Informática
XVI CONVENCIÓN Y FERIA INTERNACIONAL 2016
16TH INTERNATIONAL CONVENTION AND FAIR
Conectando sociedades
Connecting societies
La Habana, Cuba
Del 14 al 18 de marzo
Havana, Cuba
March 14th to 18th

MEJORANDO LA USABILIDAD DE BÚSQUEDAS POR TOPÓNIMOS CON TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN Y PROCESAMIENTO DE LENGUAJE NATURAL

IMPROVING USABILITY OF PLACE NAME SEARCHES THROUGH INFORMATION RETRIEVAL AND NATURAL LANGUAGE PROCESSING TECHNIQUES

Germán Carrillo[1]

Centro de Investigación y Desarrollo en Información Geográfica – CIAF, Colombia
german.carrillo@igac.gov.co, Transversal 4 No. 52B-35

RESUMEN: *Se identificó una baja usabilidad en componentes de búsqueda por topónimos de aplicaciones de Sistemas de Información Geográfica – SIG, estudiándolos en tres niveles: internacional, nacional e institucional. Por medio de una revisión de literatura en los campos de Recuperación de Información – IR, Procesamiento de Lenguaje Natural – PLN y modelos de catálogos de nombres geográficos, así como de un primer estudio de usabilidad, se determinó un conjunto de requerimientos para componentes de búsqueda por topónimos. A diferencia de la mayoría de aplicaciones SIG estudiadas, el estudio se enfoca en usuarios no expertos. Se compararon varios motores de búsqueda de software libre para analizar, indexar y ordenar información de lugares geográficos. La información de topónimos empleada es producida por instituciones afiliadas a la Infraestructura Colombiana de Datos Espaciales – ICDE, pero principalmente por el Instituto Geográfico Agustín Codazzi – IGAC, como coordinador. Después de comparar formatos de almacenamiento de geometrías, se seleccionó TopoJSON comprimido para reducir tiempos de consulta. El modelo de lugar geográfico usado toma ideas de modelos reconocidos, manteniendo el enfoque en optimización de búsquedas web. Siguiendo una metodología ágil, se desarrollaron prototipos incrementales de un nuevo componente de búsqueda. El prototipo final fue puesto a prueba en un segundo estudio de usabilidad, así como en una evaluación de desempeño, encontrando mejoras sustanciales en métricas de efectividad, eficiencia y satisfacción del usuario. El componente de búsqueda desarrollado se planea disponer en el Geoportal de la ICDE, además de embeberlo en aplicaciones SIG web del IGAC.*

Palabras Clave: recuperación de información; procesamiento de lenguaje natural; gazetteer; infraestructuras de datos espaciales; usabilidad; topónimo.

ABSTRACT: *We identified a lack of usability in place name search components of web Geographic Information Systems (GIS) applications by studying them at three levels: international, national, and institutional. By means of both, a literature review on the fields of Information Retrieval (IR), Natural Language Processing (NLP), and gazetteer models, as well as a first usability study, we determine a number of requirements for place name search components. In contrast to most GIS applications studied, we focus on non-experienced users. We compare several open source search engines for analyzing, indexing, and ranking place information in a scalable manner. Place name information used is produced by public institutions affiliated to the Colombian Spatial Data Infrastructure (SDI), but mainly by the Agustín Codazzi Geographic Institute (IGAC), as its coordinator. After a*

comparison of geometry storage formats we select zipped TopoJSON to reduce network round trip lapse. The place model we employ takes insights from recognized place models, while keeping focus on web search. Following an agile methodology, we developed incremental JavaScript prototypes of a new search component. Our final prototype is tested in a second usability study as well as in a performance evaluation, finding significant improvements in metrics of effectiveness, efficiency, and user satisfaction. We plan to implement the developed place name search component in the ICDE's Geoportal as well as to embed it in GIS applications built by IGAC.

KeyWords: information retrieval; natural language processing; gazetteer; spatial data infrastructures; usability; place name.

1. INTRODUCCIÓN

Las búsquedas por topónimos son uno de los tres servicios esenciales proveídos por toda Infraestructura de Datos Espaciales – IDE [1]. Los componentes de búsqueda por topónimos permiten a los usuarios encontrar información acerca de lugares geográficos por medio de sus nombres. El nombre del lugar es la llave de acceso a información gráfica (geometrías georreferenciadas) y alfanumérica acerca del mismo.

