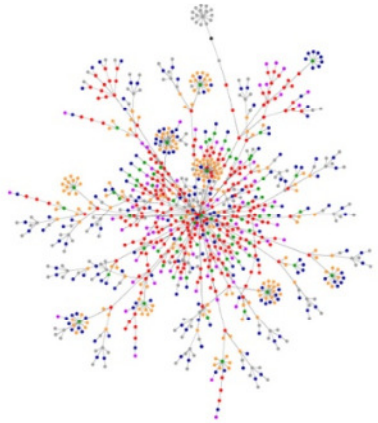


Web Crawling

Recuperación de la Información

Rodrigo A. Bartels

Sistema de Estudios de Posgrado
Universidad de Costa Rica
rodrigo.bartels@ucr.ac.cr

Introducción

- Cuando se desea crear un Sistema de RI se debe definir el Universo de Información:

Ubicación

Tipo

- Estos dos elementos definen el tipo de infor e d re ec y tipo de consultas que puede hacer el usuario.

Crawling

- Un *“crawler”* o *araña* es una herramienta computacional que navega un espacio de información de forma metódica y automatizada.

- También son conocidos como a*nts*, indexadores automáticos, *bots*, w*orms*, *Web spider*, *Web robot*.

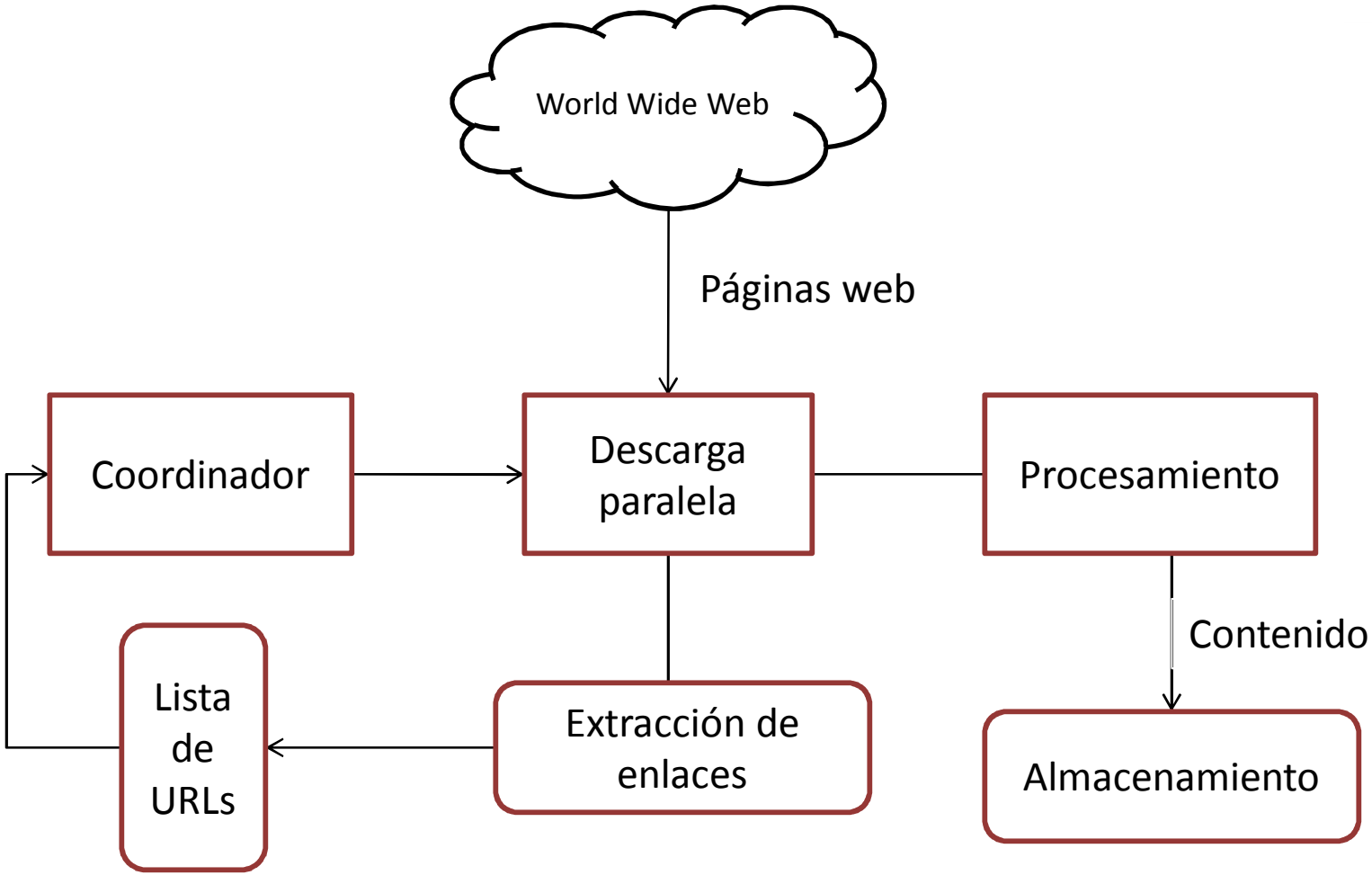
Web Crawler

- Un *web crawler* se dedica a navegar la Web y almacena la información que va recolectando.

