JIIDE 2011

UNIVERSIDAD DE CASTILLA-LA MANCHA

IDR Instituto de Desarrollo Regional

Descubrimiento y Geoprocesado de Información Espacial mediante un buscador Semántico

Javier Márquez, **David Cifuentes**,
Antonio Quintanilla, J.E Córcoles
Universidad de Castilla-La Mancha. España

Introducción

- Multitud de aplicaciones geoespaciales:
  - Cartografía de visualización raster y vectorial.
  - Geocoding – Localizadores de lugares
  - Servicios de procesamiento.
- Aplicaciones basadas en servicios Web.
- Estándares geoespaciales de la Open Geospatial Consortium (OGC)

Introducción (II)

- Miles de Servicios OGC en Internet disponible para todos los usuarios.
- Más de 1400 servicios localizados sólo en España!
- Todos los servicios de OGC en conjunto, ofrecen mucho más datos que cualquier solución privada.

Introducción (III)

- ¿Por qué la mayoría de los usuarios eligen soluciones privadas como Google Maps?
  - Fácil interacción para obtener información espacial.
  - Buena geolocalización de divisiones administrativas tales como países, ciudades; negocios, hoteles, parkings, etc.
- Problemas de este tipo de soluciones:
  - Recuperación de información de una única fuente (información limitada).
  - Sin conocimiento semántico.
  - Búsquedas basadas en palabras claves.

Descripción
INVESTIGACIÓN

- Tratar de acercar los servicios OGC al usuario de Internet:
  - Mediante consultas de usuario del tipo *"Los ríos de Madrid",* ofreciendo servicios con la información solicitada.
  - Operadores geográficos; cruzan, atraviesa, contiene, cerca...
  - Integración semántica a servicios espaciales y no espaciales.

Descripción (II)
CONTRIBUCIÓN

- Un algoritmo que actualiza contenidos mediante relaciones semánticas entre los conceptos que se obtienen de las ontologías.
- Los resultados de consultas se ordenan según criterios conceptuales y geográficos.

Arquitectura

A Nivel Internet: Información necesaria para configurar el sistema en su conjunto.

A Nivel Crawler: Servicios encargados de la búsqueda de información en Internet y actualización de índices del motor de búsqueda.

A Nivel Motor de búsqueda:

Tarea de recuperación de información de cualquier estructura (índices y taxonomías del dominio de aplicación para el conocimiento semántico).

Nivel Internet

- Servicios OGC:
  - Web Map Service (WMS): Visualización de cartografía en Raster.
  - Web Feature Service (WFS): Visualización de cartografía en vectorial.
  - Web Feature Service Gazetteer (WFS-G): Diccionarios geográficos.
- Base de datos de lugares:
  - Permiten la validación de topónimos.
  - Obtiene una jerarquía de divisiones administrativas.
    - ***Ej. Geonames.***

Nivel Crawler

- Rastreador de servicios OGC:
  - Cada enlace (link) encontrado es validado mediante peticiones **"GetCapabilities".** Similar al WSDL de los Servicios Web.
  - De los servicios válidos se obtiene del XML la información y sus operaciones.

Nivel de Motor de búsqueda

- Análisis de información, tratamiento, recuperación e indexación de contenidos.
  - **Módulo inferencia**: Analiza metadatos de servicios y peticiones a ontologías.
  - **Módulo indexación**: Manejador de estructuras de índices.
  - **Analizador de consultas**: Se encarga de extraer los datos del sistema.