En el campo del mapeo web, este tipo de componentes son usualmente subestimados e ignorados por diseñadores y desarrolladores, generando problemas de usabilidad [3], [4] para usuarios no expertos. Los especialistas en SIG están acostumbrados a escribir consultas que involucran operadores siguiendo cierta sintaxis y las herramientas SIG en la web tratan de replicarles ese esquema de trabajo.

La informática ha abordado problemas de usabilidad en componentes de búsqueda en la web por décadas [5]. Los campos de RI y PLN proveen un amplio número de técnicas para superarlos. Sin embargo, las aplicaciones SIG en la web solo son ocasionalmente optimizadas con dichas técnicas. La escogencia de un conjunto de técnicas que ayuden al usuario a ejecutar búsquedas depende de las características de los datos a buscar, las relaciones entre los mismos, los resultados esperados de las búsquedas y la visualización de resultados.

El objetivo de este artículo es compartir la metodología seguida por el CIAF para rediseñar el componente de búsqueda por topónimos, con el fin de mejorar su usabilidad y, de este modo, asegurar que la información producida y mantenida por el IGAC pueda ser descubierta y usada ampliamente. Cabe recordar que el IGAC es legalmente responsable de administrar la información de topónimos y de promover investigación alrededor de la misma.

Este artículo está estructurado de la siguiente manera: en la sección *Contenido* se presenta un diagnóstico, un listado de requerimientos para el componente, la arquitectura propuesta, el modelo de datos establecido, las técnicas de RI y PLN empleadas, y se describe y evalúa el prototipo desarrollado. Finalmente, se presentan las conclusiones del estudio y se propone trabajo futuro.

2. CONTENIDO

2.1 Diagnóstico

Para comenzar, se realizó un diagnóstico en tres niveles: Internacional, nacional e interno.

A nivel internacional, uno de los expertos en mapeo web más influyentes de los Estados Unidos, Brian Timoney [6], presentó una serie de artículos web en los que sustentó por qué no funcionan los portales de mapas. Uno de los cinco aspectos resaltados por el autor, es la poca atención que se le presta a los componentes de búsqueda basados en texto. El autor contrasta búsquedas basadas en formularios que reflejan la estructura de almacenamiento en la base de datos, contra interfaces sencillas y robustas como una caja de texto con funcionalidad de autocompletar. Agrega que los usuarios interactúan primero con este tipo de herramientas y no con el mapa o con los menús de la aplicación, como parecen pensar los diseñadores de aplicaciones SIG en la web.

A nivel nacional se compararon componentes de búsqueda de aplicaciones de entidades públicas representativas, evidenciando que varias de ellas no ofrecen un componente de búsqueda por topónimos, aunque para algunas pareciera ser muy necesario. Cuando está presente, el componente de búsqueda por topónimos no está visible en la interfaz inicial de la aplicación sino que debe ser activado navegando por algún menú. Por lo general el componente no se optimiza, sino que consiste en una caja de texto en donde el usuario ingresa términos que harán parte de una consulta en *Structured Query Language* – SQL a la base de datos, o de una consulta *GetFeature* a servicios del *Open Geospatial Consortium* – OGC. Por otra parte, las respuestas no suelen incluir un contexto que permi-

ta a los usuarios desambiguar cuando se obtienen lugares con el mismo nombre. En el mejor de los casos, se acude a componentes de búsqueda de terceros (por ejemplo, Google o ESRI), conllevando al soporte de nombres alternativos, a flexibilidad con tildes, a la presentación de resultados con contexto y a sugerencias ante nombres mal escritos. Cuando estos componentes externos no se configuran de manera apropiada surgen inconvenientes para el usuario, como la obtención de sugerencias ajenas al contexto nacional. Además, usar componentes de terceros significa trabajar con datos globales y privativos

Tabla I. Grado de ejecución de cada búsqueda por parte de cada participante en el primer estudio de usabilidad. Los símbolos representan:

√: Terminada correctamente; ~: Terminada correctamente pero por otro medio; –: Terminada pero incorrecta; ×: No terminada.

