posibles crisis derivadas de quejas de los clientes o a los bulos que se propagan (y la rapidez es clave en este asunto). En algunos casos además les ayuda a decidir aumentar o disminuir la producción de ciertos productos. También les facilita cuantificar el coste de una campaña o nueva gama de producto tras su lanzamiento. Y lo último: les ayuda a decidir sus estrategias en función de lo que se habla en la Red.

Ahí va un caso concreto: en julio de 2011 en las búsquedas de los ciudadanos en Google se dio un vuelco entre dos palabras: sexo y amor. Antes de esa fecha siempre ganaba el sexo. Pero desde ese verano le gana invariablemente el amor. Unos meses más tarde, en marzo de 2012, ING Direct lanzaba una campaña en la que varias personas le declaraban su amor a su pareja. “Por primera vez la estrategia empieza a salir de la monitorización”, dice Fernández, de Annie Bonnie. “Y empieza a haber teorías que dicen que los comentarios de la gente van a influir en las cotizaciones de las empresas en Bolsa”.

# Tipos de información

Las empresas grandes reciben una cantidad ingente de comentarios. En 2013, Repsol —por ejemplo— obtuvo 700.000 menciones. Sobre sus inversiones en Latinoamérica, su equipo de motos, su guía gastronómica. 2.500 menciones un día tranquilo. En sus oficinas de Méndez Álvaro de Madrid tres personas se dedican a tiempo completo a monitorizar, más una cuarta que lo compagina con otras tareas. Natalia Villoria, subdirectora de Reputación Corporativa, explica que tras un primer intento de llevar todo el control, han externalizado el 40% del proceso por la precisión que reciben al ponerlo en manos de expertos.

Además de contratar a agencias utilizan el *software* de la empresa francesa Augure, que en 2010 compró a la catalana iMente. Uno de sus productos estrella es su herramienta Augure Influencers, que analiza los comentarios de todas las personas que consideran influyentes. En ella están gente como Wally López y famosos, catedráticos universitarios y casi todos los periodistas del país por que lo que escriben o hacen lo ven muchos ojos. Augure elabora una ficha por cada persona en la lista destacando los temas de los que hablan con más frecuencia.

A pesar del avance, todos coinciden en que el sector está todavía muy inmaduro. Manuela Battaglini, de la Asociación de Responsables de Comunidades Online, cree que hay mucho camino que recorrer.

# Tipos de información

- **Aplicaciones útiles para la empresa que necesiten el uso de información no estructurada:**
  - Buscadores Web: Google, Bing, Yahoo, etc.
  - Dossier de prensa: conocer, presentar y agrupar las noticias de una empresa
    - http://intime.dlsi.ua.es:8080/dossierct/index.jsp?lang=es&status=probable&date=15-01-2013&newspaper=
  - Minería de opiniones: saber lo que opina los consumidores de una empresa (p.ej. en Twitter)
    - http://gplsi.dlsi.ua.es/gplsi11/content/gplsi-social-observer
  - Responder a la petición de información:
    - Comparar los precios de los productos que vende una empresa con los de los competidores

# Minería de datos

- **Características (*Data Mining*):**
  - Consiste en la extracción no trivial de información que reside de manera implícita en los datos
  - Dicha información era previamente desconocida y podrá resultar útil para algún proceso
  - Aborda la solución a problemas de predicción, clasificación y segmentación
  - Campos de estudio: Inteligencia Artificial y el Análisis Estadístico

# Minería de datos

- **Características (cont.):**
  - Conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las **bases de datos** (almacenes de datos).
    - Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos para la **toma de decisiones**
  - Actualmente está cobrando una importancia cada vez mayor la **minería de datos no estructurados** como información contenida en ficheros de texto **(minería de textos)**, en textos semi-estructurados **(**HTML, XML, **minería del marcado** o ***Markup Mining*****)**, en Internet **(minería web),** en vídeos o imágenes **(minería multimedia)**, etc.

