INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

LABORATORIO DE PROCESAMIENTO INTELIGENTE DE INFORMACIÓN GEOESPACIAL

Geocodificación semántica

Tesis

Que para obtener el grado de

Doctorado en Ciencias de la Computación

Presenta

Imelda Escamilla Bouchan

Directores de tesis

Dr. Marco Antonio Moreno Ibarra
Dr. Miguel Jesús Torres Ruíz

México, D.F., Julio de 2016

SIP-14 BIS

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 16 del mes de junio de 2016 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Geocodificación semántica"

Presentada por la alumna:

ESCAMILLA	**BOUCHÁN**	**IMELDA**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	1	2	1	0	0	0

aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de tesis

Dr. Miguel Jesús Torres Ruiz

Dr. Marco Antonio Moreno Ibarra

Dr. Oleksiy Pogrebnyak

Dr. Grigori Sidorov

Dr. Rolando Quintero Téllez

Dr. José Giovanni Guzmán Lugo

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

ESTADOS UNIDOS MEXICANOS
INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉECNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de __México___el día _21_del mes_junio__del año _2016_, la que suscribe_Imelda Escamilla Bouchán__ alumna del Programa de_Doctorado en Ciencias de la Computación__con número de registro _B121000_, adscrita al _Centro de Investigación en Computación_, manifiesta que es autora intelectual del presente trabajo de Tesis bajo la dirección del _Dr. Miguel Jesús Torres Ruiz y el Dr. Marco Antonio Moreno Ibarra__ y cede los derechos del trabajo intitulado _"Geocodificación semántica"__, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección __ebouchana10@sagitario.cic.ipn.mx__. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Imelda Escamilla Bouchán

Nombre y firma

Resumen

La capacidad humana de entender las referencias aproximadas a lugares, mediante la desambiguación del contexto y el razonamiento de las relaciones espaciales, es la clave para describir los entornos espaciales y para compartir información sobre ellos. En este trabajo, se propone un enfoque para la geocodificación que utiliza las relaciones espaciales contenidas en el texto de los *tweets*, empleando análisis semánticos y espaciales. El texto de los *microblog* tiene características especiales (por ejemplo, la jerga, abreviaturas, acrónimos, etc.) por lo tanto representa una variación especial de lenguaje natural. El principal objetivo de este trabajo consiste en asociar las relaciones espaciales que se encuentran en el texto con un lugar, para determinar la ubicación del evento descrito en el *tweet*. Se demuestra la viabilidad de la propuesta usando un corpus de 200,000 *tweets* publicados en español relacionados con eventos viales en la Ciudad de México.

Abstract

Human ability to understand approximate references to locations, disambiguated by means of context and reasoning about spatial relationships, is the key to describe spatial environments and to share information about them. In this paper, we propose an approach for geocoding that takes advantage of the spatial relationships contained in the text of tweets, using semantic and spatial analyses. Microblog text has special characteristics (e.g. slang, abbreviations, acronyms, etc.) and thus represents a special variation of natural language. The main objective of this work is to associate spatial relationships found in text with a spatial footprint, **to determine the location of the event** described in the tweet. The feasibility of the proposal is demonstrated using a corpus of 200,000 tweets posted in Spanish related with traffic events in Mexico City.

Agradecimientos

Agradesco al Instituto Politécnico Nacional y al Centro de Investigación en Computación, principalmente a mis asesores Miguel y Marco por todo su apoyo.
Agradesco a mis papas y mi hermana por estar siempre conmigo y a Vladimir que todo el tiempo me apoya y me impulsa a ser una major persona cada día.
Gracias a todos mis amigos, en especial a Luis y a todos aquellos que estuvieron involucrados directa o indirectamente para que llegara hasta aquí.
¡Gracias Totales!

Índice

Índice de Figuras

Índice de Tablas

1. Introducción

Usualmente, la forma más común de describir entornos espaciales en lenguaje natural es utilizando relaciones espaciales. Estas relaciones espaciales describen las relaciones correspondientes que existen entre los objetos en el espacio (Laurini, 2012), considerando una ciudad y las relaciones espaciales entre objetos urbanos las relaciones topológicas (Allen, 1983) y las relaciones de Egenhofer (Egenhofer, 1994) no contienen la suficiente semántica para describir las relaciones entre las características geográficas y los objetos urbanos.

Sin embargo, el Lenguaje Natural cuenta con la habilidad humana para entender las referencias aproximadas a lugares, desambiguando por medio del contexto. Por ejemplo, las referencias indirectas a lugares que se incluyen al momento de dar indicaciones, incluyendo la ubicación de lugares facilmente reconocibles o lugares representativos Delboni et al. (2007).

No obstante, cuando la descripción en lenguaje natural contiene algún tipo de argot, abreviaturas, acrónimos, faltas de ortografía y es limitada a un pequeño conjunto de caracteres, la detección de las relaciones espaciales se convierte en una tarea compleja. Tal es el caso de *Twitter,* plataforma que permite mostrar cualquier tipo de información a través de opiniones e ideas que mantienen a las personas actualizando o informando los eventos cotidianos.

Los usuarios pueden publicar y leer textos cortos, de máximo 140 caracteres de longitud o aproximadamente 25 palabras, conocidos como “*tweets*”. También pueden hacer un *Retweet* o *RT* como popularmente se conoce, que significa compartir el mensaje de su interés a los usuarios que se encuentren en su red.

Diariamente son publicados millones de *tweets* relacionados con eventos viales, tanto por usuarios particulares como por instituciones públicas o privadas. Cada *tweet* describe eventos tales como accidentes, bloqueos en vialidades, reportes de tránsito, entre otros, y su ubicación particular. Recuperar esta

ubicación del texto crudo no es una labor sencilla, por que el lenguaje natural no sigue un formato o estandar y para ser de utilidad, la ubicación de este tipo de eventos debe ser estimada con la mayor precisión posible.

En este trabajo, se propone una metodología para determinar la ubicación de los eventos descritos en los *tweets,* asociando las relaciones espaciales que se encuentran en el texto con un lugar. La viabilidad de la propuesta se demostró utilizando un corpus de 200,000 *tweets* publicados en español relacionados con eventos viales en la Ciudad de México.

1.1 Descripción del problema

Mediante el uso de las redes sociales, las personas fácilmente pueden comunicarse y publicar cualquier cosa. Por ejemplo, en los últimos años, *Twitter* se ha convertido en una plataforma de *microblogging* muy popular con más de 400 millones de *tweets* publicados diariamente[1]. Esto ha impulsado numerosos esfuerzos de investigación con diversas temáticas para explotar esta información, tales como detección de eventos (Agarwal et al., 2012; Atefeth & Khreich, 2015), monitoreo de salud (Nielsen et al., 2015), detección de emergencias (Seol et al., 2013), etc. Muchas de estas aplicaciones pueden ser beneficiadas con la información referente a la ubicación donde ocurre el evento, pero desafortunadamente esta información es muy escasa puesto que solamente el 1% de los *tweets* contienen una etiqueta geográfica (Takhteyev et al., 2012), esto significa que el 99% de los *tweets* restantes serían descartados.

Los enfoques simples para determinar la ubicación de un *tweet* no son aplicables, por ejemplo: La ubicación no puede ser estimada utilizando la dirección IP del dispositivo, ya que ni *Twitter* ni el proveedor de telecomunicaciones permitirían el acceso a esta información debido a sus políticas de privacidad.

[1] https://www.about.twitter.com/company

Por otro lado, el API de *Twitter* proporciona búsquedas de lugares empleando filtros espaciales, pero se basa únicamente en la información del perfil del usuario que hizo la publicación, que a menudo son incompletas o incorrectas (Hecht et al., 2011).

Por lo tanto, la extracción de la información geográfica por otros medios se considera un reto debido a:

1. Los nombres de lugares (también llamados topónimos) tienen que ser identificados en el texto del *tweet* o en sus metadatos.
2. Los *tweets* son cortos (máximo 140 caracteres) y comúnmente se encuentran escritos con una gran cantidad de errores gramaticales. Esto significa que las personas utilizan nombres o abreviaturas para las ubicaciones, por ejemplo "marina nal " en lugar de "Avenida Marina Nacional". El siguiente texto es un ejemplo de un *tweet* publicado comúnmente y se observa claramente el uso de abreviaturas y los errores gramaticales: "*RT:@manuelivanr: Sigue cerrado rio rhin? que alternativas tengo si vendo de marina nal por el NTE y voy a alvaro obrego*".
3. El uso de jergas sociales es común en los *tweets*. Esto significa que los usuarios emplean términos que tienen significados distintos dependiendo del grupo social al que pertenezca o de la región de donde sean originarios, por ejemplo "troca" o "picap" para hacer referencia a una camioneta.
4. El espacio limitado de aproximadamente 25 palabras que nos proporciona *Twitter*, requiere ser breve al escribir, dando lugar a la creación de un diccionario informal de palabras y abreviaturas solamente usadas en las redes sociales, por ejemplo: "OMG, LOL, RT".
5. Por la naturaleza de la información descrita en los *tweets* es común que sea incompleta existiendo dos motivos principales;

a. El primero es que el autor del mensaje no detalle la información compartida y quede incompleta, como en el caso del *tweet* "*La Diana, también fue rehabilitada*" en donde no se da ningún otro detalle del acontecimiento o el lugar.

b. En el segundo caso, suponiendo que un *tweet* tenga el tamaño máximo permitido y sea *retweeteado* se agrega **RT @usuario: @usurio2** al mensaje provocando que el mensaje original sea cortado y quede incompleto como se muestra en el siguiente *tweet*: "*RT @OrientadorVial: ZS Continúa cerrada la circulación del cruce de Eje 7 Sur y Eje 1 Poniente, por #manifestantes Alt. Eje 7-A Sur y Divis…"*

6. Por otra parte para la asignación de topónimos, existen dos problemas principales;

   Primero, un topónimo puede hacer referencia a múltiples ubicaciones geográficas (geo desambiguación), por ejemplo, en el *tweet* "*Av. Juárez totalmente lentísima. Prever*" el topónimo "*Av. Juárez*" puede hacer referencia a los más de 20 lugares diferentes en todo México con ese nombre.

   a. Segundo, un topónimo puede estar relacionado con una ubicación geográfica pero también a una persona o alguna cosa (geo / no geo desambiguación), por ejemplo, en el *tweet* "*ZO* ***#Percancevehicular*** *en Anillo* ***#Periférico*** *a la altura de Jacarandas con dirección al Poniente."* "*Jacarandas*" puede referirse a una calle en la delegación Iztapalapa, pero también puede ser el nombre de una hotel o el de un árbol.

Este tipo de desambiguación se denomina *toponym resolution* (Leidner 2004) y es uno de los mayores retos cuando se trata de información geográfica en *microblogs*.

Para resolver la ambigüedad en los topónimos, esta propuesta se auxilia del uso de ontologías. Por ejemplo para distinguir la "*Av. Juárez*" perteneciente a la delegación Benito Juárez en el Distrito Federal a la "*Av. Juárez*" ubicada en el municipio de Nezahualcóyotl en el Estado de México, la información semántica contenida en el *tweet* puede ser de gran ayuda, como identificar algún evento cercano u otra calle de referencia.

Otro problema importante en el caso de las redes sociales es que la fuente de la que proviene la información puede ser desconocida y de esta forma nadie se hace responsable del contenido publicado.

Con base en los puntos descritos anteriormente, el presente trabajo se enfoca en geocodificar eventos que aparecen en los *tweets* referentes a eventos viales en la Ciudad de México. Esta metodología consiste en la identificación de eventos, entidades geográficas y sus relaciones espaciales, auxiliándose de una técnica basada en ontologías para la desambiguación y validación de las relaciones espaciales entre las entidades geográficas, además de proponer un método para validar la calidad de la información recuperada de los *tweets*.

1.2 Solución propuesta

Se propone diseñar e implementar una metodología capaz de geocodificar textos cortos en español, identificando eventos, características geográficas y sus relaciones espaciales, auxiliandose de representaciones conceptuales y técnicas de Procesamiento de Lenguaje Natural.

La viabilidad de la propuesta se demuestra utilizando un corpus de 200,000 *tweets* publicados en español relacionados con eventos viales en la Ciudad de México.

1.3 Justificación

En la actualidad una cantidad impresionante de información se genera a través de las redes sociales como *Twitter,* proporcionando detalles de diversos eventos a nivel mundial y actualizando dicha información a cada instante, convirtiendo a cada persona emisora de un *tweet* en un sensor, capaz de recuperar ¿dónde?, ¿cómo? y ¿cuándo? ocurre un evento (Zhang & Gelernter, 2014).

