Ingeniería Investigación y Tecnología
volumen XXI (número 4), octubre-diciembre 2020 1-12
ISSN 2594-0732 FI-UNAM artículo arbitrado
Información del artículo: Recibido: 26 de noviembre de 2019, aceptado: 4 de julio de 2020

https://doi.org/10.22201/fi.25940732e.2020.21.4.029

Ingeniería
Investigación y
Tecnología

Computación paralela para la geocodificación de direcciones postales cubanas
Parallel computing for the Cuban postal addresses geocoding

Alfonso-Cantillo Ofir
Universidad Tecnológica de La Habana
José Antonio Echeverría, Cuba
Corrreo: ofir4991@gmail.com
https://orcid.org/0000-0003-2401-4565

Sánchez-Ansola Eduardo
Universidad Tecnológica de La Habana
José Antonio Echeverría, Cuba
Correo: esanchezansola@gmail.com
https://orcid.org/0000-0001-5977-1633

Resumen

Teniendo en cuenta la velocidad y cantidad de datos que se generan en la actualidad, se ha convertido en una necesidad la creación de sistemas capaces de procesarlos en breves periodos de tiempo. La información espacial no queda exenta de esto por lo que la utilización de Sistemas de Información Geográfica se ha convertido en una necesidad. GeoServer es uno de los servidores de mapas que brinda estas funcionalidades mediante servicios. Uno de estos servicios es la geocodificación, que se define como el proceso de convertir una dirección postal en coordenadas geográficas. Una petición de geocodificación puede contener una o varias direcciones, siendo esta última conocida como geocodificación por lotes, la cual puede estar compuesta por una cantidad ilimitada de direcciones. Una de las técnicas que se utiliza para aumentar la capacidad de procesamiento de los sistemas geocodificadores es la computación paralela, que utiliza múltiples elementos de procesamiento para resolver determinado problema. El objetivo del presente trabajo consiste en la disminución de los tiempos de respuesta de los algoritmos de geocodificación de direcciones por lotes. Para lograr dicho objetivo, se desarrollaron varias versiones para la geocodificación haciendo uso de la computación paralela. Cada una de estas versiones fue incluida en el servidor de mapas GeoServer como un servicio Web y fueron evaluadas, mediante la utilización de pruebas estadísticas, con el fin de obtener la versión que ofrece menor tiempo de procesamiento. Tanto para el desarrollo de las variantes como para su validación, se tuvieron en cuenta direcciones postales del territorio nacional cubano, pues es el alcance pretendido con este trabajo. A partir de la evaluación realizada, se puede afirmar que la versión del algoritmo, desarrollada de forma independiente a la base de datos de referencia, es en la que mejores tiempos se ofrecen las respuestas.
Descriptores: Algoritmo, computación paralela, dirección postal, geocodificación, GeoServer.

Abstract

Taking into account the speed and quantity of data generated today, the creation of systems capable of processing them in a short period of time has become a necessity. Spatial information is not exempt from this, so the use of Geographic Information Systems has become a necessity. GeoServer is one of the map servers that provides these functionalities through services. One of these services is geocoding, which is defined as the process of converting a postal address into geographic coordinates. A geocoding request may contain one or more addresses, the latter being known as batch geocoding and may be composed of an unlimited number of addresses. One of the techniques used to increase the processing capacity of geocoder systems is parallel computing, which uses multiple processing elements to solve a given problem. The objective of this work is to reduce the response times of the geocoding algorithms for batch addresses. To achieve this goal, several versions were developed for geocoding using parallel computing. Each of these versions was included in GeoServer as a web service and were evaluated, through the use of statistical tests, in order to obtain which of all versions offers the least processing time. Both for the development of the variants and for their validation, postal addresses of the Cuban national territory were taken into account, as it is the intended scope of this work. From the evaluation carried out, it can be affirmed that the version of the algorithm developed independently of the reference database is in which better times the answers are offered.
Keywords: Algorithm, geocoding, GeoServer, parallel computing, postal address.

Introducción

El uso de la información geográfica en el día a día de las personas, ha tomado un auge significativo en los últimos años. Servicios populares de internet como Google Maps, Bing Maps y OpenStreetMap, los cuales tienen como base el uso de datos geográficos de diversas fuentes, han provisto a la población de herramientas atractivas para la consulta y manipulación de información espacial.

Los datos geográficos son la clave para diferenciar un Sistema de Información Geográfica (SIG) de otro sistema de información. Los SIG son un potente conjunto de herramientas para recolectar, almacenar, recuperar a voluntad, transformar y presentar datos espaciales procedentes del mundo real (Burrough, 1986). Comúnmente, todas estas funcionalidades se complementan con servicios basados en la localización.

Los Servicios Basados en la Localización (LBS, por sus siglas en inglés) en su forma estandarizada OpenLS (Open Geospatial Consortium, 2008), definidos por la Open Geospatial Consortium (OGC), proporcionan al usuario información y datos basados en la posición geográfica. Estos servicios se basan, por lo general, en una red de comunicaciones y una o más tecnologías de posicionamiento, que en combinación con los SIG recogen la información y la presentan al usuario final (Sánchez, 2012). Los LBS incluyen servicios de ruta, directorio y geocodificación de direcciones postales (Open Geospatial Consortium, 2008). La geocodificación es un proceso que ayuda a organizaciones a refinar y enriquecer información existente relacionada con direcciones y localizaciones en una base de datos. Genera latitud/longitud a partir de una dirección existente, o localización, asimismo es el primer paso a tomar por cualquier aplicación que haga uso de servicios basados en la localización (Oracle, 2007).

En la actualidad, la geocodificación se ha convertido en una necesidad para múltiples análisis espaciales y para la toma de decisiones (Jiang & Stefanakis, 2018). Por esta razón, los investigadores se mantinenen publicando avances en la misma, ya sea desde el punto de vista de su desarrollo o de cómo puede utilizarse como medio para realizar análisis espaciales de mayor profundidad y valor agregado. En el primer caso se encuentra Yin *et al.* (2019) que presentan un método de geocodificación basado en la detección de objetos en imágenes a partir de un marco de Deep Learning y comparan sus resultados con sistemas de geocodificación tradicionales a partir de 22,481 direcciones. Asimismo Küçük & Avdan (2018) proponen un algoritmo basado en Procesamiento del Lenguaje Natural (PLN) que recompone una dirección a partir de una previa descomposición de los elementos significativos de la misma, eliminando errores ortográficos, abreviaturas, etcétera. Estos autores demuestran cómo, usando el API de geocodificación de Google, con las nuevas direcciones se obtienen más resultados y, a su vez, más precisión.