Módulo de inferencia

OBTENCIÓN DEL ÁREA GEOGRÁFICA MÁS ADECUADA PARA UN DETERMINADO SERVICIO

```
<Service>
  <Name>OGC:WMS</Name>
  <Title>WMS Server Madrid</Title>
  <Abstract>Server WMS SDI Madrid</Abstract>
        <KeywordList>
          <Keyword>Madrid</Keyword>
          <Keyword> wms</Keyword>
          <Keyword> SDI-MADRID</Keyword>
        </KeywordList>
</Service>
```

1) Nombre de lugar "Madrid"

Base de datos de lugares (Internet)

```
<Layer queryable="1" opaque="0" cascaded="0">
    <Name>water_bodies</Name>
    <Title>Water bodies 1:25000</Title>
    <Abstract>Water bodies in Madrid</Abstract>
    <SRS>EPSG:23030</SRS>
    <SRS>EPSG:4230</SRS>
    <SRS>EPSG:4326</SRS>
    <SRS>EPSG:32630</SRS>
    <SRS>EPSG:4258</SRS>
    <SRS>EPSG:25830</SRS>
    <LatLonBoundingBox minx="-5.57333" miny="37.9774" maxx="-0.775927" maxy="41.3338" />
    <BoundingBox SRS="EPSG:25830"
                 minx="284561" miny="4.20622e+06" maxx="686177" maxy="4.57589e+06" />
```

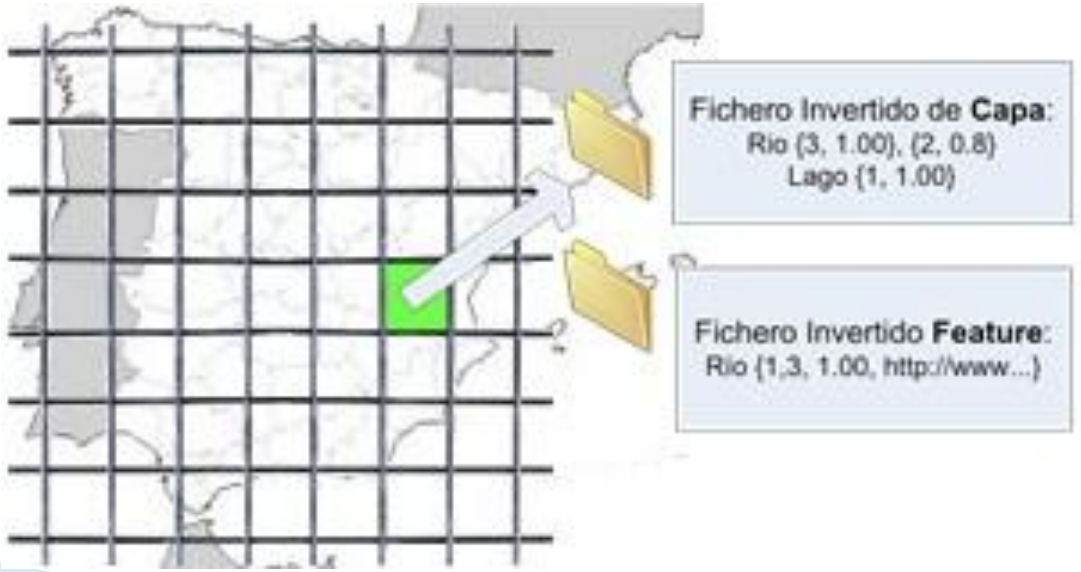
2) Análisis de metadatos dentro de las coordenadas

3) Área geográfica del servicio – Celdas afectadas -

Módulo de Indexación

INDEXACIÓN DE SERVICIOS

- Una cuadrícula abarca el área geográfica del búscador. *Ej: Península Ibérica*
- Las celdas contienen indexados contenidos y capas asociados a esa región. Servicios WMS y WFS para ese zona concreta.



Módulo de Indexación (II)

INDEXACIÓN DE SERVICIOS

```
<Layer queryable="1" opaque="0" cascaded="0">
    <Name>water_bodies</Name>
    <Title>Water bodies 1:25000</Title>
    <Abstract>Water bodies in Madrid</Abstract>
    <SRS>EPSG:23030</SRS>
    <SRS>EPSG:4230</SRS>
    <SRS>EPSG:4326</SRS>
    <SRS>EPSG:32630</SRS>
    <SRS>EPSG:4258</SRS>
    <SRS>EPSG:25830</SRS>
    <LatLonBoundingBox minx="-5.57333" miny="37.9774" maxx="-0.775927" maxy="41.3338" />
    <BoundingBox SRS="EPSG:23030"
                 minx="284561" miny="4.20622e+06" maxx="686177" maxy="4.57589e+06" />
```