Pero es de vital importancia tener esta información disponible y conocer la mayor cantidad de detalles posibles, desafortunadamente es poco usual que los usuarios compartan su ubicación o describan detalladamente la información por la limitación en el espacio de escritura de los *tweets,* dejando pasar información valiosa para investigaciones científicas, estudios de mercado o prevención y atención de desastres.

Por ejemplo, en el caso de un *tweet* que describa un accidente vial, sería relevante conocer la zona afectada de forma detallada, al igual que la información de un tramo detenido por un automóvil descompuesto para tomar las medidas pertinentes a tiempo, pero los usuarios que redactan los mensajes no son claros, confiables y no asocian coordenadas geográficas de referencia por lo que no se puede conocer la ubicación exacta del evento.

Por tal motivo es importante contar con una metodología capaz de recuperar y validar la información contenida en los *tweets*, procesarlos, desambiguarlos y no solo permitir la visualización de un lugar mencionado en el texto, sino de la zona afectada por el evento lo cual permitirá la toma oportuna de decisiones.

1.4 Objetivo

Diseñar e implementar una metodología para la geocodificación (*detección y ubicación de eventos en un espacio geográfico*) de eventos viales en textos cortos en español, identificando entidades geográficas y sus relaciones espaciales, empleando representaciones conceptuales para la desambiguación y verificación de la coherencia, además de implementar una técnica para validar la calidad de la información proveniente de los textos, con la finalidad de proporcionar certidumbre en las ubicaciones identificadas.

1.5 Objetivos Particulares

- Diseño e implementación de una metodologìa semiautomática para geocodificar eventos viales recolectados de *Twitter*, empleando una red de ontologías y las relaciones espaciales contenidas en el mensaje.
- Diseño e implementación de una ontología de relaciones espaciales, que contenga las operaciones geográficas correspondientes a cada una.
- Diseño e implementación de una ontología que represente la infraestructura vial y los Puntos de Interés o POIs de la Ciudad de México.
- Diseño e implementación de una red de ontologías, conformada por la ontología de relaciones espaciales, la de infraestructura vial y POIs y por último la ontología de Calles de la Ciudad de México generada en el trabajo de Rivera et al. (2015).
- Generación manual de cuatro diccionarios para la estandarización de los *tweets* recolectados: “Abreviaturas”,” Acrónimos”,”*Hashtags*” y ”*Nicknames*”.
- Generación manual de cuatro diccionarios para la clasificacíon de los *tweets* recolectados: “Manifestaciones”,”Accidente vehicular”,”Congestión vehicular ” y ”Obras públicas”.

- Generación de un diccionario que contiene las Expresiones Regulares que representan, las Relaciones Espaciales más usadas en los textos.
- Diseño e implementación de un *script* para validar la confiabilidad de los usuarios que publican los *tweets*.
- Evaluación de los resultados obtenidos, empleando una geocodificación realizada por un usuario y otro tipo de métricas.
- Visualizar en un mapa los eventos y las vialidades afectadas por los mismos.

1.6 Alcances y Limitaciones

La metodología será capaz de geocodificar los eventos encontrados en los *tweets*, representando cada evento con un punto, línea, o ambos según sea el caso y colocando marcadores distintos dependiendo del tipo de evento.

Como limitantes del trabajo se encuentran las siguientes:

- La geocodificación está restringida a la cobertura del callejero que es utilizado, si el *tweet* menciona lugares que no se encuentren en la red vial, no es posible ubicarlo.
- Es necesario que el *tweet* describa al menos dos objetos geográficos para poder ubicar el evento. Si solo describe un objeto, debe ser un punto para poder ubicarlo.
- Si el *tweet* contiene acrónimos o abreviaturas que no se encuentren en los diccionarios definidos, es probable que alteren la geocodificación del evento.
- La precisión cartográfica de los eventos geocodificados, depende del número de elementos geográficos que se presentan en el *tweet* y de la cartografía empleada.

1.7 Aportaciones científicas

- Se propone una ontología de relaciones espaciales basada en el trabajo de Yecheng,(2011).
- Se genera un corpus con 368,933 *tweets* referentes a eventos viales

en la Ciudad de México,en el idioma español.

- Se propone un método para la validación de la confiabilidad de los *tweets* mediante la información del usuario que lo publica.
- Se propone un algoritmo de geocodificación semántica para el idioma español, empleando una red de ontologías para validar y desambiguar la información.
- La generación de un diccionario en español que describe las relaciones espaciales más usadas y las expresiones regulares que las representan.
- La generación de cuatro diccionarios: “Abreviaturas”,” Acrónimos”, ”*Hashtags*” y ”*Nicknames*”, que representan las expresiones que hacen referencia a los elementos involucrados en los eventos viales.
- La generación de cuatro diccionarios: “Manifestaciones”,”Accidente vehicular”,”Congestión vehicular ” y ”Obras públicas”, que representan las expresiones comúnmente empleadas para describir eventos viales.

1.8 Aportaciones tecnológicas

La aportación tecnológica más imporntante de este trabajo, es la conjunción de varias tecnologías para la implementación de la metodología propuesta, cada una de las tecnologías y para que fueron empleadas, serán enlistadas a continuación:

- Se emplean las técnicas de tratamiento de lenguaje natural para el procesamiento inicial de los textos cortos, y un script de términos frecuentes para obtener la lista que será empleada para generar los diccionarios.
- Para la clasificación se realizó un script basado en *N-Gram-Based Text Categorization* (Cavnat & Trenkie, 1994).
- Para la identificación de los lugares en los *tweets*, se emplea *Named Entity Recognition*.
- Son generadas las expresiones regulares de las relaciones espaciales.

- Se implementa un SPARQL Endpoint (triple store) para almacenar la red de ontologías y realizar las consultas a las mismas.
- Por último para la visualización de los mapas, es empleado OpenLayers y GeoServer como servidor geográfico.

1.9 Aportaciones académicas

- Publicación de un artículo JCR titulado "Geocoding Tweets Approach Based on Conceptual Representations in the Context of the Knowledge Society". Publicado en International Journal on Semantic Web and Information Systems (IJSWIS) Enero 2016.
- Publicación del artículo "Geocoding of Spatial Relationships Contained in Tweets". Publicado en International Journal of Knowledge Society Research (IJKSR) Marzo 2016.
- Publicación del artículo "Geocodificación de microblogs basada en ontologías". Publicado en 9ª Conferencia Ibérica de Sistemas y Tecnologías de Información, Barcelona, España Junio 2014.
- Estancia de investigación en la Universidad Federal de Minas Gerais, Belo Horizonte, Brasil. Bajo la supervisión del Dr. Clodoveu Davis Jr. Febrero a Agosto de 2015.

1.10 Organización del documento

Este documento está organizado como sigue. En el Capítulo 2 se realiza una breve descripción de los trabajos científicos relacionados a los tópicos presentados en este trabajo. Posteriormente en el Capítulo 3 se presenta la descripción de los conceptos y herramientas necesarias para entender de manera clara este trabajo. Por otro lado, en el Capítulo 4 se explica de forma detallada en que consisten cada una de las cuatro etapas de la metodología propuesta. Los resultados experimentales, así como la discusión de los mismos, son presentados en el Capítulo 5. Finalmente, en el Capítulo 6 se exponen las conclusiones y recomendaciones para la investigación futura.

2. Estado del Arte

En este capítulo se presentan trabajos relacionados con la detección de eventos en *Twitter*. Los trabajos se dividen en tres secciones. Primero menciona de describe el panorama general de obtener ubicaciones en *Twitter*, posteriormente se describen los trabajos relacionados a la detección de diferentes tipos de eventos y por último se describen las técnicas para la identificación de ubicaciones.

2.1 Microblogging

Ross et al. (2011) ha llevado a cabo un extenso estudio acerca de *microblogging* y *Twitter* proporcionando la siguiente definición de *microblogging*:

"*Microblogging* es una variante de un blog el cual permite al usuario publicar actualizaciones cortas, proporcionando un método de comunicación innovador que puede ser visto como el híbrido de un blog, mensajería instantánea, redes sociales y notificaciones de estado. El origen de la palabra sugiere que comparte la mayoría de los elementos con los blogs, por lo que potencialmente puede describirse mediante tres conceptos clave del *blogging* (Karger y Quan, 2005): los contenidos son publicaciones cortas, estas publicaciones se mantienen unidas por un autor de contenido y las entradas individuales al blog pueden ser fácilmente agregadas juntas ".

Las pequeñas actualizaciones publicadas en los *microbloggings* están limitadas en longitud, *Twitter* por ejemplo limita sus publicaciones a 140 caracteres (aproximadamente 25 palabras) por que originalmente los mensajes de los teléfonos móviles estaban limitados a ese tamaño (Weller, 2012).

2.2 Ubicación con base en redes sociales

En los últimos diez años se ha visto un enorme cambio de paradigma de cómo se crea, mantiene y usa la información en la web. Basados en los avances tecnológicos (como la gran inclusión del acceso a internet de banda ancha), coincidiendo con el nuevo software basado en la Web, lo que permite a los usuarios participar en la creación de contenidos para la Web, cambiando el papel del usuario de ser exclusivamente el consumidor al de un consumidor y productor, un término que se ha acuñado como *prosumer* Roick O. y Heuser S. (2013).

Desde entonces, los usuarios de la Web han producido una enorme cantidad de contenido, como fotos en *Flickr*, videos en *Youtube*, publicaciones en *Twitter* o descripciones de productos y clasificaciones. Esta evolución ha sido permitida gracias a la Web 2.0 (O'Reilly 2005).

Estos acontecimientos cambiaron aún más la forma en que la información geográfica es adquirida, mantenida y distribuida a través de la Web. La creciente disponibilidad de dispositivos GPS permitió a los usuarios aficionados sin formación de levantamiento de datos geográficos a capturar una gran cantidad de información geográfica. Impulsando proyectos de colaboración exitosos tales como *OpenStreetMap* o *WikiMap*. Y grandes empresas como *Google*, *Apple* o *Nokia* actualizan sus mapas base con la información obtenida de sus usuarios para obtener información del tráfico en tiempo real (Boyd y Ellison 2008).

Actualmente los sitios de redes sociales juegan un papel clave en esta evolución y por lo tanto ha atraído a millones de usuarios en todo el mundo. *Twitter* por ejemplo tiene un promedio diario de más de 400 millones de usuarios activos al mes y envía 500 millones de *tweets* al día[2].

Entonces a los Sitios de Redes Sociales que incluyen información de la ubicación en los contenidos compartidos se llaman Redes Sociales Basadas en Localización (*Location Based Social Network LBSN*). Estas redes muestran información geográfica en un mapa o en una lista de actualizaciones de estado

[2] http://www.trecebits.com/2016/06/20/Twitter-tiene-418-millones-de-usuarios-activos-al-mes-y-envia-700-millones-de-tuits-al-dia/ Consultada el 20 de junio del 2016

ordenado por la proximidad geográfica en comparación con el concepto inverso del orden cronológico tradicional. Este fenómeno también ha sido conocido como *Locative Mobile Social Networks* (Gordon y de Souza e Silva 2011).

En general existen dos caminos de cómo la información geográfica se puede compartir en *Location Based Social Networks* (Elwood et al. 2011). Primero, la anotación de la información de ubicación proveniente de los dispositivos digitales conocido como geo etiquetado (*geotagging*) (Turner 2006). Esto convierte a fotografías, videos o *tweets* de *Twitter* en información geográfica. Un ejemplo de ello , son los usuarios que permiten se agreguen coordenadas geográficas a sus *tweets*. Esta información geográfica puede ser usada para búsquedas locales o ubicar eventos en un mapa.

La segunda forma, es mostrar las actividades cotidianas junto con la ubicación actual . Por ejemplo los usuarios en *Foursquare* pueden hacer *check-in* en un lugar determinado y compartirlo con sus amigos. Esto hace referencia a la creación de redes geosociales. Las aplicaciones que forman parte de este tipo de redes sociales incentivan al usuario con juegos o características para mostrar constantemente su ubicación. La información tal como el número de *check-ins* en cierto lugar, el número de visitas individuales, gustos y consejos, pueden indicar la popularidad de un lugar en específico y pueden emplearse para mejorar la experiencia del usuario en la ciudad (Gordon y De Souza e Silva 2011).

Dentro de los LBSN los usuarios comparten información geo referenciada de diferentes maneras. Basándose en la motivación primaria de compartir la ubicación se puede distinguir entre un LBSN con fines sociales y un LBSN impulsado.