- De forma general, el proceso empieza con una lista de URLs, llamados semillas.

- Al visitar cada URLs la aplicación identifica todos los hipervínculos contenidos en él, y los agrega a la lista de URLs por visitar.

Indexación del Web



Políticas de indexación

- Selección
  - Restricción de tipos
  - *Focused o topical Crawling*
  - El Web Oculto
- Re-visita
  - Uniforme
  - Proporcional

Políticas de indexación

- Cortesía
  - Recursos de Red
  - Sobrecarga de servidores
  - Ataques de denegación de servicios
- Paralelismo
  - Distribución del espacio
  - Proporcional

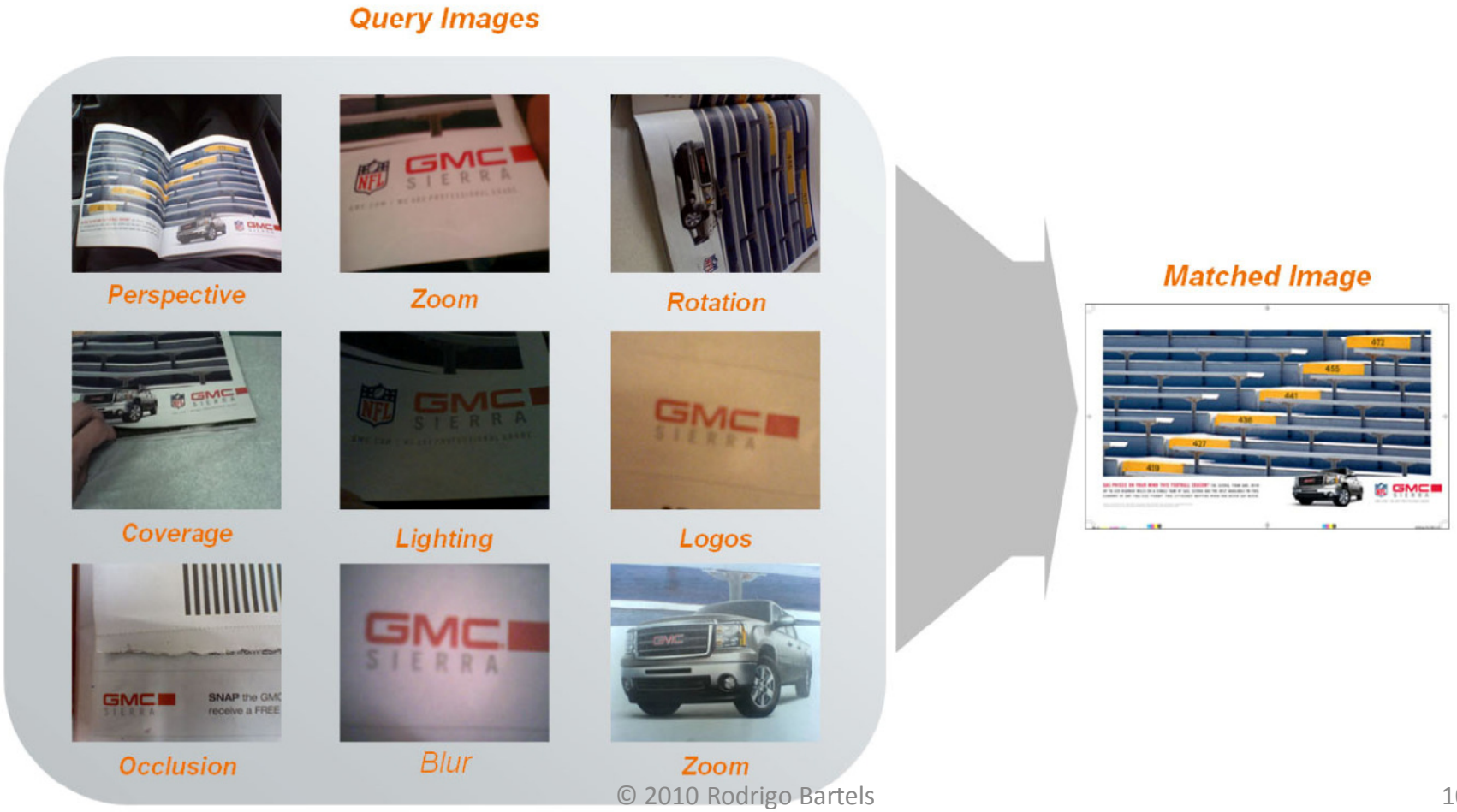
Más allá del texto…

- Actualmente las técnicas de RI no se limitan solamente a la indexación de texto.

- Se puede indexar:
  - Audio
  - Vídeo
  - Imágenes
  - Información Espacio-Temporal

Aplicaciones

- Búsqueda de Imágenes



Aplicaciones

- Espacial
  - Bienes raíces
  - Ubicación de clientes y competidores
- Temporal
  - Precios y recomendaciones
- Audio
  - Derechos de Autor