# Minería de datos

- **Proceso típico de minería de datos:**
  - Fase de integración y recopilación:
    - Selección del conjunto de datos
    - Análisis de las propiedades de los datos
  - Fase de transformación o preprocesamiento de los datos de entrada
  - Fase de minería de datos mediante la técnica seleccionada:
    - Se construye el modelo predictivo, de clasificación o segmentación
  - Fase de evaluación e interpretación de datos
  - Fase de difusión, uso y monitorización

# Minería de datos

- **Técnicas de minería de datos:**
  - Redes neuronales
  - Regresión lineal
  - Árboles de decisión
  - Modelos estadísticos
  - Agrupamiento o *clustering*

# Minería de datos

- **Ejemplos de uso de la minería de datos:**
  - Hábitos de compra en supermercados:
    - Se detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza
    - Padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano
    - Decisión: el supermercado incrementó sus ventas de cerveza colocándolas próximas a los pañales
  - Comportamiento en Internet:
    - Según perfiles de usuarios, sus clicks en la web de la empresa, etc.
  - Relación características que hacen probable el padecer una enfermedad (p. ej. Alzheimer):
    - Antecedentes familiares, hábitos del individuo, profesión, etc.

# Minería de datos

- **Tipos de aplicaciones de Minería de Datos no estructurados:**
  - Módulo 2: Recuperación de información: obtener información desde datos textuales (p. ej. la Web)
  - Módulo 3: Extracción de Información en Bases de Datos
  - Módulo 4: Clasificación y Agrupamiento de Información

# Minería de datos

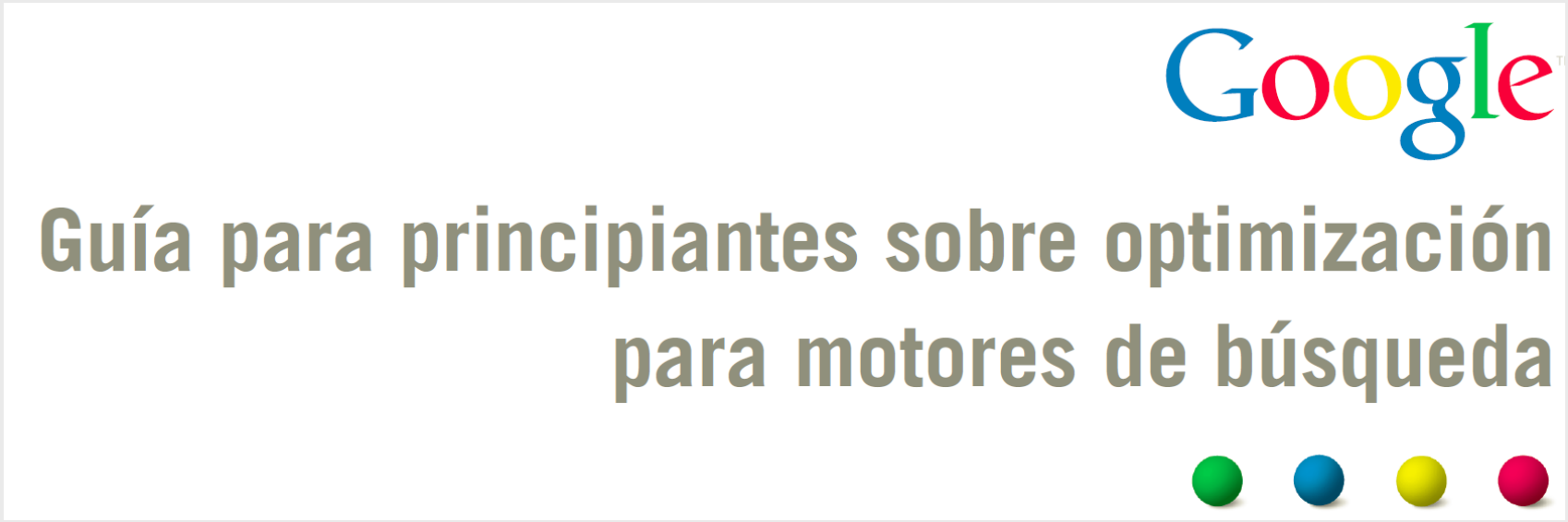
- **Tipos de aplicaciones de Minería de Datos no estructurados:**
  - Módulo 2: Recuperación de información
    - Unidad 4: Introducción a los sistemas de Recuperación de Información: Google, Bing, Altavista, etc.
    - Unidad 5: Arquitectura de los sistemas de recuperación de información
    - Unidad 6: Fase de recopilación de documentos
    - Unidad 7: Fase de indexación. Módulo de segmentación de palabras
    - Unidad 8: Fase de indexación. Módulo de segmentación de palabras. Cálculo de complejidades
    - Unidad 9: Fase de indexación. Módulo de filtrado de palabras
    - Unidad 10: Fase de indexación. Módulo de almacenamiento. Estructuras de datos utilizadas: matrices, matrices dispersas, índices invertidos, tablas hash, trie, …