El LBSN con fines sociales muestra las actividades cotidianas y muchas veces es motivada para socializar con otros usuarios mostrando su ubicación en lugares divertidos para mantener su red social. En el caso de los LBSN impulsados se muestra la ubicación a un grupo grande de personas con una razón en específico, por ejemplo notificando de un evento (Sui y Goodchild 2011).

2.2.1 Tweets georreferenciados

Desde Agosto del 2009, *Twitter* permitió incluir metadatos geográficos a los *tweets* indicando la ubicación donde fue enviado (*Twitter*, 2009). Existen dos tipos de *tweets* georreferenciados disponibles: *Lugar*, que permite a un usuario especificar manualmente una ciudad o vecindario usando un menú disponible en el software y *Ubicación Exacta*, que es un conjunto de coordenadas generalmente proporcionadas a través del GPS del dispositivo o por una triangulación celular (Twitter, 2013b).

La ubicación de los lugares debe ser manualmente seleccionada por el usuario desde una lista predefinida de ubicaciones soportada por Twitter. Es principalmente usada cuando se está "twitteando" desde una computadora de escritorio o un dispositivo fijo. Esta ubicación debe ser actualizada manualmente por el usuario, por lo que los *tweets*, de los usuarios que viajan a otro país, reflejan su última ubicación seleccionada (*Twitter*, 2013c).

En contraste, la ubicación exacta utiliza funciones de geolocalización en un dispositivo móvil para proporcionar la ubicación geográfica del usuario en el momento de enviar un *tweet*, es decir, el usuario no tiene que realizar ninguna acción para actualizar su ubicación mientras viaja; lo que significa que puede capturar una dirección precisa, como una casa o cafetería favorita (*Twitter*, 2013c).

Pero debido a los riesgos de privacidad, las funciones de geolocalización de Twitter están desactivadas y los usuarios deben habilitarlas de forma explícita en su cuenta. En un día común, sólo el 2.02% de todos los *tweets* a nivel mundial incluyen metadatos geográficos. De esta porción, el 1.8% tiene un indicador de *Lugar*, el 1.6% tiene el dato *Ubicación Exacta* y, el 1.4% tiene ambos datos.

2.3 Detección de Eventos en *Twitter*

Gutierrez et al. (2015) y Oussalah et al. (2013) establecen que el uso de la información contenida en los *tweets*, proporciona una cantidad importante de información geográfica, por que los textos comúnmente hacen referencia a lugares. El análisis de estos los *tweets* permite conocer y evaluar eventos sociales y naturales.

El tópico de la detección de eventos en *Twitter*, ha generado mucho interés en el ámbito de investigación, enfocándose en analizar eventos de todo tipo (Agarwal et al., 2102; Atefeh & Khreich, 2015).

Por ejemplo en el trabajo de Nielsen et al. (2015) se realiza un análisis detallado de la información proveniente del flujo de *Twitter* para realizar monitoreo de salud, en particular enfoca sus esfuerzos en el análisis del comportamiento de la influenza. Proponiendo un enfoque basado en *Machine Learning*, para la obtención de las ubicaciones.

Este trabajo se inspira en los métodos propuestos en Culotta (2013), que emplea un análisis de regresión múltiple con el número de *tweets* para detectar brotes de influenza y el trabajo de Broniatowski et al,. (2013), en donde propone un sistema de geolocalización para identificar donde ocurren los casos de influenza en los *tweets*, auxiliándose de los datos del perfil de usuario y de herramientas de geocodificación , los autores señalan que su sistema denominado “Carmen”, es preciso y abarca muchos lugares demostrando ser una herramienta muy útil para la mejora en la vigilancia de los casos de influenza.

Por otro lado, el trabajo de Robinson et al,. (2013) se enfoca en la detección de emergencias, principalmente en sismos, ya que propone el desarrollo de un detector sísmico para Australia y Nueva Zelanda usando *Twitter*. Este sistema monitorea los *tweets* relacionados con el sismo y determina su ubicación cuando el *tweet* se ha identificado como geográficamente cercano y el porcentaje de *retweets* es bajo una notificación por correo electrónico es generada.

Los autores mencionan que su algoritmo ha detectado 20 notificaciones desde Diciembre del 2012 identificando 17 hechos reales con solo 3 falsos positivos, también mencionan que no necesitan muchos *tweets* para generar sus alertas, solo identificar la cercanía al evento.

Otro enfoque altamente estudiado, es la detección de eventos viales. El trabajo propuesto por Gutierrez et al. (2015), propone un método para obtener información del tránsito, empleando un enfoque basado en Máquinas de Soporte Vectorial (SVM) y un algoritmos de clusterización para seguir la evolución de los eventos. Por otro lado Albuquerque et al. (2015) presenta una ontología de dominio referente a eventos viales, llamada TEDO y describe una herramienta para identificar la ubicación de los eventos en los *tweets* usando los metadatos del mismo.

Finalmente, se tienen los trabajos que se enfocan en recuperar y desambiguar los lugares mencionados en los *tweets*, tal es el caso de Lee et al. (2015), que describe un método basado en Aprendizaje Máquina para extraer y desambiguar las ubicaciones en el mensaje. Otro enfoque es el de Roller et al. (2012) que combina técnicas de *Named Entity Recognition* y Wikipedia sobre un corpus que *twitter* y obtiene una lista con los términos que representan ubicaciones.

Por otro lado, en (Graham et al., 2014), se combinan las técnicas de *Named Entity Recognition* con un *gazetteers* para asignar etiquetas conforme se reconocen los nombres en el texto. Zhang y Gelernter (2014) describen un modelo para determinar cual ubicación presenta una correcta coincidencia con la ubicación encontrada en el *tweet*, usando los metadatos y la información que extraen de la ciudad, estado y país, obteniendo una medida F de 85.22. Los autores sugieren que los errores que presentan, se deben principalmente a los errores ortográficos o la falta de información.

2.3.1 Geocodificación en *Twitter*

La poca disponibilidad de *tweets* georreferenciados, sugiere que empleando algoritmos de geocodificación se podría ampliar considerablemente el universo de *tweets* mapeables, mediante la extracción de información geográfica de la información textual de los *tweets*. Esto plantea preguntas sobre cómo obtener la mayor cantidad de información geográfica del texto, así como de qué manera tener la más alta precisión posible, además, cómo se evaluará la precisión y qué algoritmos de geocodificación lograrían los mejores resultados con el texto limitado de los *tweets*.

El texto en los *tweets* puede mencionar una o más ubicaciones (por ejemplo "Delegación Iztacalco"), este texto requiere un algoritmo de geocodificación para identificar la ubicación contenida en el texto, para desambiguar y convertir la información en coordenadas geográficas. Existen dos tipos principales de algoritmos de geocodificación:

- Geocodificación tradicional, donde todo el texto de entrada ya es una ubicación.
- Geocodificación de texto completo, donde se debe analizar de forma más detallada el texto para identificar los lugares mencionados Leetaru (2013).

El campo *Lugar* sólo contiene el nombre de la ciudad o alguna otra ubicación geográfica y representa una tarea de geocodificación tradicional. Pero un usuario puede mencionar su ubicación geográfica en el texto, por ejemplo "Yo soy un estudiante de Ciencias de la Computación viviendo en la Ciudad de México y realmente me encanta este lugar", este tipo de texto requiere otra forma de geocodificación, debido a que se tiene que analizar todo el texto para extraer las palabras que hagan referencia a un lugar geográfico y de esta forma asignarle una coordenada geográfica que lo represente.

Casi un tercio de todos los lugares en la Tierra comparten su nombre con otro lugar en el planeta, lo que significa que una referencia a “Cuauhtémoc” debe ser desambiguada por un sistema de geocodificación para determinar cuál de los 5 resultados, incluyendo calles, colonias y delegaciones con ese nombre es al que se está refiriendo. Este proceso de desambiguación es altamente dependiente del contexto y de estimaciones basadas en patrones geográficos de referencia, por lo tanto es propenso a errores. Uno de los mayores retos con la evaluación de la precisión en los sistemas de geocodificación es la falta de datos de referencia o *gold standard* contra los cuales se puedan comparar los resultados obtenidos. Por ejemplo, cuando se aplica la geocodificación de texto completo a una colección de documentos, se deben comprobar los fallos del sistema al asignar las referencias geográficas. Lo anterior es ocasionado porque se pueden confundir nombres de lugares con nombres propios y, sin una adecuada desambiguación, las ubicaciones no serían identificadas. Esta tarea es comúnmente realizada de forma manual ocupando un subconjunto aleatorio de documentos, pero es lenta y propensa a errores, por tal motivo se requieren pequeñas muestras de prueba para ser procesadas (Fischer, 2011).

Cada resultado debe ser evaluado en dos dimensiones: exactitud (el porcentaje de *tweets* recuperados a los cuales se les asignó información geográfica) y, precisión (de los recuperados con información geográfica, cuales están asignados correctamente). Un factor importante en la evaluación de la precisión, es que cuando se asigna una coordenada geográfica a un lugar identificado en el texto, se coloca como información geográfica la correspondiente al centro de la ciudad, por ejemplo al reconocer la palabra “Ciudad de México” como un lugar, la coordenada asociada a esta ubicación será la del zócalo capitalino, aunque realmente se haga referencia a un lugar alejado de este punto. De esta manera, el objetivo es encontrar una combinación de buena información y algoritmo que generen la más alta precisión y exactitud.

2.4 Técnicas para la identificación de ubicaciones

Existen varios enfoques para recuperar, o en este caso, identificar, ubicaciones. Entonces surge la cuestión ¿Cómo se identifican las palabras que representan ubicaciones?: el primer enfoque es de acuerdo a su sintaxis (NER),el segundo es por cantidad de términos, el tercero es por los objetos o personas asociados con una ubicación, el cuarto sería por coincidencia exacta de las palabras en un diccionario geográfico, el quinto consiste en la inferencia ocupando una enciclopedia de referencia, el sexto es por coincidencia probabilística entre las siglas o acrónimo de ubicación y por último el séptimo sería empleando una palabra o frase que sirva para eliminar la ambigüedad. A continuación se explica de forma breve cada enfoque.

2.4.1 Geolocalización basada en la sintanxis (NER)

La identificación de ubicaciones es un sub problema de identificación de todos los nombres de las entidades y así extraer su ubicación, usualmente tratado en el contexto de *Named Entity Recognition* (NER). Los nombres propios que representan lugares pueden ser extendidos a los idiomas, eventos o puntos de referencia asociados con los lugares tal como el "francés" o “La torre Eiffel” para Francia Vasardani, M., et al. (2013) y puede ser identificado mediante la combinación de un clasificador del vecino K más cercano con un clasificador lineal de Campos Aleatorios Condicionales para encontrar el nombre de las entidades.

En el método de Liu et al. (2012) se ha logrado una medida F1 de 78.5% para la ubicación de entidades en *tweets*. Aunque usualmente se utiliza la precisión y la exactitud para evaluar los resultados del *NER* , algunos sistemas han demostrado que el rendimiento de estas herramientas en micro texto es menor que en texto de mayor dimensión y algoritmos como Latent Dirichlet ha demostrado arrojar resultados bastante buenos para micro texto Ritter, A., et al. (2012).

Un método especializado para Twitter es el de Galernter y Mushegian (2011), que empleando NER extraian información de las ubicaciones tales como el nombre del país, el nombre de la ciudad o nombres de calles y edificios y todo obtenido de la ubicación referenciada en un tweet. Éste método ubica solo el 34% de los tweets correctamente lo cual comentan los autores es debido a las faltas de ortografía y las abreviaturas usadas comunmente en éstos textos. Una segunda versión de este trabajo, presentado en Zhang & Galernter (2014) emplea un método supervisado de aprendizaje máquina, para ponderar las características de los gazetters y seleccionar el adecuado para geocodificar el mensaje obtenido de *Twitter*.

2.4.2 Geolocalización basada en modelos de lenguaje

Kinsella et al. (2011) se basa en el enfoque de modelado del lenguaje de Ponte y Croft (1998) para crear una función que describe la distribución probabilística. El grupo de Kinsella estimó la distribución de términos asociados con una ubicación y entonces estimó la probabilidad de los *tweets* asociados a una ubicación. Su modelo de lenguaje fue exitoso en la pruebas a nivel de ciudad con una precisión del 65%, pero regresando solamente una precisión del 24% a nivel de vecindario. Por otro lado Eisenstein et al. (2010) construye un modelo para predecir la región del autor del *tweet* de acuerdo a la elección del vocabulario y la jerga seleccionada. Su modelo puede identificar la ubicación correcta del autor en un 24% de los casos. Cheng et al. (2010) por otro lado usa un modelo de lenguaje que identifica la ubicación del autor de un *tweet* con un error de 100 millas entre la ubicación encontrada y la ubicación real del autor con una precisión del 51% de los autores probados.

