



Resumen Recuperación de la Información en Internet

La presente **Unidad** trata de las aplicaciones involucradas en la **Recuperación de la Información en Internet**. La Web constituye actualmente el principal repositorio de información. La utilidad de este repositorio está directamente relacionada con la capacidad de recuperar la información contenida en el mismo.

Inicialmente hemos visto los **Sistemas de Recuperación en Internet**. Se ha revisado la **tipología** de los sistemas de recuperación en Internet: **Directorios, Motores Web y Metabuscaores**.

Posteriormente, ya centrados en **Motores Web** se ha visto cual es su **diseño, funcionamiento y componentes principales**.

La visión general de los buscadores se ha completado con la descripción y análisis de limitaciones de de varios **buscadores especiales**, según diferentes perspectivas: Cobertura temática, Adecuación a la Web Semántica y a la Web Social y Diseño del interfaz

En relación a los motores, un tema de especial relevancia por importancia comercial es el **Posicionamiento**. Se trata de los factores que afectan a la ordenación de resultados al ejecutar una consulta en un buscador Web. En principio el algoritmo de posicionamiento no es público, si bien mediante pruebas se ha identificado un gran número de parámetros que afectan al posicionamiento. Debido a su importancia nos hemos centrado en Google. Ya que acapara el 70% de las búsquedas en Internet en el contexto mundial.

Sobre los **parámetros que afectan directamente al posicionamiento** unos son específicos de los términos de consulta y otros son genéricos. No todos los parámetros tienen el mismo peso en el cálculo final de la relevancia. Los criterios son de varios tipos, destacan:

- Los de ubicación de los términos de consulta
- Los relacionados con los enlaces
- Relacionados con la tipografía
- Relacionados con la URL
- Los factores relacionados con parámetros temporales
- Además existen otros factores relacionados con el contenido, ubicación y visitas.



SISTEMAS AVANZADOS DE RECUPERACIÓN DE INFORMACIÓN

Existen además un conjunto de **factores indirectos**. Estos factores son aquellos que afectan de forma genérica a la utilidad que determinado sitio Web puede tener para responder a determinado tema. Más en concreto son factores asociados a criterios como la Usabilidad, Accesibilidad, Interoperabilidad, Interés de los contenidos o la credibilidad. De forma genérica un mayor acceso permite una mayor probabilidad de ser enlazado y por tanto mejorar el posicionamiento.

Existe sin embargo el peligro de que el buscador Web nos penalice por intentos de “engañar” al buscador. A estos factores los denominados **penalizaciones**.

Por último, hay que destacar la importancia que tienen las **herramientas SEO (Search Engine Optimization)**. Es decir las herramientas que facilitan el estudio de los parámetros que afectan al posicionamiento de un Site.

Por último se ha visto la **metodología de evaluación de los sistemas de recuperación**. Se han revisado los aspectos a considerar relativos a: diseño de interfaz, capacidad de consulta, actualización y tamaño de la base de datos relacionada, tipología de los documentos indexados y presentación de resultados. El estudio de la evaluación se complementa mediante el estudio de las **medidas estándar de recuperación de la información**.

De entre las medidas de recuperación se han presentado entre otros los conceptos de **recall** y **exhaustividad**, gráficos para mostrarlos y las diferencias dependientes de la **agrupación** o no de las consultas o de las medidas de **recall** y **exhaustividad**.

Se han visto además distintos **corpus de evaluación** y **competiciones de recuperación de documentos**.