Por otra parte, múltiples autores hacen referencia a la geocodificación como herramienta para poder realizar análisis espaciales más complejos. Tal es el caso de Präger *et al.* (2019) que presentan un estudio de viabilidad en el uso de los servicios de geocodificación online de Google y OpenStreetMap (OSM) en la detección de factores que propician la obesidad, concluyendo que su validez es razonable y que puede ser utilizado en la vigilancia de la diabetes. En Chopin & Caneppele (2019), se usa la geocodificación como medio para realizar un análisis exploratorio de la movilidad entre los agresores y las victimas en delitos de abuso infantil en Francia, mientras que en McIntire *et al.* (2018) se geocodifican las direcciones de 10,750 pacientes con cáncer de próstata del estado de Pensilvania, para crear un índice compuesto que identifica los vecindarios con mayor incidencia de dicha enfermedad.

En Cuba, particularmente en la CUJAE, entre los años 2011 y 2013 se realizaron avances respecto a la geocodificación de direcciones postales cubanas mediante la implementación de un sistema de geocodificación nacional (De Armas & Gutiérrez, 2013). A partir del año 2015, el sistema se incluyó como parte de una Infraestructura de Datos Espaciales (Sánchez, 2015) y en 2017 se comenzaron a ofrecer las funcionalidades de geocodificación a partir de GeoServer mediante la utilización de servicios Web (Girón & Sánchez, 2017). GeoServer es un servidor de código abierto escrito en Java que permite a los usuarios procesar información espacial y posee la capacidad de publicar datos de gran variedad de fuentes utilizando estándares de OGC (GeoServer, 2014).

Según los autores: Dueker (1974); Roongpiboonsopit & Karimi (2010); Cetl *et al.* (2016); Vargas & Horfan, (2013) el proceso de geocodificación de direcciones postales a partir de texto libre está dividido en varias etapas: extraer los elementos necesarios para el proceso de geocodificación, normalizar dichos elementos, estandarizarlos y realizar el proceso de geocodificación mediante la utilización de una base de datos de referencia. Estos métodos "tradicionales" para realizar el proceso de geocodificación pueden llegar a demorarse tanto en un proceso de geocodificación por lotes que su utilización no sea factible en determinadas circunstancias. En Sen *et al.* (2012) se realizó un estudio comparativo entre

diferentes servicios geocodificadores disponibles en internet (Google Maps, Bing y Yahoo Place Finder) donde se geocodificaron varias bases de datos, entre ellas una con un millón de instancias, y su geocodificación completa duró poco más de un año. El tiempo que duró la geocodificación fue, de acuerdo con los autores, muy superior a lo que necesitaban para cubrir las necesidades de geocodificación que tenían, por lo que tuvieron que desarrollar un sistema de geocodificación que utilizara técnicas de computación paralela.

Paralelismo es una propiedad de la computación en la cual las porciones de la funcionalidad a ejecutar y los datos son independientes, por lo tanto, permiten ser procesados al mismo tiempo (Foster, 1995). Existen diferentes opciones para la explotación del paralelismo, desde computadoras independientes trabajando para conseguir un mismo objetivo hasta una misma computadora que contenga varios elementos de procesamiento accediendo a la misma memoria (sistemas de memoria compartida) (Pacheco, 2011).

El objetivo del presente trabajo es disminuir el tiempo del proceso de geocodificación de direcciones postales cubanas. Para ello se realizará un diseño paralelo del algoritmo de geocodificación tomando como punto de partida los antecedentes de la geocodificación de direcciones postales cubanas, planteados anteriormente.

Geocodificación de direcciones postales en cuba

En la era de la información, los datos espaciales son considerados como uno de los valores añadidos que más se valoran de la información. La forma más común de obtener información espacial es mediante la interpretación de las direcciones, ya que una dirección postal es la forma fundamental en que las personas describen dónde se encuentra algo o alguien (Roongpiboonsopit & Karimi, 2010). En el proceso de geocodificación es necesario conocer las diferentes variantes de escritura de las direcciones postales. En De Armas & Gutiérrez (2013) se documentaron los siguientes modelos de direcciones postales en Cuba:

1. Modelo básico urbano de calles y entrecalles. Incluye tres variantes:
   - Calle principal, entre calles y división político administrativa. Por ejemplo: Avenida 27 Núm. 4207 entre calle 42 y calle 44, Playa, La Habana.
   - Intersección entre dos calles y división político administrativa. Por ejemplo: 23 y 12, Plaza de la Revolución, La Habana.
   - Calle principal y división político administrativa. Por ejemplo: Monte Núm. 852 Apto. 3, Habana Vieja, La Habana.
2. Modelo de referencia lineal, ampliamente utilizado en zonas rurales. Por ejemplo: Carretera a Viñales Km 4, José María Pérez, Pinar del Río, Pinar del Río.
3. Modelo en el que se utiliza como referencia un asentamiento poblado o un punto de interés. Por ejemplo: Los Mangos, Amancio, Las Tunas.
4. Modelo utilizado en un amplio conjunto de urbanizaciones edificadas en las últimas décadas donde solamente se utiliza como referencia el número o nombre de los edificios y el nombre de la urbanización. Por ejemplo: Edif. 674 Apto. 30, Alamar 19, Habana del Este, La Habana.

En Girón & Sánchez (2017) se presentó un componente para GeoServer, cuyo objetivo fue brindar al servidor de mapas servicios basados en la localización. Entre estos servicios se encuentra la geocodificación de direcciones postales. El proceso de geocodificación realizado por esta extensión se muestra en un diagrama de actividades en la Figura 1.