2.4.3 Geolocalización basada en coincidencias con diccionarios geográficos

Liberman et al. (2011) proporciona un estudio de los métodos de geo localización basada en texto aunque existen método específicos que se han utilizado para *Twitter*. Paradesi (2011) combina *Name Entity Recognition* y métodos basados en diccionarios geográficos en su “TwitterTagger”.

El sistema primero asigna etiquetas del *part-of-speech* para encontrar nombres propios y entonces comparar con frases nominales por *tweet* en el diccionario geográfico del *United States Geological Survey* para identificar los lugares. El sistema identifica los nombres que parecían ser lugares mediante la búsqueda de un indicador espacial, como una preposición encontrada antes del nombre de la ubicación. La investigación del "TwitterTagger" no considera que clase de lugares encuentra en los *tweets* y tampoco considera abreviaturas ni acrónimos. En algunos casos se puede extraer la ubicación del texto y hacer una búsqueda directa en un diccionario geográfico tal como Geonames y obtener las coordenadas geográficas asociadas a esa ubicación Bouillot et al, (2012).

Una variación de esta técnica es empleado en Moncla et al, (2014), en donde se propone un enfoque híbrido que combina el uso de geo-etiquetas y una búsqueda en diferentes diccionarios geográficos con un algoritmo de closterización para identificar los topónimos que representan ubicaciones locales, que no son facilmente identificadas por su nivel de detalle, obteniendo coordenadas geográficas de paradas turísticas, pequeños poblados, puentes, etc.

2.4.4 Geolocalización por la asociación con geo-etiquetas

Watanabe et al. (2011) identificó lugares locales al nivel de edificio, generando su propio diccionario geográfico de los lugares y sus coordenadas geográficas mediante la extracción de los nombres de los lugares de las geo-etiquetas de los *tweets* japoneses. Ellos usaron la información de los *tweets* geo-etiquetados para identificar los lugares mencionados en los *tweets* que no tenían etiquetas geográficas, y agrupan los *tweets* de acuerdo a las palabras clave temáticas comunes que generaron dentro de un corto período de tiempo y en un área geográfica limitada.

Su sistema detecta eventos locales con una precisión del 25.5%. Jung (2011) propuso que la ubicación de un *tweet* podría deducirse mediante la fusión de las conversaciones de *Twitter* entre las personas en un solo documento y usando las asociaciones entre los *tweets* individuales para mejorar el reconocimiento de la ubicación y otras entidades.

Otro método fue el desarrollado por Davis JR. et al,.(2011) que infería la ubicación de un usuario de *Twitter* basándose en los lugares conocidos de los seguidores del usuario, seleccionando el más popular entre las localidades de amigos.

2.4.5 Geolocalización mediante la asociación de coordenadas geográficas

Algunos sistemas de detección basados en eventos que utiliza *Twitter* pueden depender de la georreferenciaciación individual de un *tweet*, como el sistema *Mapster* (2011) y *TwitInfo* (2011). El problema es que esta característica de *Twitter* es voluntaria y pocas personas la usan actualmente, solo una pequeña fracción de los *tweets* incluyen latitud y longitud.

2.4.6 Geolocalización basada en acrónimos y abreviaturas

La geo localización de texto basada solo en abreviaturas o siglas que describen a la ubicación implica identificar primero las abreviaturas y los acrónimos y luego desambiguarlos. De los primeros artículos en tratar ésta problemática fue el de Park y Bird (2001) que consideraban la combinación de encontrar y eliminar la ambigüedad de las abreviaturas, aunque la identificación y desambiguación de siglas y acrónimos son temas de investigación que comúnmente se manejan por separado.

2.5 Calidad de lo datos

Los datos geográficos recopilados por voluntarios sin formación y sin experiencia siempre deben estar sujetos a un estricto control de calidad antes de emplear estos datos para proyectos de investigación y aplicaciones. Por lo tanto, la evaluación de la calidad de los datos es uno de los problemas de investigación más importante.

Con el fin de describir la calidad de los datos geoespaciales, varios elementos de calidad se han definido: Linaje de la información, exactitud posicional y de atributos, consistencia lógica, la integridad y la calidad temporal Van Ort (2006).

En términos de Información Geográfica Voluntaria (VGI por sus siglas en inglés) otros autores sugieren ampliar estos elementos y hacer énfasis en que la experiencia del usuario contribuye (calidad del usuario) y el linaje de las respectivas características (calidad de las características), así como la interrelación entre ambos aspectos Van Exel et al,. (2010).

Las cuestiones más importantes son la exactitud de la ubicación y las características, estos datos son comparados con conjuntos de datos gubernamentales y comerciales tradicionales o con datos provenientes de servicios colaborativos como es el caso de *OpenStreetMaps*. Al igual que los estudios anteriores sobre la calidad de los datos VGI , las áreas altamente pobladas revelan una mejor calidad general de los datos que se relaciona con una mayor actividad en esa área [Neis et al,. (2012), Haklay (2010), Girres y Touya (2010)].

Por lo tanto, las investigaciones sobre la correlación de las actividades del usuario y la calidad del dato pueden proporcionar ideas que se pueden emplear para la estimación de la calidad en áreas desconocidas. Por otra parte los estudios sobre la calidad temporal de los datos pueden proporcionar ideas de cómo el crecimiento de la comunidad de usuarios influye en la calidad de los datos a través del tiempo.

Elwood et al,. (2013) propone nuevas aproximaciones para asegurar la calidad de los datos como un objetivo de las nuevas investigaciones. Los autores sugieren un sistema de revisión con usuarios de confianza que actúan como moderadores que revisan las aportaciones. Una segunda estrategia es emplear un sistema de reglas sintáctico que permita validar por verificación cruzada las aportaciones de los usuarios. Sin embargo, se requiere investigación sobre la manera de formular y organizar un sistema de esta índole y que tipo de análisis es requerido para aplicar al sistema.

2.5.1 Representatividad de los datos

Otro tema crítico de la información generada por los usuarios es la representatividad de los datos. Puede ponerse en duda si los datos compartidos en las redes sociales pueden considerarse representativos de toda la sociedad ya que el grupo de usuarios es reducido a los usuarios digitales, por ejemplo, los jóvenes que crecieron con acceso a internet constante y utilizan la Web constantemente a lo largo de su vida Boyd, D., y Crawford, K. (2012).

Esto incluye también el problema de la brecha digital: mayoría de la actividad en las redes sociales en línea se concentra a América del Norte y Europa, mientras que en otras partes del mundo el acceso a los dispositivos digitales y el Internet está reservado a unas pocas personas. Estos hechos deben ser considerados cuando se trata de datos de las redes sociales. Necesitamos más investigación sobre estos aspectos, sobre todo en las limitaciones espaciales y sociales de la actividad del usuario y la forma en que la creciente penetración de los teléfonos móviles en los países en desarrollo influye en esta evolución Roick, O., y Heuser, S. (2013).

2.6 Discusión del estado del arte

Más de cuatrocientos millones de usuarios publican cientos de millones de mensajes cortos cada día en *Twitter* . Como resultado, este contenido ha sido utilizado por los investigadores de campos tan diversos como la epidemiología, la política, el marketing y la geografía para entender mejor, la parte espacial y medir las tendencias y los patrones sociales, económicos y políticos a gran escala . Sin embargo , gran parte de este análisis se lleva a cabo con sólo un entendimiento limitado de la mejor forma de trabajar con los contextos espaciales y lingüísticos en los que se produce la información . Como tal , ha sido necesario estudiar la fiabilidad del origen del contenido en *Twitter.*

En la investigación del estado del arte se encontró que existen retos significativos para la determinación precisa de la ubicación inmersa en los *tweets* de forma automatizada. Ninguno de los métodos descritos en ésta sección es capaz de igualar la precisión de la codificación humana.

El estilo informal de escritura, la longitud corta de los *tweets*, el uso de varios idiomas en un mismo *tweet* y la presencia de contenidos específicos no verbales tal como URL y emoticonos complican la identificación de la lengua y limitan la precisión.

Además se enfatiza la relevancia de verificar la validez y representatividad de la información obtenida de los *tweets* y de los diferentes métodos empleados para corroborar el resultado obtenido, recalcando el uso de un *gold standard* como por ejemplo la comparación con un ser humano.

3. Marco Teórico

En este capítulo, se describen algunos de los términos y herramientas que son utilizados durante el desarrollo de la metodología. Primeramente, se describen de manera general algunos conceptos importantes relacionados con *Twitter* y las relaciones espaciales. Posteriormente, se presenta una descripción general de ontologías, las metodologías de creación sus los estándares, el estándar *RDF*, *SPARQL* y Finalmente, una descripción de *GeoSPARQL*.

3.1 Twitter

Twitter es actualmente el más popular y de rápido crecimiento servicio de "*microblogging*". Encontrándose en el octavo lugar entre los sitios más populares a nivel mundial, de acuerdo a ranking "*3-month Alexa traffic*"[3].

Twitter permite mostrar cualquier tipo de información, a través de opiniones e ideas que mantienen a las personas actualizando o informando los eventos cotidianos. Los usuarios pueden publicar textos cortos, de máximo 140 caracteres de longitud conocidos como "*tweets*".

Estos "*tweets*", son publicados automáticamente de manera continua y accesible al público, en el perfil de usuario en *Twitter* e instantáneamente es enviado a la red de seguidores del usuario. Estos mensajes, pueden ser publicados a través de varios servicios de comunicación tales como, teléfonos celulares, emails, interfaces *Web* o alguna aplicación de terceros.

Twitter ha evolucionado a través del tiempo y adoptado sugerencias propuestas originalmente por usuarios, para hacer la plataforma más flexible. Esto actualmente proporciona diferentes formas de conversar e interactuar, haciendo referencia a otros usuarios en los mensajes enviados, mediante el uso de un vocabulario específico bien definido.

[3] http://www.alexa.com/siteinfo/twitter.com

Colocando el símbolo "@" antes del nombre de usuario, se crea una mención o una respuesta que hace referencia a una cuenta de usuario. Una mención es usada en cualquier lugar del mensaje, para señalar que el usuario mencionado también se encuentra registrado en *Twitter*. Una respuesta, es una mención especial desde un usuario en respuesta al mensaje de otro usuario, comenzando con "*replied-to @username*" (Honeycutt and Herring, 2009).

Twitter también permite a los usuarios reenviar o hacer un "*retweet*" a alguno de sus seguidores. Esto es comúnmente indicado en el mensaje usando el prefijo "*RT*" antes del nombre de usuario que publicó el mensaje originalmente, "*RT @username*". *Retweetear* es una práctica común, para mostrar la utilidad o el interés de la información mientras proporciona credibilidad al usuario original (Boyd et al. 2010).

Los temas en *Twitter*, pueden ser categorizados mediante un *hashtag*, el cual es una palabra representativa del tema y es precedida por un signo "#" ,conocido como almohadilla o numeral, por ejemplo "#TránsitoLento". Los *Hashtags*, fueron desarrollados para generar agrupaciones de *tweets* dependiendo de su temática, y permite a los usuarios realizar filtros por tema y seguir solamente los mensajes de los tópicos de su interés en tiempo real.

En el estudio de la información generada por esta plataforma, es importante conocer las definiciones técnicas dadas por *Twitter*[4] de los términos empleados cotidianamente en los *tweets*.

Twitter: Es una de las plataformas de *microbologging* más populares a nivel mundial, su funcionamiento principal es permitir a los usuarios publicar y leer mensajes cortos, similar a los *blogs*.

tweet: Textos cortos, de máximo 140 caracteres de longitud, publicados por los usuarios pertenecientes a *Twitter*.

[4] https://support.twitter.com/articles/352810

Mensajes Directos (DM): Son los *tweet* que puedes enviar a otros usuarios de manera personal y privada, que quiere decir esto, que estos mensajes solo los podrán leer el destinatario y el remitente y nadie más, no aparece el *timeline* o línea de tiempo pública donde aparecen todos los tweet.