Cada servicio contiene ficheros invertidos para las celdas asociadas

Water bodies, 1:25000 → ID layer, score

Water bodies → 3587, 1.0

1:25000 → 3587, 1.0

Los servicios de mapas generalmente ofrecen varias capas (ríos, imágenes de satélite, calles, etc.). Cada capa se identifica con un ID

La puntuación [0,1] es utilizada en el ranking de resultados

Módulo de Indexación (III)

EXPANSIÓN SEMÁNTICA Y PUNTUACIÓN

```
<Layer queryable="1" opaque="0" cascaded="0">
    <Name>water_bodies</Name>
    <Title>Water bodies 1:25000</Title>
    <Abstract>Water bodies in Madrid</Abstract>
    <SRS>EPSG:23030</SRS>
    <SRS>EPSG:4230</SRS>
    <SRS>EPSG:4326</SRS>
    <SRS>EPSG:32630</SRS>
    <SRS>EPSG:4258</SRS>
    <SRS>EPSG:25830</SRS>
```

```
.775927" maxy="41.3338" />
maxy=
```

Los conceptos extraídos se relacionan semánticamente mediante ontologías.

Clasificación de relaciones semánticas (hipónimos, hiperónimos, sinónimos...)

Ej:Water bodies -> Rivers, Lagoon…

No todas las peticiones a las ontologías externas devuelven resultados semánticos. Ej.1:25000.

Water bodies→ 3587, 1.0

1:25000→ 3587, 1.0

Rivers→ 3587, 0.8

Lagoon→ 3587, 0.8

Analizador de consultas

Ej: *“Ríos que cruzan Madrid “*

- La primera tarea es distinguir que corresponde al “*que”*, al “operador” y al *"dónde"*.
- El geoparser proporciona el área geográfica de interés de la búsqueda. *Madrid*
- El operador topológico obtendrá las celdas a consultar. *cruzan*
- Acceso a los ficheros invertidos para recuperar la información.
- Ordenación de elementos válidos por puntuación

Analizador de Consulta (II)

PROCESO INTERNO DE CONSULTA

Water bodies→ 3587, 0.8

1:25000→ 3587, 0.8

Rivers→ 3587, 0.64

Lagoon→ 3587, 0.64

Rivers→ 1245, 1.00

Roads→ 2376, 1.00

1) **El ámbito espacial de la consulta es "Madrid", el buscador trabaja con los ficheros invertidos asociados a las celdas de Madrid.**

2) **Dos capas ofertan información sobre "Rivers" (Ríos): Layer ID 3587, Layer ID 1245.**

3) **Layer ID 1245 (1.00) tiene una puntuación mayor a Layer ID 3587 (0.64). La puntuación 1,00 indica en este caso la capa ID 1245 es más adecuada.**