La geocodificación se realiza de forma secuencial, por ello cuando se desee procesar una petición de geocodificación por lotes, la geocodificación de una dirección comienza cuando la anterior termina. Para poder geocodificar una dirección primero se realiza un proceso de normalización mediante el cual se eliminan los caracteres extraños que pueda contener, después se extraen los elementos útiles para el proceso de geocodificación, se establece el nivel de precisión que se quiere (calle, esquina, reparto o provincia) y, por último, se procede a buscar la dirección en la base de datos de referencia. De acuerdo con los experimentos realizados, la configuración ideal para la geocodificación por lotes sería el procesamiento de cinco direcciones en cada lote. Esto trae consigo que sea necesario un pre-procesamiento de la información para dividirla en lotes pequeños cuando se quieran geocodificar grandes cantidades de direcciones con la configuración recomendada por el autor, lo que conllevaría a un esfuerzo mayor por parte del usuario y, por ende, un consumo mayor de tiempo en todo el proceso.

El componente al que se hace referencia, hace uso del sistema gestor de bases de datos Oracle para el almacenamiento de la base de datos de referencia necesaria en la geocodificación de direcciones postales. Parte de la lógica del proceso de geocodificación se encuentra anclada a la base de datos mediante procedimientos al-

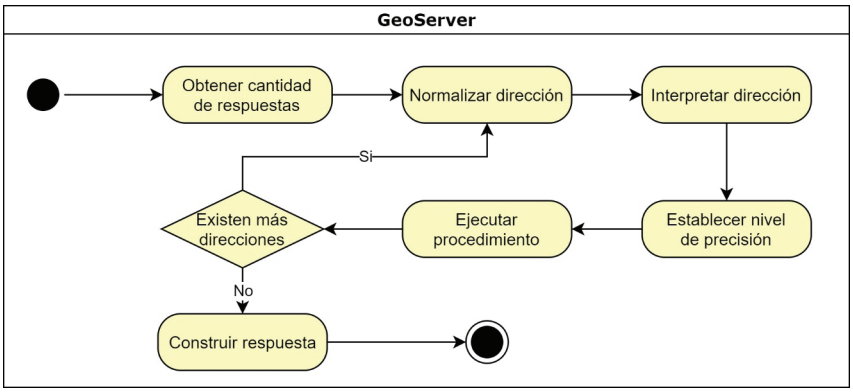

Figura 1. Versión secuencial de la geocodificación de direcciones postales

macenados de acuerdo con lo que plantea el autor. Esta lógica se encuentra duplicada en una base de datos PostgreSQL, con su extensión espacial PostGIS, debido a la necesidad de contar con una fuente de datos que utilice tecnologías libres. Esto trae consigo diversos problemas, por ejemplo, si se necesitara utilizar otro sistema de base de datos o actualizar los actuales, sería necesario no solo migrar o actualizar los datos utilizados en el proceso de geocodificación, sino que habría que actualizar parte de la lógica del proceso en todos los sistemas donde se encuentre, lo que conllevaría un esfuerzo mayor por parte de los desarrolladores.

Particularidades de la geocodificación de direcciones postales

La geocodificación de direcciones, según el estándar OpenLS, se puede realizar en lotes de más de una dirección. Una petición que sigue el estándar OpenLS y que contiene más de una dirección a geocodificar se muestra en la Figura 2 (izquierda).

Cuando se recibe una petición como la que se muestra en el código anterior, se geocodifica en un primer instante la dirección: *"60 # 109 Apto 2 E/ Ave 1 A Y Ave 1, Miramar, Playa, La Habana"* y cuando finaliza ese proceso comienza la geocodificación de la dirección: *"340 E/ 107 Y 109, Rpto Alt Naranjal, Munic Matanzas, Matanzas"*. Para cada una de las direcciones contenidas en el lote se debe ofrecer una respuesta al usuario.

Como se puede observar en la Figura 2 (derecha), la respuesta al cliente está compuesta por dos etiquetas GeocodeResponseList dentro de las cuales se encuentran las direcciones geocodificadas para cada una de las 2 direcciones enviadas al servidor para su geocodificación. Se le proporciona una respuesta al cliente que contiene la posición geográfica en formato longitud/latitud y la dirección encontrada escrita en formato libre de acuerdo con el modelo de direcciones cubano. Estas direcciones no dependen una de la otra para su geocodificación, por lo que este proceso se pudiera desarrollar en paralelo.

Tomando como punto de partida el sistema de geocodificación que se plantea en Sánchez (2015), se diseñaron varias versiones del componente de acuerdo con si se emplea o no computación paralela en su funcionamiento, o si la lógica del proceso de geocodificación se encuentra anclada o no a la base de datos de referencia para la geocodificación. Dichas versiones se muestran en la Tabla 1.

Diseño paralelo del algoritmo para la geocodificación de direcciones postales

Los algoritmos de geocodificación secuenciales planteados realizan el proceso de geocodificación sobre una dirección a la vez. Esto trae consigo demoras en el proceso de geocodificación que pueden ser mitigadas utilizando un modelo de paralelismo de datos. Utilizando el paralelismo de datos se puede aplicar el algoritmo de geocodificación sobre una determinada cantidad de direcciones en forma simultánea.

Este diseño se muestra en el diagrama de la Figura 3. El proceso se inicia mediante una aplicación cliente, encargada de construir una petición de geocodificación utilizando el estándar OpenLS para su posterior procesamiento y haciendo uso del algoritmo paralelo implementado para GeoServer. Utilizando herramientas y bibliotecas nativas de Java se identifica la cantidad de elementos de procesamientos que contiene físicamente el servidor sobre el cual se está ejecutando GeoServer, por lo tanto, el algoritmo de geocodificación. Esta cantidad de elementos de procesamiento será el número de direcciones simultáneas que podrán ser procesadas por el algoritmo.