***Retweet* (RT):** Si al usuario le gusta lo que publica **@xxxx** y cree que vale la pena compartirlo con sus seguidores selecciona la opción *retweet* y lo publica. Si quiere agregarle un comentario también los puede hacer, cuidando que no sobre pase los 140 caracteres. El **RT** (símbolo como popularmente se conoce a esta opción) sirve también para medir el impacto que tienen las publicaciones de un usuario en sus seguidores y el valor que estos le dan. Mientras más RT consiga, más popularidad y respeto logrará tener en el portal.

***Timeline* (línea de tiempo):** Es la columna de la izquierda la de mayor tamaño, donde aparecen todas las publicaciones que hacen tus seguidos y tus seguidores de manera pública, así como tus actualizaciones.

Seguir (*Follow*): Es la acción por la cual puedes agregar a alguien a tu cuenta de *Twitter* y poder tener acceso a sus publicaciones (siempre y cuando sean públicas). Si la otra persona decidí también seguirte podrás compartir con ellos, además, la opción de *retweet* y de mensaje directo (más adelante más información).

Seguiendo (*Following*): Son las personas a las cuales sigues en *Twitter*, que quiere decir esto, son personajes que puedes agregar a tu cuenta con la finalidad de poder acceder a sus *tweet* (siempre que sean públicos). Tú elijes, puedes seguirlos porque son conocidos (famosos) o porque te gusta lo que publican. Sin embargo, es bueno aclarar que al agregarlos a tu cuenta no necesariamente quiere decir que ellos también te seguirán.

Seguidores (*Followers*): Son personas que han decidido seguirte y te han agregado a su cuenta de *Twitter*, con ellos podrás compartir tanto los *tweet* públicos como tener la facilidad de hacer *retweet* a cualquiera de sus publicaciones y comunicarte con ellos vía mensaje directo (más adelante le daré más detalles).

@Mención ó @Responder (@*Replay*): es una manera para comunicarte con los otros usuarios o de personalizar el mensaje; es decir, es una manera de dirigirte de manera personal a otros usuarios, pero de una manera pública.

***Hashtag*:** Son como etiquetas, palabras acompañadas del símbolo "#", que siempre va delante, y que tienen la función de agrupar los tweet que hablan sobre un determinado tema, con el propósito de hacer más fácil su ubicación al momento de realizar una búsqueda.

3.2 Relaciones espaciales

Las descripciones espaciales, detallan en donde las cosas están ubicadas en el espacio. En general se componen de tres elementos diferentes: El "*locatum*", el "*relatum*" y su relación espacial codificada como preposiciones espaciales, definiendo una región en la cual el "*locatum*" es ubicado (Landau and Jackendoff, 1993).

Una representación sintáctica comúnmente usada para estas descripciones es:

$$\big[[verbo\ principal]\ preposición\ espacial\big]FN$$

Los corchetes indican que los elementos en la descripción son opcionales. La Frase Nominal FN, representa el "*relatum*", que puede ser un simple sustantivo ("estación"), un sustantivo compuesto ("Estación Balderas"), o una frase compleja que contenga varias frases nominales y relaciones ("Yo estoy cerca de la Estación Balderas"). En esta estructura, se tiene el sujeto, "Yo", que representa el "*locatum*" y se describe la relación a la Frase Nominal "Estación" con la preposición "cerca".

Las descripciones espaciales emplean adverbios como, delante, detrás, arriba, allí, entre otros, especificando el "*relatum*" solo implícitamente y es requerido el conocimiento del contexto para una correcta interpretación. También se incorporan relaciones espaciales a los verbos como por ejemplo, "hasta cruzar", que pueden ser interpretados como dos estructuras separadas.

3.2.1 Relaciones espaciales en los Sistemas de Información Geográfica

En los Sistemas de Información Geográfica (SIG), definen a las relaciones espaciales usualmente, como aquellas que describen las relaciones relativas a los objetos en el espacio y se encuentran contenidas en las coordenadas (Laurini, 2012). Estas relaciones pueden ser divididas en topológicas, de dirección, de distancia y difusas (Yecheng, 2011).

3.2.1.1 Relaciones topológicas

La representación de las relaciones topológicas es uno de los problemas básicos en SIG. RCC (*Region Connection Calculus*) (Randell et al., 1992) y el modelo 9-intersección (Egenhofer and Herring, 1990) son modelos ampliamente aceptados basados en la teoría de la topología del conjunto de puntos. En estos modelos, cada entidad espacial es tomada como una abstracción de un punto, línea o región. Sin embargo, no es fácil distinguir expresiones de relaciones espaciales en lenguaje natural, lo cual es afectado por el contexto y el conocimiento de las personas.

En el trabajo de (Yecheng, 2011) se definen algunas de las relaciones topológicas mayormente usadas, (ver Tabla 1).

Tabla 1. Relaciones topológicas, definidas por (Yecheng, 2011).

Tipo de Relación	Tipo de Entidad geográfica
Cross	Línea, Línea
	Línea, Región
Contains	Región, Región
Intersect	Línea, Línea
Touch	Región, Línea
	Región, Región

3.2.1.2 Relaciones de dirección

Las relaciones de dirección, necesitan un marco de referencia para asignar una dirección. Si el marco de referencia cambia, también la dirección cambia. Existen dos tipos de marco de referencia:

- Marco de referencia relativo; Está basado en un objeto de tipo punto, que es la referencia y divide el espacio en cuatro regiones de dirección, enfrente, atrás, izquierda y derecha.
- Marco de referencia absoluto; Usualmente se refiere a un sistema de coordenadas geográficas definido por la ubicación de los polos. Norte, Sur, Este y Oeste son las cuatro direcciones cardinales de este marco.

En los SIG, la orientación relativa de los objetos geográficos son típicamente descritos por direcciones de cardinalidad, por lo que este enfoque es usualmente más usado.

3.2.1.3 Relaciones de distancia

Las relaciones de distancia, pueden expresarse de forma cualitativa mediante términos en lenguaje natural (como “cerca” y “lejos”) y cuantitativamente mediante unidades de distancia (metros, millas, kilómetros etc.) e incluso por unidades de costo- tiempo (litros de gasolina, costo de peaje, tiempo de traslado, etc.).

3.2.1.4 Relaciones difusas

Las relaciones difusas, son relaciones que describen inclusión, por ejemplo *está en*. También pueden describir otro tipo de combinaciones entre las entidades geográficas tales como *al lado de* o *a continuación*. Al igual que en las relaciones descritas anteriormente, estas relaciones pueden establecerse entre elementos con un mismo tipo de información, o bien entre tipos distintos.

3.3 Ontologías

Existen varias definiciones de lo que es una ontología, nosotros presentamos las dos definiciones que consideramos son complementarias:

1. "Una ontología define los términos y relaciones básicas que componen el vocabulario de un área temática, así como las reglas para combinar términos y relaciones para definir la extensión del vocabulario".(Neches, 1991).
2. “Una ontología es una especificación explicita de una conceptualización”. (Gruber, 1995).

Para desarrollar una ontología, es necesario hacerlo bajo un dominio de discurso y un marco teórico, que definirá un vocabulario del cual se derivan los conceptos, relaciones, instancias, constantes, atributos, axiomas, y reglas.

Estos conceptos se describen a continuación:

- **Conceptos**.- Son objetos o entidades, considerados desde un punto de vista amplio. Los conceptos de una ontología están normalmente organizados en taxonomías en las cuales se pueden aplicar mecanismos de herencia.
- **Relaciones**.- Las relaciones representan un tipo de asociación entre conceptos del dominio. Si la relación une dos conceptos se denomina relación binaria. Una relación binaria relevante es Subclase-de, que se utiliza para construir taxonomías de clase, como se ha especificado anteriormente.
- **Instancia**.- Se utilizan para representar individuos en la ontología. Las relaciones también se pueden instanciar.
- **Constantes**.- Son valores numéricos que no cambian en un largo período de tiempo.

- **Atributos**.- Los atributos describen propiedades, se pueden distinguir dos tipos de atributos de instancia y de clase.
    - Los atributos de instancia describen propiedades de las instancias de los conceptos, en las cuales toman su valor, éstos se definen en un concepto y se heredan a sus subconceptos e instancias.
    - Los atributos de clase describen conceptos y toman su valor en el concepto en el cual se definen. Estos atributos no se heredan ni a los subconceptos ni a las instancias.
- **Axiomas**.- Los axiomas son expresiones lógicas siempre verdaderas que suelen utilizarse para definir restricciones en la ontología.
- **Reglas**.- Las reglas se utilizan normalmente para inferir conocimientos en la ontología, tales como valores de atributos, instancias de relaciones, etc. [(Gruber, 1995), (Corcho et al., 2005)].

Para el diseño de una ontología existen varios criterios que se deben tener en cuenta:

- **Claridad y objetividad**.- La ontología debe proporcionar el significado de términos definidos proporcionando definiciones objetivas y documentadas en el lenguaje natural.
- **Completitud**.- Una definición expresada por una condición necesaria y suficiente es preferida por una definición parcial.
- **Coherencia**.- Permite que se puedan realizar inferencias y que éstas sean consistentes con las definiciones ya preestablecidas.
- **Maximiza la extensibilidad monotónica**.- Los términos generales nuevos o especializados deben incluirse en la ontología de tal forma que no requiera revisión de definiciones existentes.
- **Mínimo compromiso ontológico**.- Hace pocas afirmaciones acerca del mundo a ser modelado, lo cual significa que la ontología debe ser especifica tanto como sea posible el significado de sus términos , dando libertad a la ontología para especializar e instanciar.

- **Principio de distinción ontológica**.- Las clases de una ontología deben ser disjuntas. El criterio utilizado para aislar las propiedades principales consideradas a ser invariantes para una instancia de una clase se llama criterio de identidad.
- **Diversificación de jerarquías**.- Si el conocimiento es suficiente es representado en la ontología y con muchas formas o criterios de clasificación, para que facilite el introducir nuevos conceptos y heredad propiedades de diferentes puntos de vista.
- **Modularidad**.- Minimiza el acoplamiento entre módulos.
- **Minimizar la distancia semántica entre conceptos hermanos**.- Los conceptos similares son agrupados y representados como subclases de 1 clase y deben definirse utilizando las mismas primitivas, mientras que los conceptos menos similares son separados en la jerarquía.
- **Estandarización de nombres.-** para evitar inconsistencias en la ontología, así como la confusión al momento de realizar inferencias en la misma (Buriano et al., 2006).

Las ontologías también se categorizan dependiendo del grado de formalidad que tengan.

En Alta informalidad.- Son las ontologías que están expresadas en lenguajes natural, un ejemplo de ello son los glosarios.

Semi-informalidad.- Estas ontologías estas estructuradas y restringidas por el lenguaje natural y son usadas en orden de reducir la ambigüedad de éstas.

Semi-formal.- Estas ontologías están expresadas en un lenguaje artificial definido formalmente como los lenguajes de marco.

Formalmente riguroso.- Son las ontologías que están precisamente definidas con semántica formal, ejemplo de ello son los lenguajes basados en lógica [(Uschold, Gruninger, 1996), (Roche, 2003)].

Y pueden ser clasificadas también, de acuerdo al tipo de conocimiento que será transmitido por la ontología:

- **Ontologías genéricas**.-Estas ontologías cubren conceptos generales definidos independientemente del dominio de la aplicación y que puede ser usada en varios dominios.
- **Ontologías de dominio**.- Estas ontologías están especificadas para un dominio en particular y cubren conceptos genéricos del mismo, lo que permite que sean reutilizadas en tareas diferentes que están relacionadas con el dominio de discurso.
- **Ontologías de aplicación**.- Estas ontologías usan el conocimiento especifico para una tarea en particular, que incluye conocimiento especifico de expertos para la aplicación, por lo general estas ontologías no se pueden reutilizar.
- **Meta-ontología**.- Esta ontología especifica la representación del conocimiento usada para definir lo conceptos del dominio y ontologías genéricas [(Roche, 2003), (Guarino, 1995)].

3.4 Metodologías para la construcción de ontologías

Para implementar una ontología es necesario elegir la metodología adecuada, dependiendo del objetivo y el uso de la misma. En el trabajo de (Guzmán et al., 2012), se describen las diferentes metodologías que existen para el diseño e implementación.

CYC.- Publicada por Lenat y Guha en 1990, describen de manera general los pasos para la construcción de ontologías; El primero, consiste en extraer manualmente el conocimiento común que esta implícito en diferentes fuentes; para después cuando se tenga suficiente conocimiento en la ontología adquirir nuevo conocimiento común usando herramientas del procesamiento de lenguaje natural o aprendizaje computacional.