	Participante				
Búsqueda	**1**	**2**	**3**	**4**	**5**
1 Pueblorrico	~	–	–	–	~
2 Usiacurí	~	~	~	×	~
3 Chigüiro	~	–	√	√	~
4 Río Nuquí	×	√	√	×	√
5 Resguardo Indígena Vaupés	~	√	√	√	×
6 Departamento Vaupés	~	√	–	×	×
7 Nariño	√	√	–	√	√
8 Nariño, Antioquia	√	√	–	√	√
9 Nariño, Leiva	×	–	–	–	×
10 Código 08	×	√	√	×	√

Finalmente, el diagnóstico interno consistió en un primer estudio de usabilidad [7] de una de las aplicaciones SIG desarrolladas en el CIAF, el SIG de Áreas de Reglamentación Especial – SIG-ARE. El objetivo fue detectar, identificar y caracterizar posibles problemas de usabilidad presentes en el componente de búsquedas de dicha aplicación. Se emplearon las técnicas: pensar en voz alta, observación y cuestionario [2] para obtener datos de efectividad, eficiencia y satisfacción del usuario empleando cinco participantes [3]. En cuanto a los resultados, se destaca la baja efectividad durante la prueba. Solamente el 40% del total de búsquedas asignadas (cincuenta búsquedas en total) a los cinco participantes fueron terminadas exitosamente (ver Tabla I). Es una cifra considerable dada la sencillez de la tarea desde el punto de vista del usuario. Asimismo, es de resaltar que se registraron en total 405 comentarios o expresiones negativas. Esto es, en promedio, cada participante mencionó 81 comentarios negativos mientras realizaba sus 10 búsquedas. Por último, como medida de eficiencia, cada participante tomó en promedio 50.4 minutos en realizar las 10 búsquedas.

2.2 Requerimientos

Con base en el diagnóstico realizado y en una revisión exhaustiva de literatura, se definieron los siguientes requerimientos para el componente de búsqueda por topónimos a desarrollar:

- Soportar búsquedas por nombres geográficos sin necesidad de especificar el tipo. Esto es, permitir ingresar términos de búsqueda y devolver coincidencias de varios tipos geográficos. En el ámbito de bases de datos relacionales, esta funcionalidad equivaldría a realizar búsquedas en varias capas a la vez (por ejemplo, buscar en departamentos, municipios, otros administrativos, ríos, etc.).
- Garantizar la flexibilidad del componente en cuanto a mayúsculas y minúsculas y en cuanto a acentos (tildes) y otros caracteres especiales (por ejemplo la letra 'ñ').
- Soportar nombres alternativos (sinónimos) de lugares geográficos. La búsqueda por el nombre oficial, o por nombres alternativos o vernaculares (esto es, no oficiales, pero ampliamente usados) deberá llevar al usuario a obtener el mismo lugar geográfico como resultado.
- Soportar búsquedas por códigos.
- Soportar funcionalidad de autocompletado mientras se escribe la consulta, con el fin de proporcionar al usuario ayudas visuales para definir su consulta.
- Soportar funciones personalizadas para ranking de resultados. Entre otros factores, dichas

funciones podrán tener en cuenta:

- Cercanía a un área de estudio.
- Cercanía a la extensión actual del mapa[1].
- Temas relacionados con el proyecto. Por ejemplo, para un SIG del Ministerio de Educación Nacional, el tema "escuelas" tendría mayor relevancia que otros temas.
- Calidad del nombre geográfico, esto es, los nombres oficiales pueden tener mayor peso que los nombres alternativos.
- Importancia del lugar geográfico. Por ejemplo, una capital de departamento debería estar en una ubicación más alta del ranking que un municipio que no lo es.
- Población del lugar. Esto es, a mayor población, un lugar debería estar mejor ubicado en el listado de resultados.

- Soportar resultados con contexto, para brindar al usuario mayores herramientas de decisión a la hora de desambiguar entre lugares geográficos con el mismo nombre. El contexto, en este ámbito, se refiere a agregar al nombre geográfico buscado, otros lugares con mayor nivel en jerarquía de contenencia.
- Soportar sugerencias a búsquedas con nombres de lugares mal escritos.
- Incluir en la interfaz de usuario enlaces a búsquedas anteriores.
- Incluir en los resultados de búsqueda enlaces a nombres geográficos relacionados. Por ejemplo, para el lugar Cundinamarca puede mostrarse un enlace a Bogotá, su capital.

2.3 Arquitectura del Buscador por Topónimos

Los requerimientos establecidos conducen a cambiar al paradigma de la RI [8]. Si bien en la recuperación de datos es posible implementar algunos de los requerimientos, la RI tiene técnicas que permiten solucionar cada uno de ellos. En la RI se recupera información con mayor flexibilidad (para ello se hace un análisis de términos almacenados en la base de datos y de términos de búsqueda ingresados por el usuario), se involucra PLN y se retornan resultados ordenados por relevancia [9], [10].

Después de un proceso de comparación y selección de herramientas y formatos para la implementación del nuevo buscador por topónimos, se optó por el siguiente conjunto de herramientas de software y formatos de datos (ver Figura 1):

- Motor de búsquedas: ElasticSearch, el cual permite implementar técnicas de RI y PLN.
- Librería JavaScript para manejo del Modelo de Objetos del Documento – DOM: jQuery.