Posteriormente, las direcciones se normalizan con el objetivo de eliminar cualquier carácter especial que pu-

https://doi.org/10.22201/fi.25940732e.2020.21.4.029


```
<?xml version="1.0" encoding="UTF-8" standaolone="no"?>
<XLS version="1.2" xmlns="http://www.opngis.net/xls">
    <RequestHeader/>
    <Request maximunResponses="10"
             methodName="GeocodeService"
             requestID="1.2">
        <GeocodeRequest>
            <Address>
                <freeFormAddress>
                    60 #109 APTO 2 E/ AVE 1A y Ave 1,
                    MIRAMAR, PLAYA, LA HABANA
                </freeFormAddress>
            </Address>
                    <Address>
                <freeFormAddress>
                    340 E/ 107 Y 109, RPTO ALT NARANJAL,
                    MUNIC MATANZAS, MATANZAS
                </freeFormAddress>
            </Address>
        </GeocodeRequest>
    </Request>
</XLS>
```

```
<GeocodeResponse>
    <GeocodeResponseList numberOfGeocodedAddresses="1">
        <GeocodedAddress>
            <Address>
                <freeFormAddress>
                    Calle 60 entre Avenida 1 y Avenida 1A,
                    Miramar, Playa, La Habana
                </freeFormAddress>
            </Address>
            <Point srsName="EPSG:4326">
                <pos>-82.435526 23.116071</pos>
            </Point>
        </GeocodedAddress>
    </GeocodeResponseList>
    <GeocodeResponseList numberOfGeocodedAddresses="1">
        <GeocodedAddress>
            <Address>
                <freeFormAddress>
                    Calle 346 entre Calle 107 A y Calle 109,
                    Alturas de Naranjal, Matanzas
                </freeFormAddress>
            </Address>
            <Point srsName="EPSG:4326">
                <pos>-81.598757 23.037505</pos>
            </Point>
        </GeocodedAddress>
    </GeocodeResponseList>
</GeocodeResponse>
```

Figura 2. Formatos de geocodificación de múltiples direcciones. Izquierda: petición, derecha: respuesta

Tabla 1. Versiones del sistema de geocodificación de direcciones postales

Versión	Características
1	Lógica del proceso de geocodificación anclado a la base de datos de referencia, con un diseño secuencial del sistema.
2	Lógica del proceso de geocodificación independiente a la base de datos de referencia, con un diseño secuencial del sistema.
3	Lógica del proceso de geocodificación anclado a la base de datos de referencia, con un diseño paralelo del sistema.
4	Lógica del proceso de geocodificación independiente a la base de datos de referencia, con un diseño paralelo del sistema.

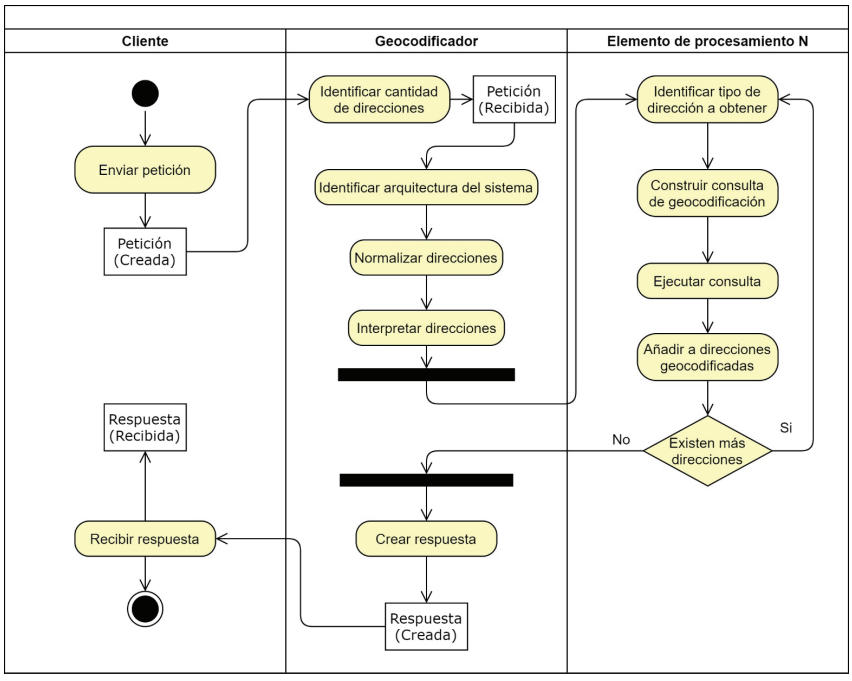

Figura 3. Versión paralela de la geocodificación de direcciones postales utilizando N elementos de procesamiento

dieran contener y así evitar algunos errores en su interpretación. A cada dirección se le realiza un proceso de interpretación donde se le extraen sus componentes o TOKENS, utilizando un módulo de interpretación de direcciones. Luego, el conjunto de direcciones se divide en la forma más homogénea posible para su geocodificación, utilizando cada elemento de procesamiento del sistema. De forma simultánea, sobre cada conjunto de direcciones, el algoritmo permite establecer el tipo de dirección a geocodificar y se construye una solicitud de geocodificación para ser comparada con los datos de referencia. Esta solicitud de geocodificación es la que puede variar de acuerdo con la versión del sistema que se esté empleando debido a que la lógica de este proceso puede encontrarse anclada o no a la base de datos de referencia.

Finalmente, del conjunto de direcciones posibles se selecciona la que mayor concordancia presente de acuerdo con la dirección de entrada y se culmina el proceso obteniendo una lista de direcciones geocodificadas en el mismo orden, las cuales se recibieron en los datos de entrada.

En la Figura 4 se muestra un diagrama de estructuración en capas que muestra la arquitectura de los componentes fundamentales del sistema mediante paquetes y subsistemas de acuerdo con el nivel de reutilización que pudieran tener.

El sistema está compuesto por varios paquetes y cada uno de estos contiene una o más clases dedicadas, en conjunto, al proceso de geocodificación de direcciones postales cubanas. El paquete "request" contiene la gestión de la configuración necesaria para que el componente pueda procesar una petición de geocodificación, incluyendo los parámetros necesarios para que el componente se pueda conectar a la base de datos de referencia. El paquete "geocode" es el contenedor de las clases y métodos necesarios para el correcto procesamiento de una dirección postal. El paquete "query Builder" es el contenedor de la totalidad de la lógica del proceso de geocodificación para las versiones 2 y 4 del sistema; finalmente en la versión 3 es la encargada de ejecutar la petición a la base de datos de referencia para que continúe con el proceso de geocodificación.

El paquete "connection" modelado en la Figura 4 es el encargado del proceso de conexión a cada una de las bases de datos de referencia, tanto en Oracle como en Postgre, mientras que el subsistema "OpenLSCore" contiene todas las clases que describen los elementos que se encuentran en la especificación OpenLS de OGC necesarias para la correcta construcción e interpretación de los XML de petición y respuesta a los LBS. El subsistema "AddressParser" provee al sistema la capacidad de interpretar direcciones postales cubanas y obtener los elementos de interés de cada dirección para su posterior geocodificación (Girón & Sánchez, 2017).

