Acceso y Recuperación de datos en la Web

En los últimos años, Internet ha evolucionado de un almacén de documentos a un repositorio de datos. Los procesos para aprovechar esta web de datos son analizados en este tema. A diferencia de las páginas web, en la web de datos la necesidad de valorar la calidad de los datos para un fin determinado, saber interpretar los datos a capturar y manipularlos para optimizar su rendimiento es básica.

En este tema veremos el ciclo de vida asociado a la captura y gestión de estos datos. Este ciclo de vida es, en esencia, similar a otros ciclos de vida asociados a la gestión de información. Esto es, ir asignando relaciones y restricciones a los datos, de manera que se pueda obtener nueva información. Estos datos por sus características, tienen similitudes con varias disciplinas, como son: BigData, Web Semántica y los sistemas de gestión de la información.

En este tema se verán las distintas etapas de estos datos: planificación, captura de datos y su calidad, limpieza y normalización, y enriquecimiento e integración con otras fuentes. Las siguientes fases, es decir el análisis estadístico, quedan fuera de los objetivos del tema.

Las distintas serializaciones, en la que se pueden encontrar los datos que encontramos en la Web, serán ejemplificadas en la presentación.

Especial énfasis será puesto a la limpieza de datos. En esta limpieza y en la normalización de datos se pueden aplicar distintas técnicas: algoritmos de similitud de cadenas (p.e. distancia de Levenshtein), similitud fonética, validación mediante recursos externos, crowdsourcing, análisis estadístico, etc.

J.Morato y V.Palacios

J.Morato y V.Palacios