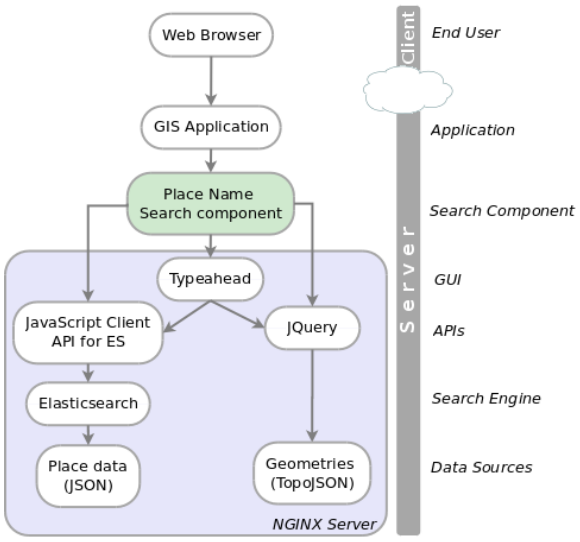

Figura. 1: Arquitectura Buscador por Topónimos.

- API ElasticSearch para JavaScript para facilitar la integración entre software de servidor y de cliente. De hecho, una de las formas en las cuales ElasticSearch provee su API JavaScript, es con base en jQuery.
- Servidor de aplicaciones web: NGINX, ampliamente usado en la web y recomendado como proxy de ElasticSearch.
- Software del cliente (interfaz gráfica): Plugin de jQuery llamado jQuery#typeahead, desarrollado y liberado por Twitter.
- Formatos de datos: Se emplean formatos JSON (para atributos alfanuméricos) y TopoJSON comprimido para almacenar geometrías.

2.4 Modelo de Datos

El buscador por topónimos tiene una estructura de datos conocida como gazetteer, nomenclátor o catálogo de nombres geográficos. Esta estructura se basa en una tabla principal que corresponde al lugar

[1] Este es un factor empleado por *Google Maps* para sus funciones de relevancia.

geográfico. Sus componentes fundamentales son un nombre, un tipo y una geometría [11]. Existen diferentes modelos de gazetteer, cada uno agrega atributos relevantes para su objetivo. En este caso, el objetivo principal del modelo de datos es la optimización de búsquedas por topónimos. Por ello se almacena la información alfanumérica en un archivo JSON, mientras que la geometría en WGS84 es codificada en TopoJSON y se comprime posteriormente en formato GZIP.

Los campos de la tabla lugar corresponden a (ver Tabla II) nombres (nombre oficial, alternativo, sinónimos, coloquiales y de despliegue), códigos (código DANE[2] y Código Único de Entidad), tipos (oficial y de despliegue), posición (punto) y extensión (rectángulo), padres (lugares superiores jerárquicamente), peso (relacionado con la importancia del tipo de lugar), fuente de los datos (autor, fecha, escala y enlace al metadato), un campo *suggest* (propio de ElasticSearch para mostrar sugerencias) y un campo para almacenar relaciones o enlaces entre lugares (por ejemplo, "capital de", "está compuesto por" o "limita con"). Hay otros campos en el modelo, como etnia, área o enlaces web de interés (por ejemplo, página oficial y página en Wikipedia), pero los listados en la Tabla II son los principales.

Tabla II: Campos que componen la tabla Lugar y sus respectivos tipos.

Tabla Lugar	
Campo	**Tipo de campo**
tipo	String
tipo_despliegue	String
codigo	String
codigo_dane	String
codigo_unico_entidad	String
nombre	String
nombre_alternativo	String
sinonimo	String
nombre_coloquial	String
nombre_despliegue	String
nombre_despliegue_alternativo	String
padre	JavaScript object
posicion	geo_point
extent	geo_shape
peso	Float
fuente	JavaScript object
suggest	Completion
relaciones	JavaScript object

Se requiere que el esquema de la tabla sea dinámico pues cada tipo de lugar contiene atributos particulares (por ejemplo, algunos lugares tienen atributo población). Implementar un esquema extensible en el paradigma relacional puede ser complejo, mientras que paradigmas como NoSQL[3] lo permiten fácilmente [12]. Así mismo, se requiere redundancia (algo que el paradigma relacional elimina a través del proceso de normalización) para optimizar las búsquedas. NoSQL lo permite, soportando bases de datos basadas en documentos [13], que suelen estar en formato JSON. Por esto, se implementó una base de datos NoSQL basada en documentos.