Uschold y King.- Publicada por Uschold y King en 1995, empleado en el Modelo Enterprise, donde recrean una serie de pasos que permiten plasmar y especificar los conocimientos que se tiene sobre un dominio especifico, centrando sus esfuerzos en la forma en la cual representar conocimientos.

Grüninger y Fox.- Publicada por Grüninger y Fox en 1995, fue desarrollada paralelamente de la metodología del Uschold y King y fue usada para construir las ontologías del proyecto *TOVE* (Toronto Virtual Enterprise); Este enfoque utiliza un conjunto de preguntas en lenguaje natural, llamadas cuestiones de competencia, para determinar el ámbito de la ontología y extraer los conceptos principales, sus propiedades, relaciones y axiomas.

Kactus.- Surge en 1996, usando el dominio de las redes eléctricas para desarrollar ontologías como parte del proyecto *Spirit KACTUS*. Esta metodología usa una base de conocimiento por medio de un proceso de abstracción.

Methontology.- Esta metodología es desarrollada por (Fernández et al., 1997). Esta metodología define actividades para la planificación del proyecto, la calidad del resultado, la documentación, etc. La metodología esta compuesta por once tareas definidas en el trabajo de (Corcho et al., 2005) mostradas en la Figura 3.1.

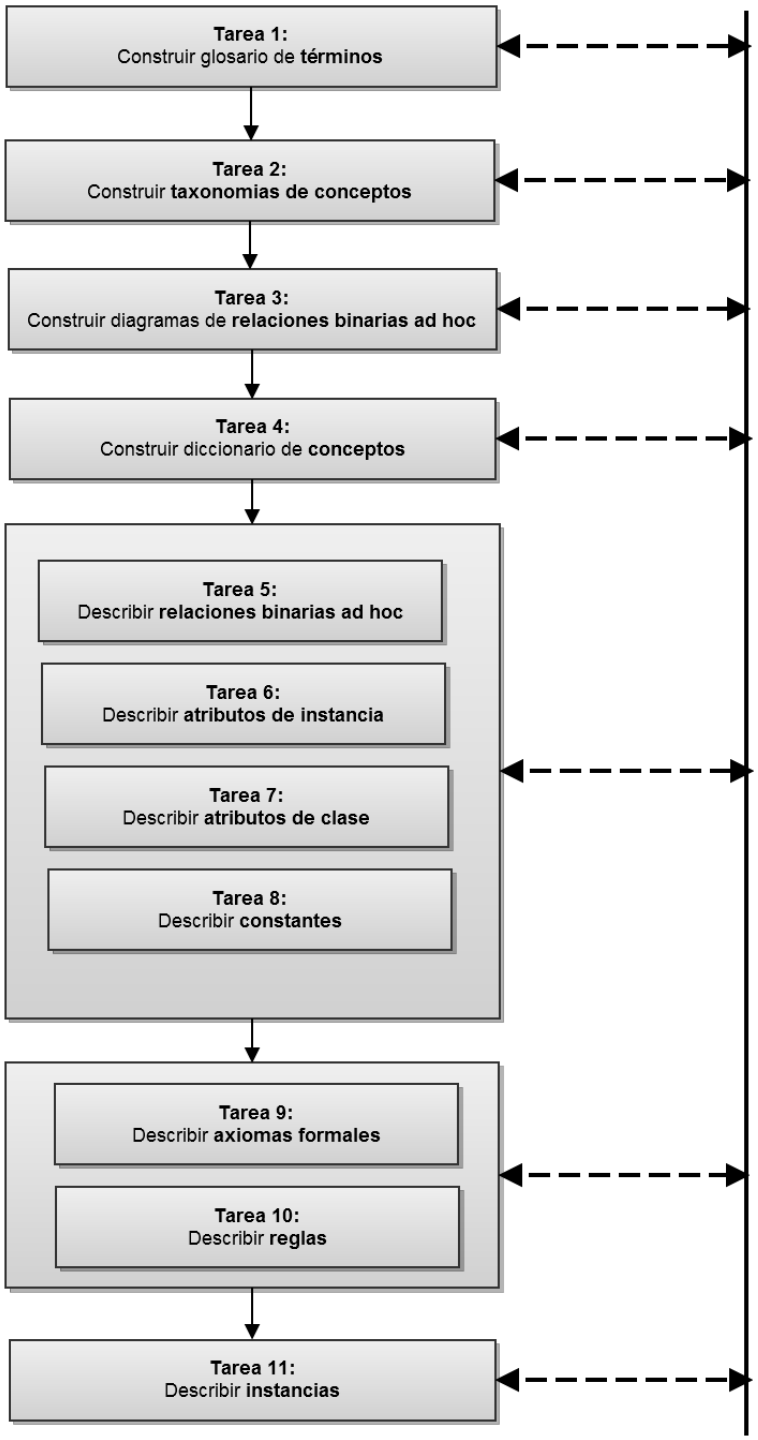

Figura. 3.1. Tareas de la actividad de Conceptualización según Methontology (Corcho et al., 2005).

Sensus.- Surge en 1997, como un nuevo método para construir ontologías, la cual constituye un enfoque *top-down* para derivar ontologías especificas del dominio a partir de grandes ontologías. En ésta se identifican un conjunto de términos semilla que son relevantes para el dominio particular. Tales términos se enlazan manualmente a una ontología de amplia cobertura, con lo cual el usuario puede seleccionar los términos más relevantes y así acotar la ontología *Sensus*. Después, el algoritmo devuelve un conjunto de términos estructurados jerárquicamente para describir un dominio, que puede ser usado como para la base de conocimiento.

ON-TO- KNOWLEDGE.- Desarrollada por (Sure et al., 2003), esta metodología aplica ontologías a la información disponible electrónicamente, para mejorar la calidad de la gestión de conocimiento en organizaciones grandes y distribuidas. Además, incluye la identificación de metas que deberían ser conseguidas por herramientas de gestión de conocimiento, y está basada en el análisis de escenarios de uso y en los diferentes papeles desempeñados por trabajadores de conocimiento y accionistas en las organizaciones.

NeOn.- Diseñada por (Suárez et al., 2012), esta metodología diseñada para la construcción de redes de ontologías esta basada en escenarios. Que se apoyan en los aspectos de colaboración de desarrollo de ontologías y en la reutilización, así como en la evolución dinámica de las redes de ontologías en entornos distribuidos.

Las claves de la Metodología *NeOn,* son un conjunto de nueve escenarios para la construcción de ontologías y redes de ontologías, como se muestra en la Figura 3.2, haciendo hincapié en la reutilización de los recursos ontológicos y no ontológicos, la reingeniería y la fusión, y teniendo en cuenta la colaboración y el dinamismo. El Glosario de Procesos y Actividades identifica y define aquellos procesos y actividades involucrados en el desarrollo de las redes de ontologías.

Directrices metodológicas para diferentes procesos y actividades del proceso de desarrollo de la ontología de la red, tales como la reutilización y la reingeniería de los recursos ontológicos y no ontológicos, la especificación de los requisitos de la ontología, la localización de la ontología, la programación, etc. Todos los procesos y actividades se describen con (a) una tarjeta llena, (b) un flujo de trabajo, y (c) ejemplo.

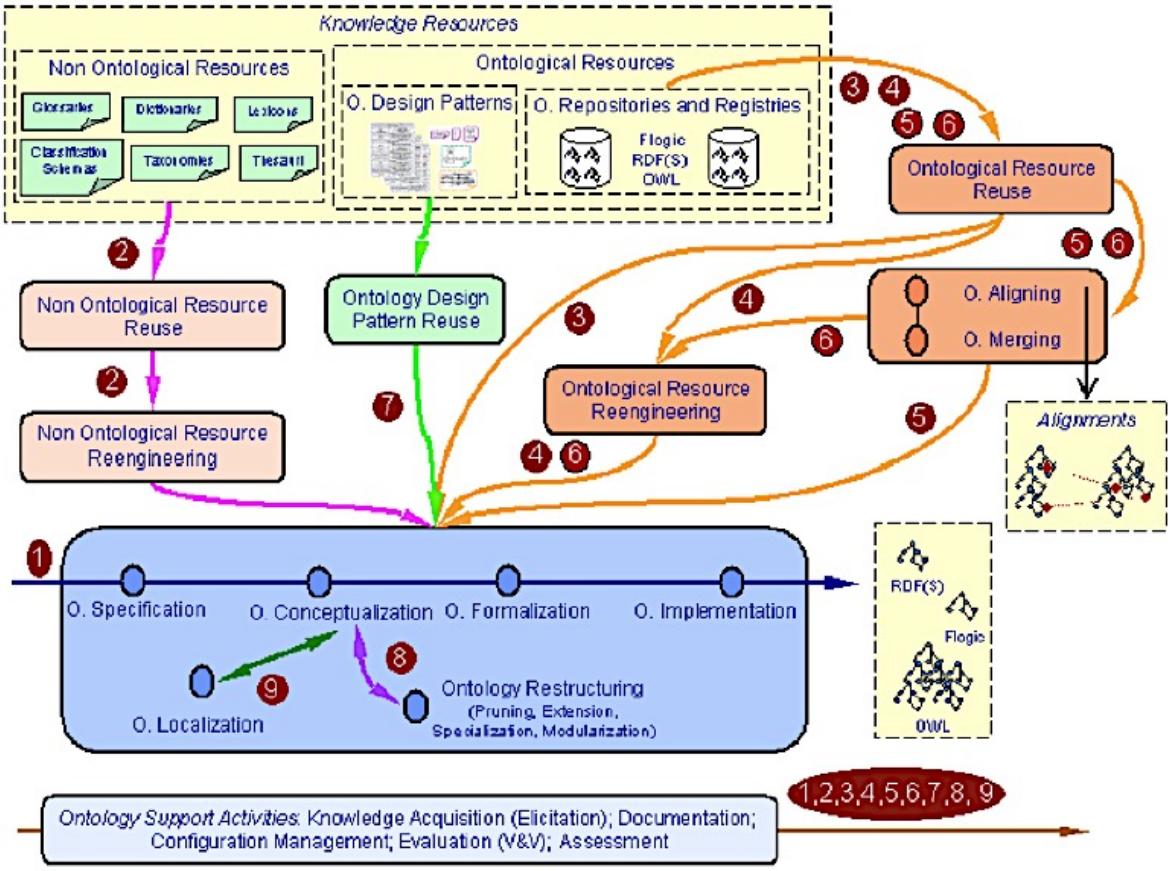

Figura. 3.2. Escenarios para la construcción de ontologías y ontologías de red, usando NeOn (Suárez et al., 2012).

Escenario 1: Desde la especificación de la aplicación. La red de ontologías es desarrollada sin volver a utilizar los recursos existentes. Los desarrolladores deben especificar los requisitos de la ontología. Después de eso, se asesora para llevar a cabo una búsqueda de recursos potenciales para ser reutilizados. A continuación, la actividad de planificación se debe realizar, y los desarrolladores deben seguir el plan.

Escenario 2: La reutilización y reingeniería de los recursos no ontológicos (NOR). Los desarrolladores deben llevar a cabo el proceso de reutilización NOR para decidir, de acuerdo con los requisitos de la ontología, que *NORs* pueden ser reutilizados para construir la red de la ontología. A continuación, los *NORs* seleccionados deben ser volver al proceso de re-ingeniería ontológicas.

Escenario 3: La reutilización de los recursos ontológicos. Los desarrolladores utilizan recursos ontológicos (ontologías como un conjunto de módulos ontológicos, y/o declaraciones) para construir redes de ontologías.

Escenario 4: La reutilización y re-ingeniería de los recursos ontológicos. Los desarrolladores de ontologías reutilizan los recursos y reorganizar los recursos ontológicos.

Escenario 5: La reutilización y la fusión de los recursos ontológicos. Este escenario se produce cuando varios recursos ontológicos en el mismo dominio que se seleccionan para su reutilización, y los desarrolladores desean crear un nuevo recurso ontológico con los recursos seleccionados.

Escenario 6: Reutilización, la fusión y re-ingeniería de los recursos ontológicos. Los desarrolladores de ontologías reutilizan, combinan y reorganizan los recursos-ontológicos. Este escenario es similar al Escenario 5, pero en este caso los desarrolladores deciden reorganizar el conjunto de recursos combinados.

Escenario 7: Reutilización de los patrones de diseño de ontologías (ODPs). Los desarrolladores de ontologías acceden a repositorios de reutilización ODPs.

Escenario 8: Reestructuración de recursos ontológicos. Los desarrolladores de ontologías reestructuran (modularizan, podan, extienden y / o especializan) recursos ontológicos que deben integrarse posteriormente en la red de ontologías.