# Minería de datos

- **Tipos de aplicaciones de Minería de Datos no estructurados:**
  - Módulo 2: Recuperación de información (cont.)
    - Unidad 11: Fase de búsqueda. Modelos booleano, vectorial y probabilístico
    - Unidad 12: Fase de búsqueda. Modelos basados en pasajes
    - Unidad 13: Fase de búsqueda. Expansión de la pregunta
    - Unidad 14: Fase de presentación de resultados
    - Unidad 15: Evaluación de los sistemas de Recuperación de Información
    - Unidad 16: Sistemas de *Question Answering* o Búsqueda de Respuestas
    - Unidad 17: Sistemas de Recuperación de Información multimedia

# Minería de datos

- Módulo 2: Recuperación de información (cont.)




- https://static.googleusercontent.com/media/www.google.com/es//intl/es/webmasters/docs/guia_optimizacion_motores_busqueda.pdf
- **¿Necesitas optimización en buscadores (SEO)?**
  - **https://support.google.com/webmasters/answer/35291?hl=es**

# Minería de datos

- Módulo 2: Recuperación de información (cont.)
  - Vídeos de Olga Carreras sobre posicionamiento web:
    - Conceptos básicos: https://www.youtube.com/watch?v=YpJcR1UWqMY
    - Técnicas básicas (1): https://www.youtube.com/watch?v=YpJcR1UWqMY
    - Técnicas básicas (2): https://www.youtube.com/watch?v=YpJcR1UWqMY

# Minería de datos

- Módulo 2: Recuperación de información (cont.)

TECNOLOGÍA

## Google tiene un nuevo reto: la inteligencia artificial

- El gigante tecnológico ha unido fuerzas con la Universidad de Oxford.
- El reto, aplicar la inteligencia artificial en reconocimiento de imágenes y procesamiento del lenguaje natural.
- Google ya adquirió la empresa DeepMind, en enero, por 650 millones de dólares.

Google adquirió la startup DeepMind en enero por 650 millones de dólares. Muchos creen que su objetivo es desarrollar equipos que **piensan como seres humanos** y poder aplicarlo a su negocio de búsquedas.

# Minería de datos

- **Tipos de aplicaciones de Minería de Datos no estructurados:**
  - Módulo 3: Extracción de Información en BD
    - Unidad 18: Introducción a los sistemas de Extracción de Información
    - Unidad 19: Arquitectura de los sistemas de Extracción de Información
    - Unidad 20: Módulo de análisis léxico
    - Unidad 21: Módulo de análisis sintáctico
    - Unidad 22: Módulo de reconocimiento de entidades
    - Unidad 23: Módulo de análisis semántico
    - Unidad 24: Módulo de resolución de correferencias
    - Unidad 25: Módulo de análisis contextual
    - Unidad 26: Módulo de extracción, rellenado y almacenamiento de plantillas
    - Unidad 27: Ejemplos de sistemas de EI

# Minería de datos

- **Tipos de aplicaciones de Minería de Datos no estructurados:**
  - Módulo 4: Clasificación y Agrupamiento de Información
    - Unidad 28: Sistemas de clasificación de información
    - Unidad 29: Clasificación basada en vocabulario
    - Unidad 30: Clasificación utilizando árboles de decisión
    - Unidad 31: Clasificación utilizando sistemas de reglas
    - Unidad 32: Problema del overfitting. Sistemas de poda
    - Unidad 33: Sistemas de agrupamiento de información
    - Unidad 34: Sistemas de agrupamiento de información en la Recuperación de Información
    - Unidad 35: Sistemas de agrupamiento de información particionales (algoritmo k-mean)
    - Unidad 36: Sistemas de agrupamiento de información jerárquicos