2.5 Técnicas de RI y PLN empleadas

En el motor de búsquedas se empleó un índice invertido [14], ampliamente usado para búsquedas web. Se definió un análisis de términos que consistió en eliminar caracteres acentuados, puntuación, preposiciones y artículos. Preposiciones y artículos son *Stop Words* [15], o palabras tan comunes que poco contribuyen a la recuperación de información.

Por otro lado, las jerarquías de lugares (por ejemplo, "Acacías, Meta") se analizaron diferente a los nombres ("Acacías"). Las primeras no se dividen para el indexado, lo que corresponde a la técnica de PLN "frases compuestas y estadísticas" [15].

La relevancia de los resultados se define con base en factores como el grado de coincidencia, la categoría del nombre del lugar (coincidencias con el nombre oficial aparecen primero que coincidencias con un nombre coloquial), el tipo de lugar (los departamentos aparecen primero que los centros poblados) y la población del lugar.

2.6 Prototipo

El prototipo mostrado en la Figura 2 corresponde a una aplicación en JavaScript que integra el componente Typeahead de Twitter y el motor de búsquedas ElasticSearch. Se han generado interfaces entre el buscador por topónimos y tres librerías para mapas en la web: Leaflet, OpenLayers 2 y OpenLayers 3, facilitando la integración del componente de búsquedas con cualquiera de esas tres librerías.

[2] Código que el Departamento Administrativo Nacional de Estadística – DANE asigna a las entidades político-administrativas.

[3] No solo SQL.

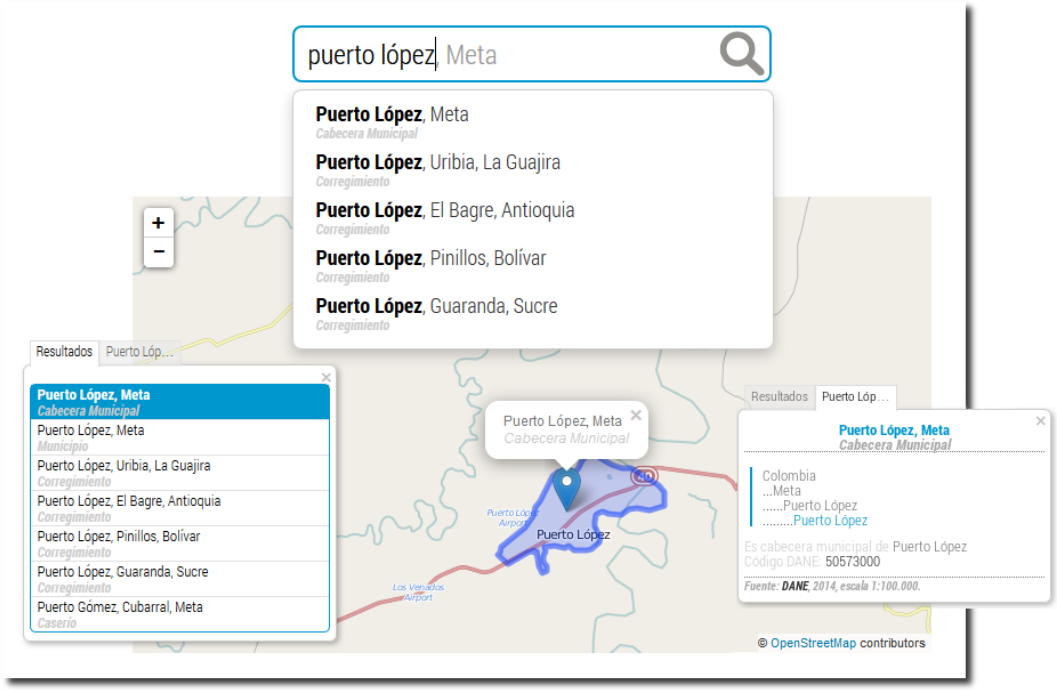

Figura 2: Prototipo del Buscador por Topónimos desarrollado.

El prototipo del buscador por topónimos puede ser accedido en Internet[4], disponiendo cerca de once mil lugares a nivel nacional. Las técnicas de RI y PLN usadas permiten realizar búsquedas como "la capital de Antioquia", "Costa Rica, Valle", "Corregimiento de Nariño en el Municipio de Leiva (Nariño)", "la ciudad milagro", "Medallo", "C/marca" y "05001"[5].