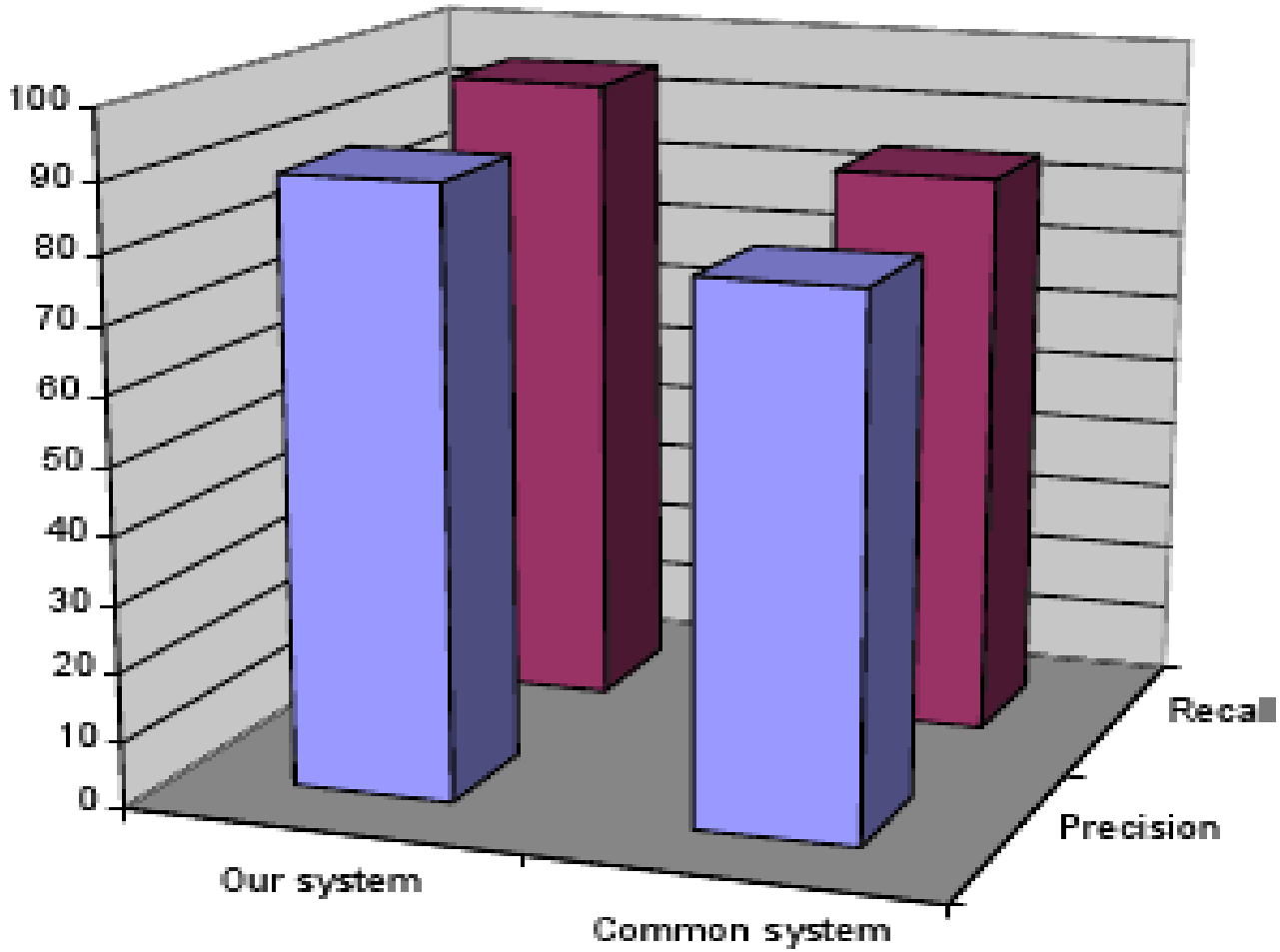
4) **La Layer ID 1245 se recuperará en primera posición, y Layer ID 3587 en segunda posición. El resto de indexadas no se recuperan en esta búsqueda.**

Experimentos

- El rendimiento del sistema se ha comparado con motores de búsqueda tradicionales basados en palabras clave.
  - Se han examinado decenas de consultas de la forma *<qué, operador, dónde>*
  - Se han indexado un centenar de capas de diferentes servicios.