Como el sistema se basa en un modelo de paralelismo de datos, los paquetes diseñados corresponden prácticamente con los de un sistema secuencial. La diferencia radica en que, al recibirse una petición con múltiples direcciones, se crean tantos hilos de ejecución como elementos de procesamiento que se hayan identificado para procesar dichas direcciones. Este proceso ocurre en el paquete "geocode" y se extiende hacia los paquetes "queryBuilder" y "connection", pues las cla-

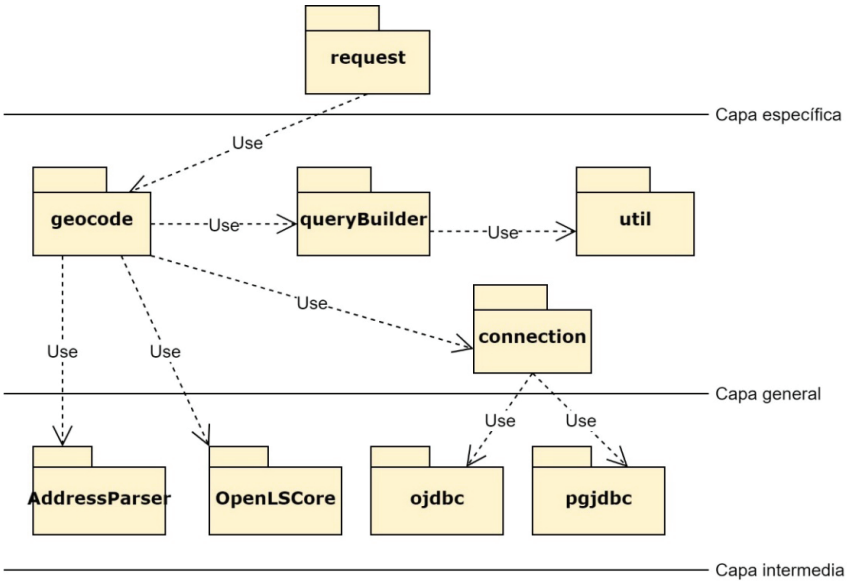

Figura 4. Vista de la arquitectura del software desde un punto de vista de reutilización

ses de estos dos últimos son instanciadas para cada hilo de ejecución.

Evaluación

Luego del desarrollo del software es necesaria la realización de un conjunto de pruebas con las que se puedan comprobar el comportamiento del componente ante determinadas circunstancias. Con el objetivo de comparar el comportamiento de las cuatro versiones del sistema de geocodificación de direcciones postales cubanas, en cuanto al tiempo de procesamiento, se identificaron varios factores controlables y no controlables para la ejecución de las pruebas. A continuación, se definen los factores controlables identificados:

- Cantidad de direcciones enviadas: Es la cantidad de direcciones que se envía en el fichero de la petición al servicio de geocodificación.
- Cantidad de direcciones recibidas: Es la cantidad de direcciones que se reciben en el fichero de la respuesta que se obtiene del servicio de geocodificación para cada dirección solicitada.
- Cantidad de ejecuciones (Ciclos): Es la cantidad de veces que se repetirá la prueba. Al recibir la respuesta a una petición, se envía la nueva petición correspondiente a la siguiente ejecución.

Dentro de los factores no controlables se señalan:

- Tráfico de la red: Las fuentes de datos se encuentran ubicadas en unidades de la red del centro donde se ejecutan los experimentos. En el momento en que se realizan las peticiones no se puede controlar el tráfico que existe en la red.
- Procesos del sistema operativo: En el momento en que se realizan las peticiones al servidor hay procesos ejecutándose en el sistema operativo del servidor que consumen recursos y no pueden ser detenidos.

De acuerdo con los factores controlables del experimento se define como la cantidad de direcciones enviadas en una petición como 10,000, la cantidad máxima de direcciones de respuesta serán 10 para cada una de las direcciones enviadas y se realizarán 10 réplicas de la prueba para cada versión del sistema.

La unidad experimental que se utilizará para el despliegue de los sistemas desarrollados contiene un microprocesador Intel (R) Core (TM) i5-4200U @ 1.60HGz 2.30 GHz, 8 GB de memoria Ram, 1 TB de almacenamiento HDD y utiliza el sistema operativo Windows 8.1 Pro. La base de datos de referencia a utilizar se encuentra en 2 servidores en Real Application Cluster de Oracle que cuentan con 2 CPU AMD(R) Opteron @ 2.60GHz, 6 GB de memoria Ram, 70 GB de almacenamiento HDD y utilizan el sistema operativo SUSE Linux Enterprise Server 11. 64 bits.

Los resultados de la ejecución de la prueba utilizando las versiones secuenciales y las versiones paralelas del sistema se muestran en la Tabla 2.

Con el objetivo de demostrar que las nuevas variantes del sistema mejoran el tiempo de ejecución de la variante original y establecen comparaciones entre ellas de acuerdo con los resultados obtenidos, se utilizarán pruebas estadísticas. Dentro de las pruebas de hipótesis se encuentra la prueba de Kruskal-Wallis, que se utiliza para determinar si las medianas de dos o más muestras difieren. Esta prueba se utiliza para comprobar si existen diferencias entre los tiempos de ejecución de cada una de las variantes del algoritmo. Para ello se definen como hipótesis las siguientes:

Tabla 2. Tiempo de ejecución de cada variante del sistema en minutos

Petición	Versión 1 (min)	Versión 2 (min)	Versión 3 (min)	Versión 4 (min)
1	200.7251	187.1409	52.02323	45.50525
2	201.1681	188.0125	47.1555	48.82272
3	197.1450	191.1899	46.88175	40.2178
4	215.1457	198.088	47.26227	40.6072
5	213.394	197.3034	47.17183	40.57968
6	199.4715	188.8933	47.37605	40.74093
7	214.8107	201.9182	47.24465	49.06213
8	212.0311	203.3355	52.40212	45.75048
9	197.2248	187.1072	48.23388	41.58937
10	195.5374	193.0839	47.97737	41.40887

- H_0: Los sistemas son iguales en cuanto a tiempo de respuesta.
- H_1: Al menos uno de los sistemas es diferente al resto en cuanto a su tiempo de respuesta.