2.7 Evaluación

La evaluación del prototipo desarrollado se llevó a cabo de dos maneras. Primero, un segundo estudio de usabilidad permitió comparar métricas de eficiencia, efectividad y satisfacción. Segundo, una prueba de rendimiento permitió medir el tiempo que tarda el componente desarrollado en retornar respuestas a las búsquedas, y compararlo con otros servicios similares basados en GeoServer.

2.7.1 Segundo estudio de usabilidad

Las condiciones del segundo estudio de usabilidad fueron las mismas del primero. Esto es, se tuvo el mismo número y perfil de los participantes (cabe aclarar que los participantes fueron distintos), las mismas tareas y las mismas técnicas de evaluación para recoger datos de eficiencia, efectividad y satisfacción del usuario. Para este segundo estudio se incorporó el buscador por topónimos desarrollado en la interfaz de uno de los proyectos SIG del CIAF.

En esta ocasión, los resultados fueron más alentadores, observando mejoras significativas en las métricas analizadas. El porcentaje de búsquedas exitosamente realizadas fue del 80% (ver Tabla III), esto es, el doble del primer estudio de usabilidad. Por otro lado, el número de comentarios negativos fue de 39 en total, casi 8 por participante al realizar sus diez (10) búsquedas.

Varios participantes incluso mencionaron comentarios positivos relativos al buscador por topónimos, concretamente, se resaltó su rapidez, la flexibilidad con tipos geográficos y con errores ortográficos, el soporte de nombres y códigos para realizar búsquedas, la información de contexto (por ejemplo,

[4] Se accede en la dirección de Internet: http://186.154.197.60:82/app/prototipo.html

[5] Código DANE de Medellín.

población del lugar) y la demarcación clara del resultado en el mapa.

Tabla III. Grado de ejecución de cada búsqueda por parte de cada participante en el segundo estudio de usabilidad. Los símbolos representan:

√: Terminada correctamente; ~: Terminada correctamente pero por otro medio; –: Terminada pero incorrecta; ×: No terminada.

	Participante				
Búsqueda	**1**	**2**	**3**	**4**	**5**
1 Pueblorrico	–	√	√	–	–
2 Usiacurí	–	√	–	–	–
3 Chigüiro	√	√	√	√	√
4 Río Nuquí	√	√	√	√	√
5 Resguardo Indígena Vaupés	√	√	√	√	√
6 Departamento Vaupés	√	√	√	√	√
7 Nariño	√	√	√	√	√
8 Nariño, Antioquia	√	√	√	√	–
9 Nariño, Leiva	√	√	√	√	√
10 Código 08	√	√	×	√	×

El segundo estudio de usabilidad permitió observar que varios problemas fueron superados con el nuevo buscador por topónimos. No solamente los participantes pudieron terminar más búsquedas exitosamente, sino que estuvieron notablemente más conformes con el funcionamiento de la herramienta.

Con respecto a la eficiencia, el tiempo que tardaron los cinco participantes del primer estudio de usabilidad en realizar las diez búsquedas, fue de 252 minutos en total, mientras que los cinco participantes del segundo estudio solamente tardaron 121 minutos (ver Figura 3). Es decir, el aumento en productividad fue notorio, ahorrando más de la mitad del tiempo con el nuevo buscador por topónimos.

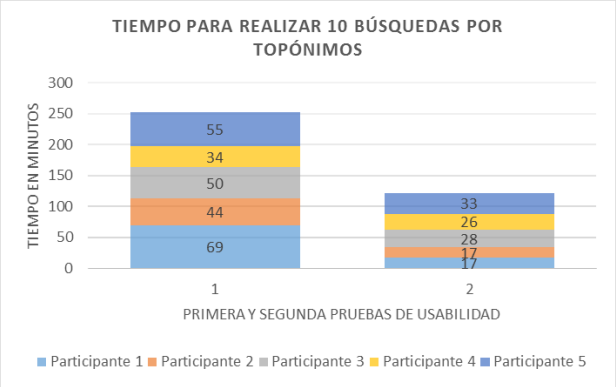

Figura 3: Tiempo en minutos empleado por participante para realizar las 10 búsquedas asignadas en cada prueba de usabilidad.

2.7.2 Prueba de rendimiento

La prueba de rendimiento tuvo por objetivo medir cuantitativamente el tiempo de respuesta de varios componentes de búsquedas por topónimos. Concretamente se compararon: el componente nativo de búsquedas SIG-ARE (que fue el proyecto SIG del CIAF que se tomó para los estudios de usabilidad), el nuevo componente desarrollado y, por último, consultas directas a servicios *Web Feature Service* - WFS de GeoServer[6].