Escenario 9: Localización de recursos ontológicos. Los desarrolladores de ontologías adaptan una ontología a otras lenguas y la cultura las comunidades, obteniendo así una ontología multilingüe.

3.5 Modelo *RDF*

RDF[5] es un modelo estándar para el intercambio de datos en la Web, el cual tiene características que facilitan la fusión entre diferentes esquemas, y tiene soporte para la evolución de esquemas sobre el tiempo sin que se requiera que todos los datos del consumidor sean cambiados.

También extiende la estructura de vinculación de la Web para usar *URIs* en los nombres de las relaciones entre los objetos, que sirve para relacionar dos objetos y formar unta tripleta. Usando este modelo, es posible estructurar y semi-estructurar datos para ser mezclados, expuestos o compartido desde diferentes aplicaciones.

La estructura de vinculación es de forma directa, empleando un grafo etiquetado donde los bordes representas los links nombrados entre dos recursos, representado por los grafos de un nodo. Donde cada nodo esta conformado por tres elementos:

- **Sujeto**.- Puede ser el nodo inicial, una instancia, una entidad o una característica.
- **Predicado**.- Puede ser un verbo, una propiedad, un atributo, una relación, un miembro, un enlace o una referencia.
- **Objeto**.- Puede ser un valor, un nodo final o algún valor no literal que pueda ser usado como un sujeto.

Estos tres valores conforman una tripleta en RDF, como se ve en la Figura 3.3 y puede ser representada mediante una URI (*Uniform Resource Identifier*) o identificador de recursos únicos. Los sujetos y objetos son llamados nodos y pueden ser representados como un nodo en blanco.

[5] http://www.w3.org/RDF/

Los objetos pueden también ser representado como un valor literal. También un mismo nodo puede tener el rol de sujeto en una arista, y en otras ser un objeto.

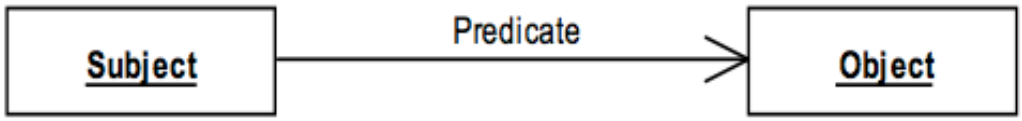

Figura. 3.3. Tripleta en RDF

3.6 Lenguaje de consulta SPARQL

SPARQL es el lenguaje de consultas para *RDF,* así como un protocolo con especificaciones para realizar consultas remotas.

SPARQL Language[6] es un lenguaje, que es utilizado para realizar consultas a grafos en *RDF* a través de patrones de coincidencias. Este lenguaje permite incluir patrones básicos conjuntivos, filtros por valores, patrones opcionales y patrones de disyunción .

El protocolo *SPARQL* contiene una interfaz gráfica llamada *SparqlQuery* ,que permite realizar consultas. Para realizar consultas al protocolo SPARQL se necesita implementar un *Endpoint*., este es un servicio del protocolo *SPARQL* el cual esta definido conforme al estándar *SPROT*. Que permite a usuario realizar consultas a una base de conocimiento empleando el lenguaje *SPARQL*. La respuesta, puede ser retornada en diferentes formatos.

Tanto las consultas, como la presentación de resultados, debe ser implementada y recuperada en alguna aplicación independiente para poder ser interpretadas por un usuario, estas consultas utilizan como base los archivos RDF que se encuentran almacenados en un *Endpoint*, conocido también como *triple store.*

[6] http://www.w3.org/TR/rdf-sparql-query/

Un *triple store* es una base de datos especialmente diseñada para el almacenamiento y recuperación de tripletas, siendo una tripleta una entidad de datos compuesta de sujeto-predicado-objeto, como "Imelda tiene 29" o "Imelda conoce Vladimir". Al igual que una base de datos relacional, se almacena la información en un *triple store* y lo recupera a través de un lenguaje de consulta.

A diferencia de una base de datos relacional, *un triple store* está optimizado para el almacenamiento y recuperación de tripletas. Además de las consultas, las tripletas usualmente se pueden importar ó exportar utilizando el formato *RDF* ó en algún otro formato como *HTML* o *JSON* (Candillier et al., 2007).

3.7 GeoSPARQL

GeoSPARQL[7] es un estándar de la OGC, el cual define una extensión de funciones para *SPARQL* [W3r], un conjunto de reglas *RIF*[8] y un núcleo *RDF/OWL*[9] vocabulario para información geográfica basada en el *General Feature Model* (ISO 19109), *Simple Features* [ISO 19125-1], *Feature Geometry* (ISO 19107) y SQL MM (ISO/IEC 13249-3).

El estándar *GeoSPARQL* representa soporte y consultas de datos geoespaciales en la Web semántica. Éste define un vocabulario para representar datos en RDF, y una extensión para el lenguaje de consulta *SPARQL* para procesar datos geoespaciales.

El estándar sigue un diseño modular, que comprende diferentes componentes, los cuales se enumeran y describen a continuación:

1. **Un *core.-*** Define una clase *top-level* de *RDFS/OWL* para objetos espaciales.
2. **Un vocabulario topológico.-** Define propiedades *RDF* para afirmar y consultar las relaciones topologías entre objetos espaciales.

[7] http://www.opengeospatial.org/standards/geosparql
[8] http://www.w3.org/TR/rif-core/
[9] http://www.w3.org/RDF/

3. **Una geometría.-** Define tipos de datos *RDFS* para serializar datos geométricos, propiedades *RDF* geométricamente relacionadas y funciones topológicas no espaciales para objetos geométricos.
4. **Una topología geométrica.-** Define funciones topológicas de consulta.
5. **Una vinculación *RDFS.-*** Define un mecanismo de coincidencias implícitas en tripletas *RDF* que son derivadas en un *RDF* o en un esquema *RDFS*.
6. **Un *query rewriter***.- Define reglas, para transformar un patrón de tripletas a una relación topológica entre dos objetos en una *query* equivalente, que envuelve concretamente funciones de consulta geométrica y topológica.

Cada una de estas componentes forman parte de los requerimientos para *GeoSPARQL*, el cual está diseñado para integrarse a sistemas cualitativos basados en razonamiento espacial y sistemas cuantitativos basados en computo espacial.

4. Metodología

En este capítulo se describe, de manera general la metodología propuesta, y de manera detallada cada una de las etapas que la componen. Describiendo las técnicas, estándares y protocolos que fueron implementados en cada etapa.

4.1 Metodología propuesta

La metodología propuesta en este trabajo se enfoca en la detección y ubicación de eventos viales en los *tweets* de la zona metropolitana de la Ciudad de México y se compone de cuatro etapas principales (ver Figura 4.1):

1. **Recolección y confiabilidad:** En esta etapa, los *tweets* publicados son recolectados y es evaluada la confiabilidad de la fuente, de esta manera se descartan las publicaciones que no vengan de una fuente confiable. El resto de los *tweets* es almacenado en nuestra base de datos. Por otra parte, se calculan y ponderan los términos frecuentes empleados en los *tweets* y son generados los diccionarios para la etapa de clasificación (Manifestaciones, Accidentes vehiculares, Congestión vehicular y Obras públicas) , los diccionarios para la etapa de estandarización (Abreviaturas, Acrónimos, *Hashtags* y *Nicknames*) y el diccionario de relaciones espaciales.

2. **Estandarización y Clasificación:** En esta etapa se realiza la estandarización de los *tweets*, empleando los diccionarios de Abreviaturas, Acrónimos, *Hashtags* y *Nicknames* generados en la etapa anterior. Posteriormente, se emplea el enfoque de *N-Gram-Based Text Categorization* (Cavnar & Trenkle, 1994) y los diccionarios de Manifestaciones, Accidentes vehiculares, Congestión vehicular y Obras públicas para clasificar los *tweets*. Finalmente, se almacenan en la Base de Datos.

3. **Geocodificación Semántica:** Esta etapa consiste de tres tareas: la primera es la obtención de los nombres referentes a ubicaciones en los *tweets* utilizando *Named Entity Recognition (NER)* (Grishman & Sundheim, 1996), con estos nombres se genera una consulta a nuestro *Endpoint* para obtener su componente geográfica. La segunda tarea se encarga de la identificación de las Expresiones Regulares en el texto de los *tweets* las cuales representan relaciones espaciales, con estas expresiones se genera una consulta en el *Endpoint* para obtener la operación u operaciones espaciales que se deben ejecutar. Por último, la tercera tarea utiliza el resultado de las consultas anteriores para generar otra consulta con los objetos geográficos y las operaciones espaciales que serán aplicadas, de esta manera, el *Endpoint* retornará los segmentos de vialidad que son afectados por el evento.

4. **Visualización y Evaluación:** Por último, en esta etapa los segmentos de vialidad afectados son visualizados en un mapa, representados por diferentes colores de acuerdo al tipo de evento y actualizados cada 5 minutos. Por otro lado, se compara la geocodificación obtenida con una geocodificación realizada por un usuario, de esta forma, se obtiene una línea de comparación. De igual forma, se realiza una comparativa con *Google Maps* y finalmente, son evaluadas las medidas de *precision*, *recall* y la medida *F*.

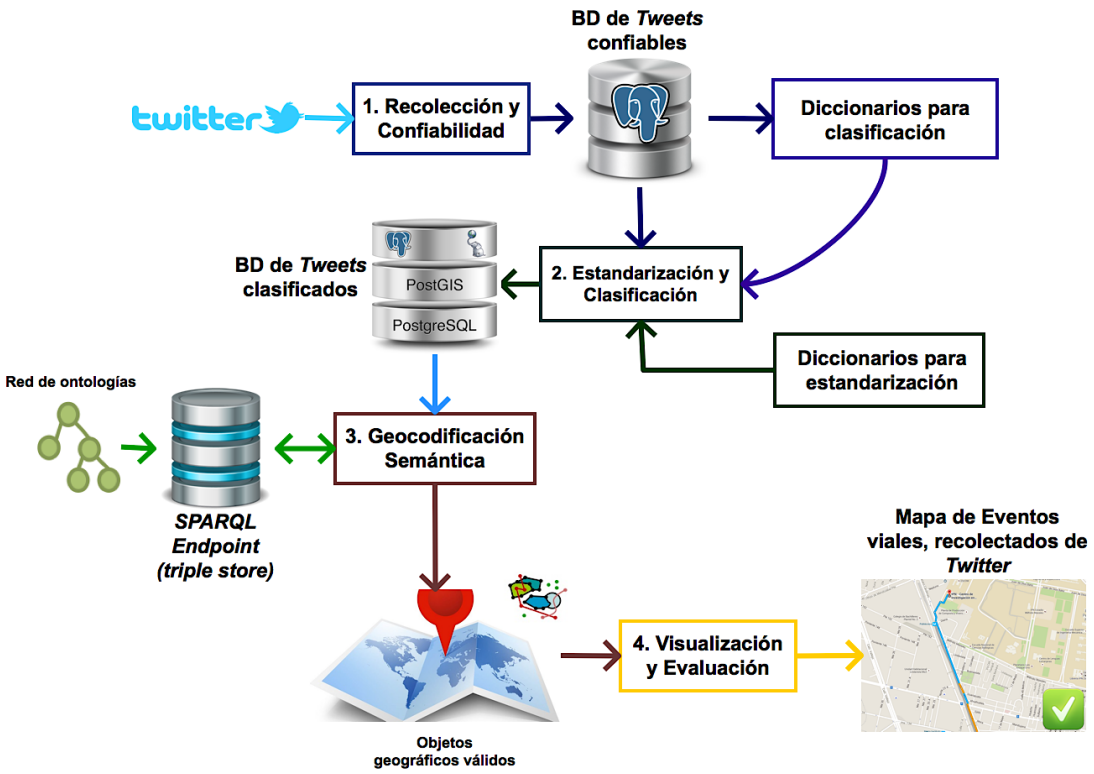

Figura 4.1. Metodología propuesta.

4.2 Recolección y Confiabilidad

En el flujo de *tweets* publicados diariamente, muchos hacen referencia a eventos viales, pero no se pueden considerar confiables todos ellos, por ese motivo en éste módulo, la primera tarea que se realiza es recolectar los *tweets* relacionados a los eventos viales y la información de los usuarios que los publican empleando *python* y el API de *tweepi* , todo esto se almacena en una base de datos temporal.

