uc3m | Universidad Carlos III de Madrid

Sistemas de Recuperación de Información: Crawlers

Los motores de recuperación constituyen los principales sistemas de recuperación en la Web. En este tema se verá la arquitectura básica de un motor de recuperación en la Web. Concretamente, el tema se centra en los Crawlers o arañas.

Un crawler en una aplicación diseñada para visitar recursivamente la Web. Los crawler se diseñan para visitar hiperenlaces de forma sistemática, de forma eficiente e, idealmente, aplicando una política de acceso respetuosa con los deseos de los propietarios de los sitios web.

Se muestran, además, distintos conceptos básicos relacionados con los crawlers. Como los ficheros robots.txt y sitemaps.txt.

El crawler es el paso previo al almacenamiento de la información. Esto se realiza, típicamente, descargando el texto y metadatos de la página y almacenándolos en una base de datos, para su posterior consulta.

Finalmente, se facilitan enlaces, en distintos lenguajes de programación, para implementar un motor de recuperación.

J.Morato y V.Palacios