H_0 corresponde a la hipótesis nula y H_1 es la hipótesis alternativa. Se utiliza un $\alpha = 0.05$ (nivel de significación). En la Tabla 3 se muestra el resultado de realizar la prueba usando el software Minitab 16.

Al realizar la prueba se obtuvo un valor p = 0.000 como resultado de un redondeo. Como este valor es menor que α se rechaza la hipótesis nula y se acepta la hipótesis alternativa, por lo que se puede concluir que los sistemas difieren en cuanto al tiempo de respuesta.

Una vez comprobado que los tiempos de respuesta de los sistemas difieren entre sí se pasa a comprobar que las nuevas implementaciones (versiones 2, 3 y 4) ofrecen mejores prestaciones que la versión 1 en cuanto a tiempo de ejecución. La demostración se realizará apoyándose de la prueba de hipótesis Mann-Whitney.

Cuando se realizan dos o más comparaciones entre muestras para determinar cuál de ellas es menor, mayor o igual que las otras, la probabilidad de rechazar una hipótesis nula que en realidad es verdadera (error de tipo 1) es mayor que α debido a la repetición de la prueba con un nivel de significancia determinado, por lo que se debería utilizar un método que ajuste el nivel de significancia. Según García & Lara (1998) uno de los métodos que tratan este problema es el método de Bonferroni que plantea ajustar el nivel de significancia a α/m siendo m la cantidad de comparaciones que se realizarán.

En un primer momento se realizarán las comparaciones necesarias para demostrar que todas las versiones nuevas del sistema de geocodificación son más rápidas que la versión original. Para ello se define un $\alpha = 0.05$ y el número de comparaciones a realizar 3 por lo que $\alpha_{ajustada} = 0.017$.

Comparación entre la versión 1 y la versión 2

Para determinar si existen diferencias entre los tiempos de ejecución de las variantes 1 y 2 se definen como hipótesis las siguientes:

- H_0**:** Las versiones 1 y 2 son iguales en cuanto a tiempo de respuesta.
- H_1: La versión 1 del algoritmo presenta tiempos de ejecución mayores que la 2.

El resultado de la prueba se muestra en la Tabla 4.

Al realizar la prueba se obtuvo un valor p = 0.0086, que es menor que el $\alpha_{ajustada} = 0.017$ por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

De forma similar se utilizó la prueba de hipótesis de Mann-Whitney para realizar comparaciones entre la versión 1 y las versiones 3 y 4 del sistema y en todos los casos se obtuvo un valor p menor que $\alpha_{ajustada}$ por lo que

Tabla 3. Resultado de la ejecución de la prueba de Kruskal-Wallis

Prueba de Kruskal-Wallis: Tiempo *vs*. Versión del Sistema				
Versión	N	Mediana	Clasificación del promedio	Z
Versión 1	10	200.95	33.7	4.12
Versión 2	10	192.14	27.3	2.12
Versión 3	10	47.32	13.9	-2.06
Versión 4	10	41.50	7.1	-4.19
General	40		20.5	
H = 32.46		GL = 3	P = 0.000	

Tabla 4. Prueba Mann-Whitney para comparar las versiones 1 y 2 del sistema de geocodificación

Prueba de Mann-Whitney e IC: Versión 1; Versión 2		
Versión	N	Mediana
Versión 1	10	200.95
Versión 2	10	192.14
La estimación del punto para ETA1-ETA2 es 10.12		
95.5 El porcentaje IC para ETA1-ETA2 es (3.42;18.95)		
W = 137.0		
Prueba de ETA1 = ETA2 vs. ETA1 > ETA2 es significativa en 0.0086		

todas las versiones nuevas del sistema disminuyen los tiempos de ejecución del sistema de geocodificación de direcciones postales.

Con el objetivo de verificar cuál de las tres nuevas versiones ofrece menor tiempo de respuesta se realiza un procedimiento similar al realizado anteriormente.

Para determinar si existen diferencias entre los tiempos de respuesta de las variantes 2, 3 y 4 del sistema se realizará una prueba Kruskal-Wallis donde las hipótesis definidas son las siguientes:

- H_0: Los sistemas son iguales en cuanto a tiempo de respuesta.
- H_1: Al menos uno de los sistemas es diferente al resto en cuanto a su tiempo de respuesta.

Donde H_0 es la hipótesis nula y H_1 es la hipótesis alternativa. Se utiliza un $\alpha = 0.05$. En la Tabla 5 se muestra el resultado de realizar la prueba.

Al realizar la prueba se obtuvo un valor p = 0.000 como resultado de un redondeo. Como este valor es menor que α se rechaza la hipótesis nula y se acepta la hipótesis alternativa, por lo que se puede concluir que los sistemas difieren en cuanto al tiempo de respuesta.

Con el objetivo de encontrar cuál de las tres versiones ofrece los mejores tiempos se va a realizar un procedimiento de comparaciones múltiples donde se define como nivel de significancia $\alpha = 0.05$ y la cantidad de comparaciones a realizar 3, luego $\alpha_{ajustada} = 0.017$.

Comparación entre la versión 2 y la versión 3

Para determinar si existen diferencias entre los tiempos de ejecución de las variantes 2 y 3 se definen como hipótesis las siguientes:

- H_0: Las versiones 2 y 3 son iguales en cuanto a tiempo de respuesta.
- H_1: La versión 2 del algoritmo presenta tiempos de ejecución mayores que la 3.

El resultado de la prueba se muestra en la Tabla 6.

Al realizar la prueba se obtuvo un valor p = 0.0001, que es menor que el $\alpha_{ajustada}$ por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

Comparación entre la versión 2 y la versión 4

Para determinar si existen diferencias entre los tiempos de ejecución de las variantes 2 y 4 se definen como hipótesis las siguientes:

- H_0: Los sistemas son iguales en cuanto a tiempo de respuesta.
- H_1: La versión 2 del algoritmo presenta tiempos de ejecución mayores que la 4.

