# MINERÍA WEB.

Etzioni definió la Minería Web como "*el uso de técnicas de Minería de Datos para descubrir y extraer información automáticamente desde el World Wide Web*"

El proceso de Minería Web consta de una serie de subtareas:

1. **Descubrimiento de las fuentes** → consiste en localizar la información que se encuentra alojada en los distintos documentos y servicios que ofrece la web.
   La recuperación de dicha información es tanto la perteneciente a fuentes textuales cono cualquier otro tipo de documento de hipertexto, es decir, se recupera información como pdf, xml, html, correo electrónico…
   Para llevar a cabo esta actividad se utilizan los "Índices de documentos web", es decir los llamados Buscadores, que recuperan los documentos relevantes a través de los procesos de recuperación de información (Information Retrieval).
   Se puede mencionar como buscadores, los siguientes:
   a. Google (http://www.google.es)
   b. AltaVista (http://www.altavista.com)
   c. Excite, que comercializa con WebCrawler (http://www.webcrauler.com)
   d. Lycos (http://www.lycos.com)
   e. Yahoo (http://www.yahoo.com)

2. **Selección y preprocesado de la información** → consiste en extraer automáticamente la información desde las fuentes.
   Para poder extraer la información necesaria, existen algunos sistemas que extraen la información a través de las FAQ (preguntas más frecuentes)

3. **Generalización** → en esta etapa se utilizan técnicas de Minería de Datos adaptadas a la Minería Web, como por ejemplo, reglas de asociación o agrupamiento, procesos de recuperación de información (Information Retrieval)…

4. **Análisis** → consiste en el desarrollo de técnicas y herramientas que permitan que las personas puedan utilizar la información que ha sido obtenida a través de las técnicas de Minería de Datos.
   Parra ello se utilizan técnicas estadísticas y de visualización

# MINERÍA WEB. DISCIPLINAS

Algunas disciplinas relacionadas con el proceso de Minería Web son:

1. **Procesos de Recuperación de Información** (Information Retrieval) → sus principales actividades son:
    a. la indexación de texto
    b. búsqueda de documentos útiles en una colección
    c. modelización de documentos
    d. categorización de documentos
    e. clasificación de documentos
    f. Visualización de filtrados
    g. Interfaz de usuario

    En definitiva, todas aquellas actividades que tengan que ver con la selección de documentos relevantes

2. **Procesos de Extracción de Información** (Information Extraction) → su principal objetivo es la extracción de hechos relevantes a partir de documentos.
    Existen dos tipos de extracción de información:
    a. A partir de textos no estructurados: son textos escritos en lenguaje natural y que requieren la necesidad de procesados lingüísticos como:
        - Análisis sintáctico
        - Análisis semántico
        - Análisis del discurso
    b. A partir de datos semi-estructurados: utilizan etiquetas html, es decir, requieren el uso de la meta-información.
        Cabe destacar el uso de técnicas de Minería de Datos ya que no se suelen utilizar sistemas manuales

# MINERÍA WEB. CLASIFICACIÓN

La clasificación que se puede realizar de la Minería Web es la siguiente:

1. **Minería del Contenido de la web** → describe la información útil que poseen los documentos que se encuentran en la web, tanto el contenido textual como el contenido gráfico, pasando por las imágenes, audio o video. Su origen se encuentra en el procesamiento del lenguaje natural y en la recuperación de información

   En relación con la recuperación de información, la minería del contenido de la web, se mejora la información que los buscadores ofrecen a los usuarios que demandan dicha información.

   En relación con las bases de datos, la minería de contenido de la web permite que el usuario realice preguntas mucho más sofisticadas en el caso de realizar una búsqueda con palabras clave

2. **Minería de la Estructura de la web** → analiza la estructura más profunda de los enlaces en la web, es decir, la relación entre los diferentes sitios web

   Un buen ejemplo del uso de Minería de la estructura de la web, es la ofrecida por los distintos buscadores como Google o AltaVista, que ofrecen dentro de sus herramientas de búsqueda los llamados "PageRank" (**MyGoogle PageRank:** http://www.mygooglepagerank.com/pagerank.php)

3. **Minería del Uso de la Web** → analiza la información sobre los accesos web disponibles en los servidores web, esto es, datos que derivan de la interacción del usuario con la web

   Para ello, de la interacción del usuario con la web se extraen patrones de comportamiento para conocer sus preferencias de navegación con el fin de mejorar las páginas adaptando la interfaz del sitio web en cuestión.

   Dos de las aplicaciones más importantes son:

   a. Patrones de navegación
   b. Perfiles de usuario