Las pruebas se llevaron a cabo empleando el software JMeter para realizar diversas peticiones, involucrando varios tipos de lugares, varios tipos de geometrías y diferente número de usuarios concurrentes (desde 1 hasta 200).

Por cuestiones de disponibilidad de equipos, los servidores sobre los que están implementados GeoServer y SIG-ARE cuentan con 10 y 8 GB respectivamente, mientras que el servidor sobre el cual se implementó el nuevo componente de búsquedas tan solo cuenta con 4GB. Este factor es relevante al analizar los resultados de la prueba.

Los resultados de la prueba de rendimiento permitieron concluir lo siguiente:

- Tanto GeoServer como el nuevo buscador presentaron tiempos de respuesta muy superiores al servicio nativo de SIG-ARE, puesto que este, a pesar de estar basado en consultas a GeoServer, agrega lógica de procesamiento a

[6] Se incluyó GeoServer pues es el servidor geográfico empleado por SIG-ARE.

la respuesta.

- Al comparar el nuevo buscador contra GeoServer, se observa que este último responde más rápido en consultas alfanuméricas (sin geometría) y en consultas con geometría cuando los polígonos no son grandes. El nuevo buscador superó a GeoServer retornando lugares con geometrías pesadas (ver por ejemplo la Figura 4), como algunos departamentos, municipios, parques nacionales y resguardos indígenas, incluso contando con menos de la mitad en memoria RAM que el servidor donde se implementó GeoServer.
- El nuevo buscador es más rápido que GeoServer cuando se descarta el análisis de términos de búsqueda. Por ejemplo, búsquedas por código DANE (que no se benefician de técnicas de RI ni de PLN) son significativamente más rápidas en el nuevo buscador.

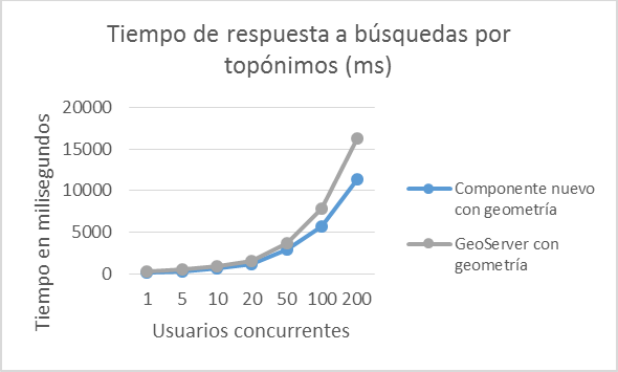

Figura 4: Tiempo de respuesta de GeoServer y del nuevo buscador para el lugar "Parque Nacional Natural Sierra Nevada de Santa Marta".

Los resultados anteriores permiten observar que el paradigma de búsquedas con técnicas de RI y PLN, sobre un motor de búsquedas basado en documentos, puede superar en desempeño al paradigma tradicional de servicios web del OGC.

3. CONCLUSIONES

Este artículo describe la metodología seguida por el CIAF para mejorar la usabilidad de buscadores por topónimos usando técnicas de RI y PLN.

El diagnóstico realizado permite observar que este tipo de componentes son usualmente ignorados en aplicaciones SIG en la web y, por ello, existe una oportunidad para que el IGAC sea referencia en el tema de búsquedas por topónimos (nacional y regionalmente), no solamente disponiendo el buscador desarrollado en sus aplicaciones, sino también brindando un servicio de búsquedas que pueda ser empleado por entidades a nivel nacional.

Durante el proyecto descrito en este artículo se realizaron las siguientes actividades:

- Se elaboró un listado de requerimientos para buscadores por topónimos, que puede ser utilizado como base por entidades a nivel nacional e internacional.
- Se adoptó un modelo de datos con base en modelos de gazetteer internacionales, pero enfocado en búsquedas óptimas en Internet, favoreciendo redundancia de datos y esquemas dinámicos en lugar de normalización y esquemas fijos.
- Se emplearon técnicas de RI y PLN como indexado invertido, remoción de puntuación y de caracteres acentuados, *Stop Words*, frases compuestas y estadísticas, y definición de relevancia.
- Se desarrolló un prototipo en JavaScript para búsquedas por topónimos, haciendo uso de un motor de búsquedas libre y de formatos abiertos como JSON y TopoJSON.
- Se llevaron a cabo dos estudios de usabilidad que permitieron comparar el beneficio de implementar técnicas de RI y PLN en búsquedas por topónimos para aplicaciones SIG en la web.