Como se observa en la Tabla 7, se obtuvo un valor p = 0.0001, que es menor que el $\alpha_{ajustada}$ por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

Tabla 5. Prueba Kruskal-Wallis para comparar las versiones 2, 3 y 4 del sistema de geocodificación

Prueba de Kruskal-Wallis: Tiempo *vs*. Versión				
Versión	N	Mediana	Clasificación del promedio	Z
Versión 2	10	192.14	27.3	4.40
Versión 3	10	47.32	13.9	–0.70
Versión 4	10	41.50	7.1	–3.70
General	30		15.5	
H = 22.34		GL = 2	P = 0.000	

Tabla 6. Prueba Mann-Whitney para comparar las versiones 2 y 3 del sistema de geocodificación

Prueba de Mann-Whitney e IC: Versión 2; Versión 3		
Versión	N	Mediana
Versión 2	10	192.14
Versión 3	10	47.32
La estimación del punto para ETA1-ETA2 es 144.03		
95.5 El porcentaje IC para ETA1-ETA2 es (139.97;150.71)		
W = 155.0		
Prueba de ETA1 = ETA2 vs. ETA1 > ETA2 es significativa en 0.0001		

Tabla 7. Prueba Mann-Whitney para comparar las versiones 2 y 4 del sistema de geocodificación

Prueba de Mann-Whitney e IC: Versión 2; Versión 4		
Versión	N	Mediana
Versión 2	10	192.14
Versión 4	10	41.50

La estimación del punto para ETA1-ETA2 es 148.85
95.5 El porcentaje IC para ETA1-ETA2 es (145.68;156.50)
W = 155.0
Prueba de ETA1 = ETA2 vs. ETA1 > ETA2 es significativa en 0.0001

Comparación entre la versión 3 y la versión 4

Para determinar si existen diferencias entre los tiempos de ejecución de las variantes 3 y 4 se definen como hipótesis las siguientes:

- H_0: Las versiones 3 y 4 son iguales en cuanto a tiempo de respuesta.
- H_1: La versión 3 del algoritmo presenta tiempos de ejecución mayores que la 4.

El resultado de la prueba se muestra en la Tabla 8.

Al realizar la prueba se obtuvo un valor p = 0.0057, que es menor que el $\alpha_{ajustada}$ por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

De forma gráfica se puede observar el comportamiento respecto al tiempo de las diferentes versiones del sistema en la Figura 5, donde se observa una diferencia significativa entre las versiones secuenciales del sistema (versiones 1 y 2) respecto a las versiones paralelas del mismo (versiones 3 y 4).

De acuerdo con las versiones 3 y 4 del sistema en la Figura 6 se puede observar que en determinadas instancias el tiempo de procesamiento con ambos sistemas paralelos fue similar, pero en la mayoría de las ejecuciones, el tiempo que se demoró el proceso de geocodificación con la versión 4 fue inferior que con la versión 3.

Como resultado de las pruebas realizadas se obtuvo que la versión 4 es la que más reduce los tiempos de respuesta del sistema de geocodificación. Dicha versión corresponde a la implementación paralela con la lógica del proceso de geocodificación independiente a la base de datos de referencia. No obstante, se recomienda la realización de otras pruebas para verificar cómo se comporta la versión 4 del sistema con lotes pequeños de direcciones, incluyendo la geocodificación de una sola dirección.

Asimismo, es recomendable realizar una nueva comparación, similar a la realizada en De Armas & Gutierrez (2013), entre la mejor versión obtenida en el presente trabajo y un servicio online de geocodificación, como el API de geocodificación de Google o Nominatim de OSM, con el fin de verificar la calidad de las direcciones encontradas y al mismo tiempo, comparar los tiempos de respuesta de cada servicio; ya que, en este momento, dicha comparación no fue posible debido a la calidad de los datos en Cuba de estos servicios. Esta carencia trae consigo una alta probabilidad de que se queden muchas direcciones sin encontrar, lo que falsearía los resultados para comparar.

Conclusiones

A partir de la realización de este trabajo se puede concluir que los objetivos y tareas planteadas fueron cumplidos. Además, se arriba a las siguientes conclusiones:

- La utilización de la geocodificación de direcciones postales en lote es muy utilizada en el mundo, creciendo cada vez más las necesidades de geocodificación.

Tabla 8. Prueba Mann-Whitney para comparar las versiones 3 y 4 del sistema de geocodificación

Prueba de Mann-Whitney e IC: Versión 3; Versión 4		
Versión	N	Mediana
Versión 3	10	47.319
Versión 4	10	41.50

La estimación del punto para ETA1-ETA2 es 6.274
95.5 El porcentaje IC para ETA1-ETA2 es (1.650;6.954)
W = 139.0
Prueba de ETA1 = ETA2 vs. ETA1 > ETA2 es significativa en 0.0057

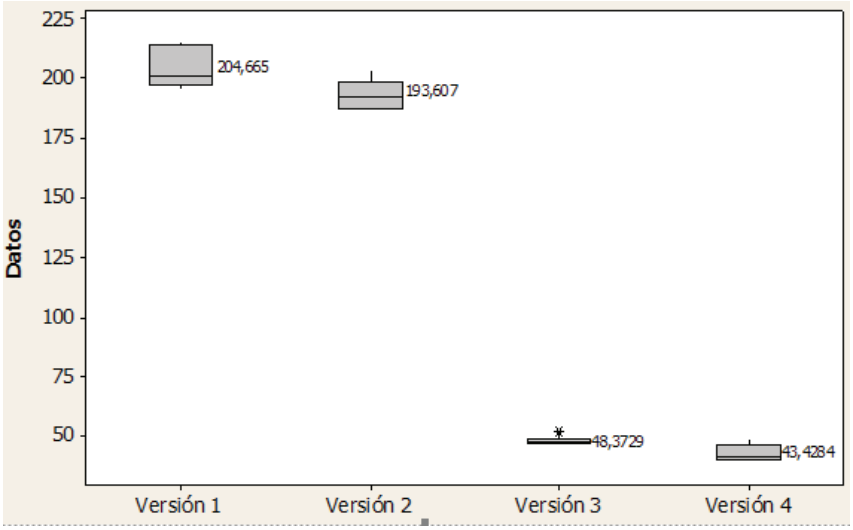

Figura 5. Gráfica de caja del tiempo de procesamiento de las versiones del sistema

