

Recuperación de información geográfica basada en múltiples formulaciones y motores de búsqueda

Geographic Information Retrieval based on multiple formulations and search engines

José Manuel Perea Ortega

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
jmperea@ujaen.es

Resumen: Este trabajo está relacionado con el área de la Recuperación de Información Geográfica (Geographical Information Retrieval, GIR). En él se ha diseñado, desarrollado y evaluado un sistema de recuperación de información geográfica, aplicando para ello diferentes técnicas de Procesamiento del Lenguaje Natural (PLN), recursos de conocimiento geográfico y herramientas de recuperación de información. El sistema desarrollado es modular, lo que permite la integración de otros módulos o recursos de conocimiento. La principal novedad del sistema consiste en utilizar una arquitectura basada en la creación de diferentes subconsultas a partir de la consulta original y también en la aplicación de varias herramientas de recuperación y técnicas de fusión de resultados, consiguiendo con ello mejoras en la variabilidad y en la cobertura de los documentos recuperados.

Palabras clave: Recuperación de información geográfica, reformulación de consulta, recuperación de información, procesamiento del lenguaje natural

Abstract: This work is related to the area of Geographic Information Retrieval (GIR). In this work, a GIR system has been designed, developed and evaluated, applying different techniques of Natural Language Processing (NLP), several information retrieval engines and using external geographic knowledge resources. The developed system is modular, allowing the integration of other modules or knowledge resources. The main novelty of the system is to use an architecture based on creation of different subqueries from the original query, the use of three textual information retrieval tools and different fusion techniques of document lists, achieving improvements in variability and coverage of the documents retrieved.

Keywords: Geographic Information Retrieval, Query Reformulation, Information Retrieval, Natural Language Processing

1. *Introducción*

Hoy día la información geográfica es almacenada en una amplia variedad de medios y tipos de documentos. En las últimas décadas, la tecnología utilizada para acceder a este tipo de información se ha centrado en la combinación de mapas digitales y bases de datos, que es lo que caracteriza a la mayoría de los sistemas de información geográfica. Sólo en los últimos años se ha prestado especial atención en desarrollar sistemas automáticos que traten específicamente de recuperar esa información geográfica presente, por ejemplo, en el recurso inmenso de documentos no estructurados que componen la

web (Larson, 1996; Purves et al., 2007). Actualmente, la aplicación de métodos usuales de recuperación de información en la tarea GIR es satisfactoria para algunas consultas geográficas, pero con bastantes limitaciones. La tarea GIR está relacionada con la mejora en la calidad de la información geográfica recuperada por un motor de búsqueda, es decir, su objetivo es mejorar la recuperación de información utilizando algún tipo de razonamiento geográfico tanto en las consultas como en los documentos (Jones y Purves, 2008). En esta memoria se ha desarrollado un prototipo de sistema GIR basado en la integración de varios motores de recuperación de informa-

ción y la formulación de diferentes subconsultas a partir de la consulta original. Esta arquitectura ha sido evaluada utilizando el foro de evaluación GeoCLEF.

2. Principales aportaciones

A continuación se exponen las principales contribuciones de este trabajo:

Diseño y desarrollo de un prototipo de sistema GIR modular. Se ha demostrado que los resultados obtenidos con este prototipo utilizando el marco de evaluación GeoCLEF han superado, en valor promedio de precisión y cobertura, los mejores resultados alcanzados en tres de las cuatro ediciones.

Diseño y desarrollo de una herramienta para la detección y reconocimiento de entidades geográficas (Geo-NER). Esta herramienta está basada en la utilización de recursos externos como *GeoNames* y *Wikipedia*, aplicando cierto razonamiento geográfico para detectar las *geo-entidades*, utilizando patrones sintácticos y características geográficas. Para su evaluación se han utilizado las consultas geográficas de GeoCLEF, demostrando que Geo-NER consigue mejorar de forma significativa la precisión y la cobertura obtenidas con otras herramientas NER como LingPipe o GATE, en porcentajes que varían entre el 18 % y el 40 %.

Técnicas de reformulación de consulta orientadas a consultas geográficas. Se ha demostrado la efectividad que supone utilizar estas subconsultas en el sistema GIR propuesto, ya que han permitido recuperar documentos que no se conseguían recuperar con la consulta original.

Estudio y evaluación del comportamiento de diferentes herramientas RI para la tarea GIR. Se ha evaluado el comportamiento en el sistema propuesto de tres de las herramientas RI más utilizadas en la actualidad, como son Terrier, Lemur y Lucene.

Estudio y evaluación de varios algoritmos de fusión de listas de documentos, aportando una estrategia basada en normalización RSV que consigue los mejores resultados. Se han evaluado diferentes algoritmos tradicionales de fusión de listas, tales como normalización RSV, RoundRobin, CombSum o CombMNZ. El objetivo principal de la fusión en el sistema

propuesto ha sido mantener la cobertura alcanzada por el uso de los diferentes motores RI, provocando el mínimo impacto en el valor MAP de la lista fusión generada como resultado final.

Función de reordenación basada en la relevancia léxica, semántica y geográfica entre consulta y documento. Se ha propuesto una función de reordenación por relevancia de los documentos recuperados por cada motor RI y que han sido fusionados en una única lista. Esta función de *reranking* está basada en el cálculo de tres similitudes entre documento y consulta: léxica, semántica y geográfica.

3. Información adicional

Tesis doctoral en Informática realizada en la Universidad de Jaén por José Manuel Perea Ortega, bajo la dirección de los doctores L. Alfonso Ureña López y Manuel García Vega. La defensa tuvo lugar el día 13 de septiembre de 2010 ante un tribunal formado por los catedráticos Manuel Palomar Sanz (Univ. de Alicante) y Francisco Javier Ariza López (Univ. de Jaén), así como por los doctores María Teresa Martín Valdivia (Univ. de Jaén), José Antonio Troyano Jiménez y Víctor Jesús Díaz Madrigal (Univ. de Sevilla). La calificación obtenida fue *Sobresaliente Cum Laude* por unanimidad.

Bibliografía

- Jones, Christopher B. y Ross S. Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.
- Larson, R. 1996. Geographic information retrieval and spatial browsing. En Smith y M. Gluck, editores, *Geographic Information Systems and Libraries: Patrons, Maps and Spatial Information*, páginas 81–124.
- Purves, Ross S., Paul Clough, Christopher B. Jones, Avi T. Arampatzis, Benedicte Bucher, David Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, y Bisheng Yang. 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745.