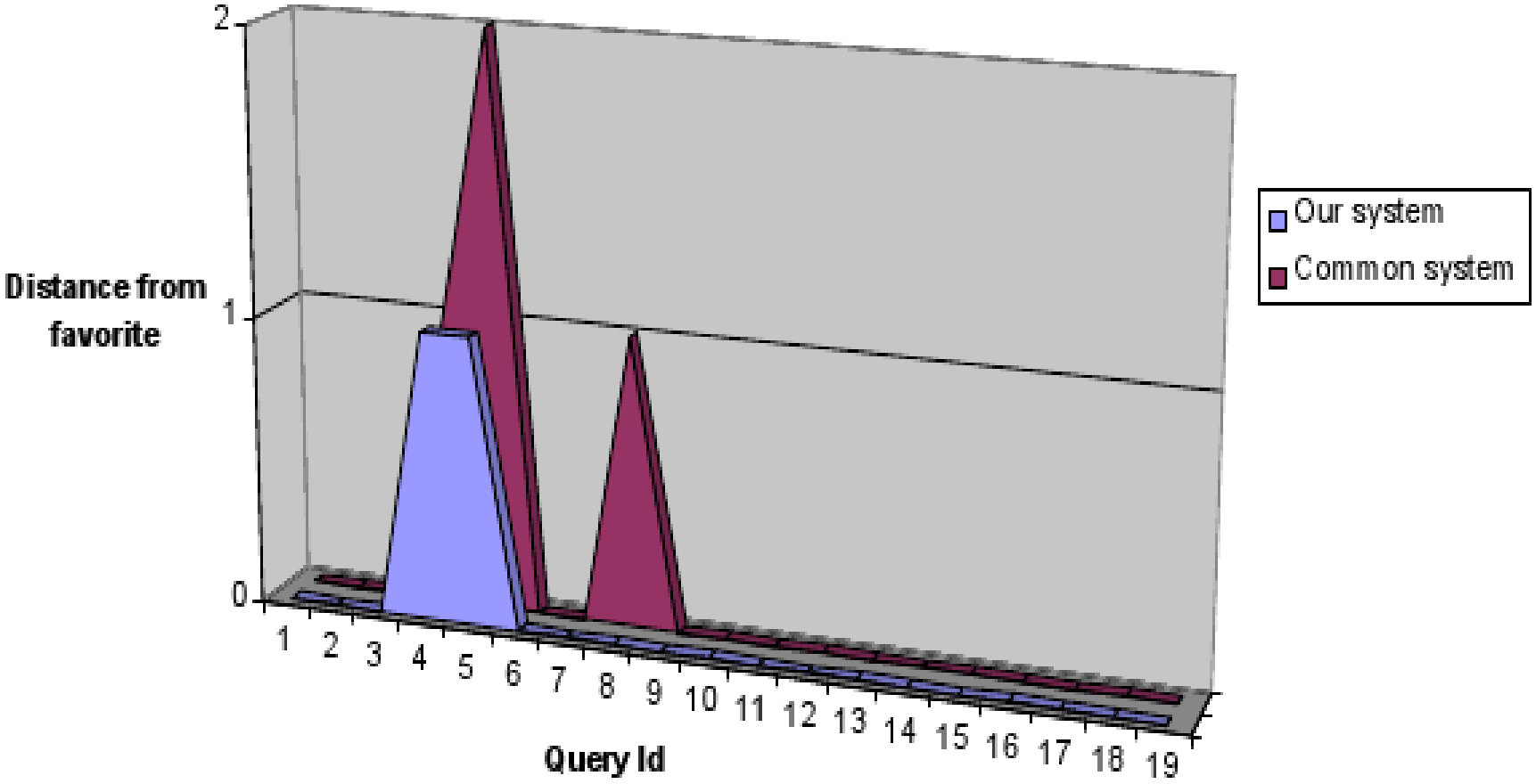
Rendimiento General

(precisión y recuperación)



Rendimiento de resultados

RANKING



Conclusiones

- Los resultados pre-calculados asociados a capas, mediante aspectos semánticos y geográficos mejoran el ranking.
- La utilización de ontologías del dominio junto con los metadatos del servicio, mejora la recuperación de información y precisión.
- La propuesta une conceptos relacionados con la tradicional IR y la Web Semántica, intentando acercar la información geoespacial al usuario.

Trabajos futuros

- Diseñar aplicaciones para diferentes propositos basadas en el motor de búsqueda.
- Mejorar la estructura del árbol de indexación para consultas cuyos ámbitos sean muy grandes.
- Creación de ontologías propias de dominios concretos.

¿Alguna pregunta?

Gracias por su atención