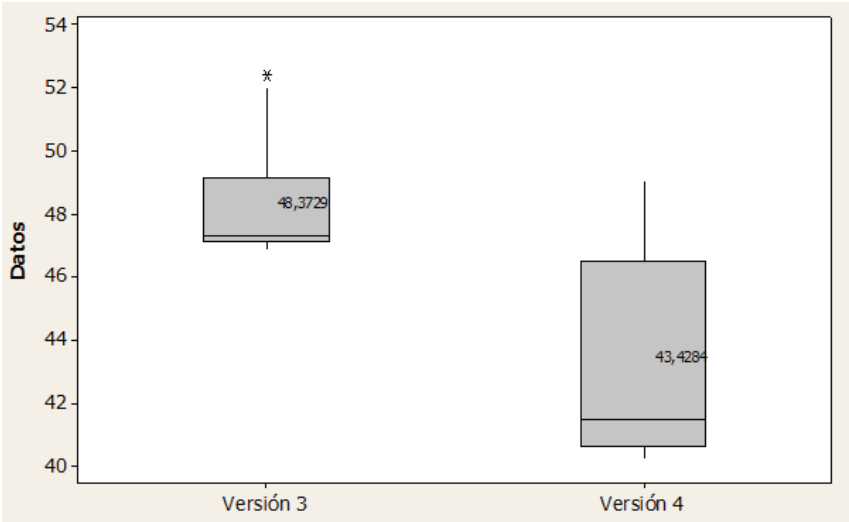

Figura 6. Gráfica de caja del tiempo de procesamiento de las versiones 3 y 4 del sistema

- La utilización de técnicas de computación paralela es una variante válida para disminuir el tiempo de procesamiento de la geocodificación de direcciones en lotes.
- Se implementaron tres variantes de geocodificación de direcciones postales cubanas, dos con un diseño paralelo y otra con un diseño secuencial, que permiten disminuir los tiempos de respuesta y mantener la calidad y cantidad de direcciones encontradas de los servicios anteriores.
- La versión paralela del sistema de geocodificación postales cubanas, que contiene la lógica del proceso de geocodificación independiente a la base de datos, es la que más reduce los tiempos de procesamiento de direcciones postales en lote.

Referencias

Burrough, P. A. (1986). *Principles of geographical information systems for land resources assessment* (Vol. 12). New York: Clarendon Press.

Cetl V. Kliment T. &Jogun T. (2016). A comparison of address geocoding techniques-case study of the city of Zagreb, Croatia. *Survey Review Ltd*, 11. https://doi.org/10.1080/00396265.2016.1252517

Chopin, J. & Caneppele, S. (2019). Geocoding child sexual abuse: An explorative analysis on journey to crime and to victimization from French police data. *Child Abuse & Neglect*, 91, 116-130. https://doi.org/10.1016/j.chiabu.2019.03.001

De Armas, C. J. & Gutierrez, A. A. (2013). Deployment of a National Geocoding Service: Cuban Experience. *URISA Journal*, *25*.

Dueker, K. J. (1974). Urban Geocoding. *Annals of the association of american geographers*, 64.

Foster, I. (1995). *Designing and building parallel programs: Concepts and tools for parallel software engineering*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

García, J. & Lara, A.M. (1998). *Diseño estadístico de experimentos. Análisis de la varianza.*

GeoServer. (2014). GeoServer User Manual Release.

Girón, L. & Sánchez, E. (2017). *Servicios basados en la localización para GeoServer.* Paper presented at the VIII Convención Agrimensura, La Habana.

Jiang, W. & Stefanakis, E. (2018). What 3Words Geocoding Extensions. *Journal of Geovisualization and Spatial Analysis, 2*(1), 7. https://doi.org/10.1007/s41651-018-0014-x

Küçük, D. & Avdan, U. (2018). Address standardization using the natural language process for improving geocoding results. *Computers, environment and urban systems,* 70, 1-8. https://doi.org/10.1016/j.compenvurbsys.2018.01.009

McIntire, R. K., Keith, S. W., Boamah, M., Leader, A. E., Glanz, K., Klassen, A. C. & Zeigler-Johnson, C. M. (2018). A Prostate cancer composite score to identify high burden neighborhoods. *Preventive Medicine,* 112, 47-53. https://doi.org/10.1016/j.ypmed.2018.04.003

Open Geospatial Consortium. (2008). OpenGIS Location Services (OpenLS): Core Services. In E. Marwa Mabrouk (Ed.): OGC.

Oracle. (2007). *Oracle Spatial 11g: Administración avanzada de datos espaciales para aplicaciones empresariales*. United States of America.

Pacheco, P. S. (2011). *An introduction to parallel programming*. Burlington, USA: Morgan Kaufmann Publishers.

Präger, M., Kurz, C., Böhm, J., Laxy, M. & Maier, W. (2019). Using data from online geocoding services for the assessment of environmental obesogenic factors: a feasibility study. *International Journal of Health Geographics,* 18(1), 13. https://doi.org/10.1186/s12942-019-0177-9

Roongpiboonsopit, D. & Karimi, H. A. (2010). Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science,* 24(7), 1081-1100. https://doi.org/10.1080/13658810903289478

Sánchez, E. (2015). *Servicios basados en localización para una infraestructura de datos espaciales.* (Tesis en opción al título de máster en Informática Aplicada). Instituto Superior "José Antonio Echeverría" (ISPJAE), Habana, Cuba.

Sánchez, L. (2012). *Servicio de directorio para consultas basadas en la localización.* (Trabajo de diploma para optar por el título de Ingeniería Informática). Instituto Superior Politécnico José Antonio Echeverría, Habana, Cuba.

Sen Xu, Soren Flexner & Vitor Carvalho. (2012, Septembre 18-12, 2012). *Geocoding billions of addresses: toward a spatial record linkage system with big data.* Paper presented at the GIScience in the Big Data Age (GIScience 2012), Colombus, OH, USA.

Vargas, J. A. & Horfan, D. (2013). Proceso de Geocodificación de direcciones en la Ciudad de Medellín, una técnica determinística de georreferenciación de direcciones. *Ing. USBMed,* 4, 15.

Yin, Z., Ma, A. & Goldberg, D. W. (2019). A deep learning approach for rooftop geocoding. *Transactions in GIS,* 23(3), 495-514. https://doi.org/10.1111/tgis.12536