La mejora en usabilidad de las búsquedas por topónimos se determinó con base en efectividad, eficiencia y satisfacción del usuario. La efectividad se duplicó, pasando del 40% al 80% de búsquedas realizadas exitosamente. La satisfacción del usuario mejoró notoriamente, pasando de 405 a 39 comentarios negativos totales expresados hacia el sistema. Por último, la eficiencia aumentó, por un lado, ahorrando más de la mitad del tiempo en la realización de la prueba, y por el otro, mejorando significativamente el rendimiento del componente nativo de SIG-ARE, e incluso superando a GeoServer en consultas que involucraron geometrías pesadas (polígonos grandes) a pesar de diferencias en hardware a favor de GeoServer.

En cuanto a trabajo futuro, se espera consolidar y disponer la información de topónimos del IGAC, adoptar el buscador por topónimos en las aplicaciones del IGAC y ofrecer un servicio de búsqueda por topónimos a otras instituciones nacionales.

En caso de desear mayor información acerca de las etapas del proyecto, así como del material generado durante el mismo, favor contactar al autor.

4. AGRADECIMIENTOS

El autor quiere agradecer a Juan Manuel Higuera, Dayanna Jiménez, Andrés Briceño y Andrés Guarín, por sus valiosas sugerencias, ideas y preguntas, las cuales ayudaron a darle forma a este proyecto. Asimismo, gracias a Mireya Ruiz y Luis Aguilar por su colaboración en temas logísticos y de organización; a los 10 participantes de los estudios de usabilidad; y, en general, a los compañeros de trabajo por su apoyo durante el proyecto.

5. REFERENCIAS BIBLIOGRÁFICAS

1. Rodríguez, A., López, E., Abad, P. y Sánchez, A.: Modelo de Nomenclátor de España v.1.2., Consejo Superior Geográfico, Infraestructura de Datos Espaciales de España, 2006.

2. Lewis, J.: "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use", International Journal of Human-Computer Interaction, 1 ed., pp. 57 - 78, Vol.7, 1995.

3. Nielsen, J.: Usability Engineering, Morgan Kaufmann, California, 1993.

4. ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability, 1998.

5. Singhal, A.: "Modern information retrieval: A brief overview", IEEE Data Eng. Bull, Vol.24, pp. 35 - 43, 2001.

6. Timoney, B.: MapBrief: Why Map Portals Don't Work, 2013. Available online: http://mapbrief.com/2013/02/05/why-map-portals-dont-work-part-i/ (accessed on 27 October 2015).

7. Rubin, J.: Handbook of usability testing: how to plan, design, and conduct effective tests, Wiley, 1994.

8. Salton, G.: Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1968.

9. Saracevic, T.: "Relevance: a review of and a framework for the thinking on the topic", Journal of the American Society for Information Science, Vol.26, pp. 321 - 343, 1975.

10. Hill, L.: Georeferencing: The geographic associations of information, MIT Press, 2009.

11. Hill, L.: Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, pp. 280 - 290, 2000.

12. Marcus, A.: "The NoSQL ecosystem", The architecture of open source applications, pp. 185 - 205, 2011.

13. Weissmann, H.L.: Deconstructing NoSQL: the quest for better definitions. 2013. Available online: http://www.itexto.com.br/devkico/en/?p=14 (accessed on 27 October 2015).

14. Cong, G., Jensen, C. y Wu, D.: "Efficient retrieval of the top-k most relevant spatial web objects", Proceedings of the VLDB Endowment, Vol.2, pp. 337 - 348, 2009.

15. Brants, T.: "Natural Language Processing in Information Retrieval", CLIN, 2003.

6. SÍNTESIS CURRICULARES DE LOS AUTORES

Germán Carrillo nació en Bogotá, Colombia. Es Ingeniero Catastral y Geodesta (2005) de la Universidad Distrital de Bogotá, Especialista en Sistemas de Información Geográfica (2008) de la misma Universidad y Máster of Science en Geoinformática (2013) de la Universidad de Münster, Alemania. Trabaja actualmente como asesor e investigador en el Instituto Geográfico Agustín Codazzi. Es administrador del portal web GeoTux, que promueve y comparte software libre en Geomática. Es miembro con derecho a voto de la fundación OSGeo, colaborador de QGIS y colaborador del paquete Maptools de R. Es autor y mantenedor de paquetes de QGIS, R y pgAdmin3. Reside en la Transversal 4 No. 52B-35, Bogotá, Colombia, y puede ser contactado en la dirección de correo electrónico:

german.carrillo@igac.gov.co