La información recolectada de las cuentas consta de: id de usuario (número único asignado por *Twitter* a cada usuario), el nombre público de la cuenta de *Twitter*, la ubicación publicada en el perfil, la fecha de creación, el número de seguidores, el número de amigos y el número de *tweets* publicados. Por otro lado la información recolectada de los *tweets* consta de: id del *tweet* (número único asignado por *Twitter* a cada mensaje), mensaje, usuario que lo publica, si es un *retweet* , número de *retweets,* usuario que hizo el *retweet* y por último fecha y hora

de publicación. Con esta información se calcula el *Nivel de Confiabilidad del Usuario (NCU).*

Primero se ordenan todos los usuarios de mayor a menor por cantidad de seguidores, a los cuales llamaremos usuarios principales up . Se selecciona el usuario principal up con más seguidores y, de manera aleatoria es seleccionado un seguidor, al cual llamaremos u . N es el número de *tweets* publicados por u que mencionan al usuario up y $upTweet$ son los *tweets* publicados por un seguidor de u que mencionan a up ver Figura 4.2.

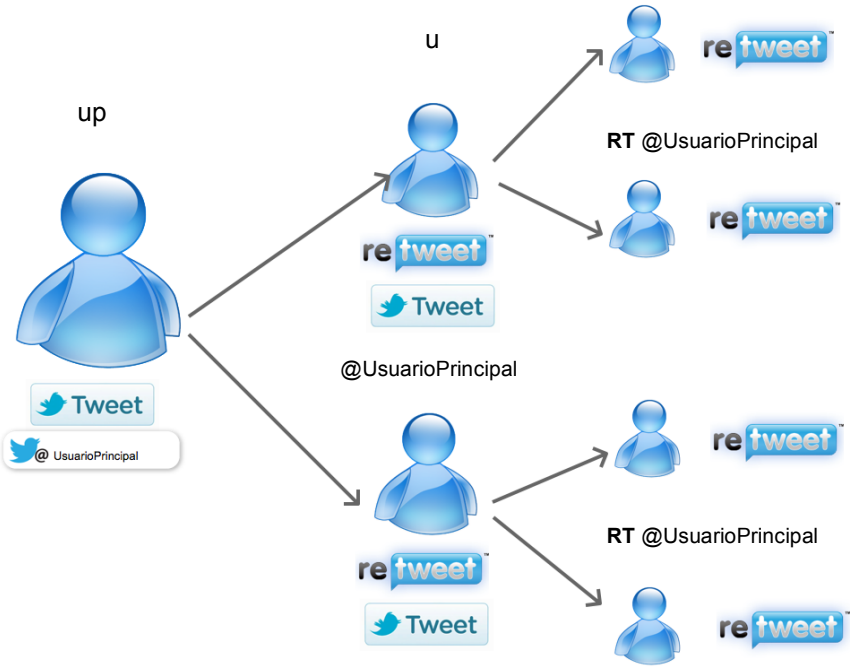

Figura 4.2. Descripción de los elementos para el NCU.

Ahora, se calcula el promedio de *tweets* que fueron *retwitteados* por u y los seguidores de u, como se muestra en la ecuación 1.

$$AvgRT(u, up) = \frac{1}{N}\sum_{i=1}^{N} RT(u, upTweet_i) \qquad (1)$$

Con estos datos es posible calcular $Umencion(u, up)$ que consiste en el conjunto de todos los *tweets* publicados por los seguidores de up en donde lo mencionen, este cálculo se realiza mediante la ecuación 2.

$$Umencion(u, up) = \frac{Tweet(u, upTweet)}{Tweet(u)} \quad (2)$$

En donde $Tweet(u)$ es el número total de *tweets* publicados por el seguidor u y $Tweet(u, upTweet)$ es el número de veces que menciona a up en sus *tweets*.

Por lo tanto, el *Nivel de Confiabilidad del Usuario (NCU)* se define como: El promedio de *retweets* $AvgRT$ de los seguidores de u ,que mencionan a los usuarios principales y el conjunto de todos los *tweets* de u que mencionan a up, mostrada en la ecuación 3.

$$NCU(up) = AvgRT(u, up) \times Umencion(u, up) \quad (3)$$

El *NCU* es colocado por cada usuario y se almacena en una tabla en nuestra Base de Datos, posteriormente se recuperan todos los *tweets*, el usuario que los publica, si son *retweets* y el usuario que hizo el *retweet*.

Primero se descartan los *retweets* que tengan el mismo contenido y el *NCU* más bajo, después se ordenan los *tweets* restantes de mayor a menor *NCU* y se descartan todos los *tweets* que sean publicados por usuarios con menos de 1000 seguidores o con un valor de NCU menor a 0.5 que es la mitad del valor máximo que es 1, para este trabajo se considera el valor de 1000 seguidores como el mínimo para considerarse confiable debido a que después de hacer un análisis de diversas cuentas, el promedio de seguidores es de 10,000 y 1000 es el 10%, por último, los *tweets* restantes son almacenados en una tabla en la Base de Datos para ser procesados en la siguiente etapa.

4.3 Estandarización y Clasificación

Después de analizar el conjunto de *tweets* recolectados, se encuentra una cantidad considerable de variaciones en la forma de escribir el nombre de las vialidades, las abreviaturas, el nombre de puntos de interés y el uso frecuente de *nicknames* o apodos para hacer referencia a cualquier tipo de lugar, es por eso que se desarrolló un *script* en *python* que encuentra los N-gramas más frecuentes. Un N-grama es una N-palabra segmentada de una oración más grande (Cavnar & Trenkle, 1994).

Se obtienen los N-gramas de cada *tweet* y se ordenan por número de ocurrencias, los N-gramas con más ocurrencias son seleccionados y se colocan manualmente en alguno de los diccionarios que se muestran en la Figura 4.3.

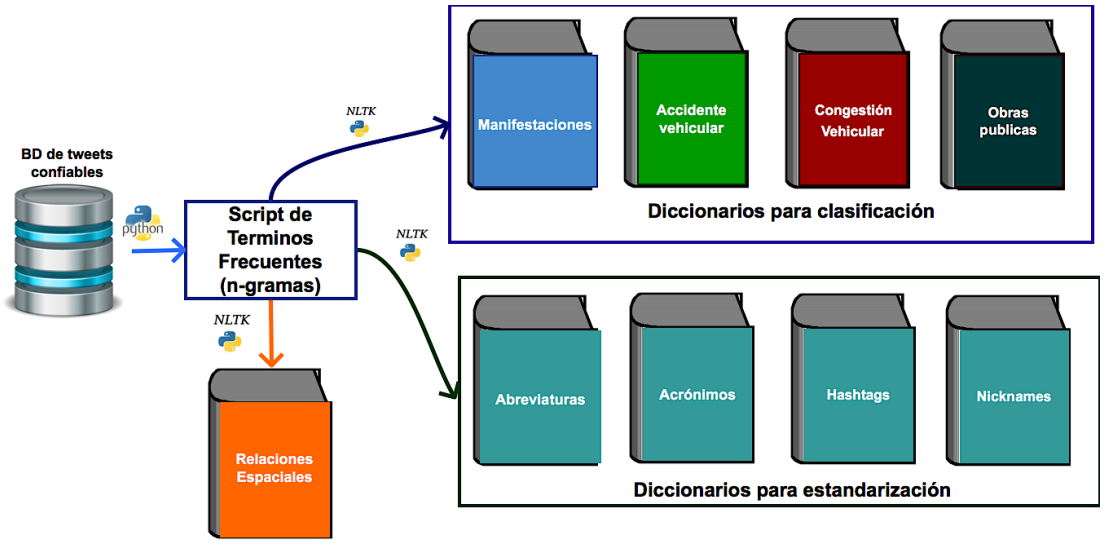

Figura 4.3. Creación de los Diccionarios con los Términos Frecuentes en los *tweets*.

De la lista de N-gramas más frecuentes, nosotros identificamos manualmente 190 abreviaturas, 73 acrónimos, 756 *hashtags* y 2,500 *nicknames* usados comúnmente.

Por otro lado, se obtuvieron 840 términos referentes a eventos viales que fueron clasificados en los diccionarios de Manifestaciones, Accidente vehicular, Congestión vehicular y Obras públicas respectivamente, dependiendo de sus características. Finalmente se identificaron 146 combinaciones de palabras y preposiciones que representan las relaciones espaciales usadas habitualmente. Cada diccionario se almacenó en un archivo de texto, con el objetivo de agilizar el procesamiento de la información.

Después de obtener los diccionarios, inicia el proceso de estandarización del conjunto de *tweets* en la base de datos, remplazando el texto original, por el que se encuentra en los diccionarios de Abreviaturas, Acrónimos, *Hashtags* y *Nicknames*. Por ejemplo, en todos los casos que una vialidad contenga la

abreviatura “Av.” es remplazada por la palabra “Avenida” o en el caso del *hashtag* “#Periférico” es sustituido por “Periférico”. También, son eliminados todos los hipervínculos, se quitan los acentos y los espacios en blanco. En nuestro caso los números y preposiciones se mantienen en el texto, por que serán utilizados posteriormente, ver Figura 4.4.

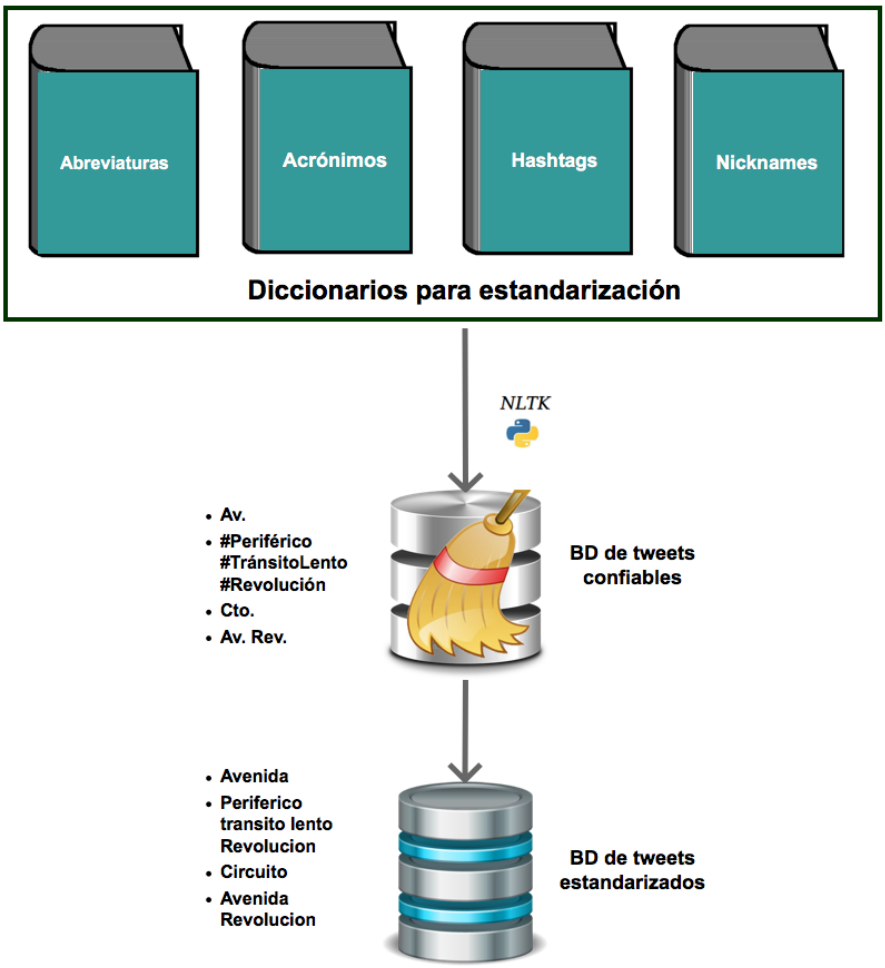

Figura 4.3. Proceso de Estandarización de los *tweets*.

Para la Clasificación, empleamos el enfoque de *N-Gram-Based Text Categorization* (Cavnar & Trenkle, 1994). El algoritmo se basa en el cálculo y la comparación de perfiles, empleando la frecuencia de N-gramas.

Primero, se calculan los perfiles del conjunto de entrenamiento, el cual está compuesto del diccionario de Manifestaciones, Accidente Vehicular,

Congestión Vehicular y Obras públicas. Después, se calcula el perfil del *tweet* que se quiere clasificar. Finalmente, se calcula la distancia entre el perfil del *tweet* y los perfiles del conjunto de entrenamiento, seleccionando la categoría que tenga el perfil con la menor distancia al perfil del *tweet* . Los pasos principales se muestran en la Figura 4.4a y en la Figura 4.4b se muestra de una manera general el funcionamiento.

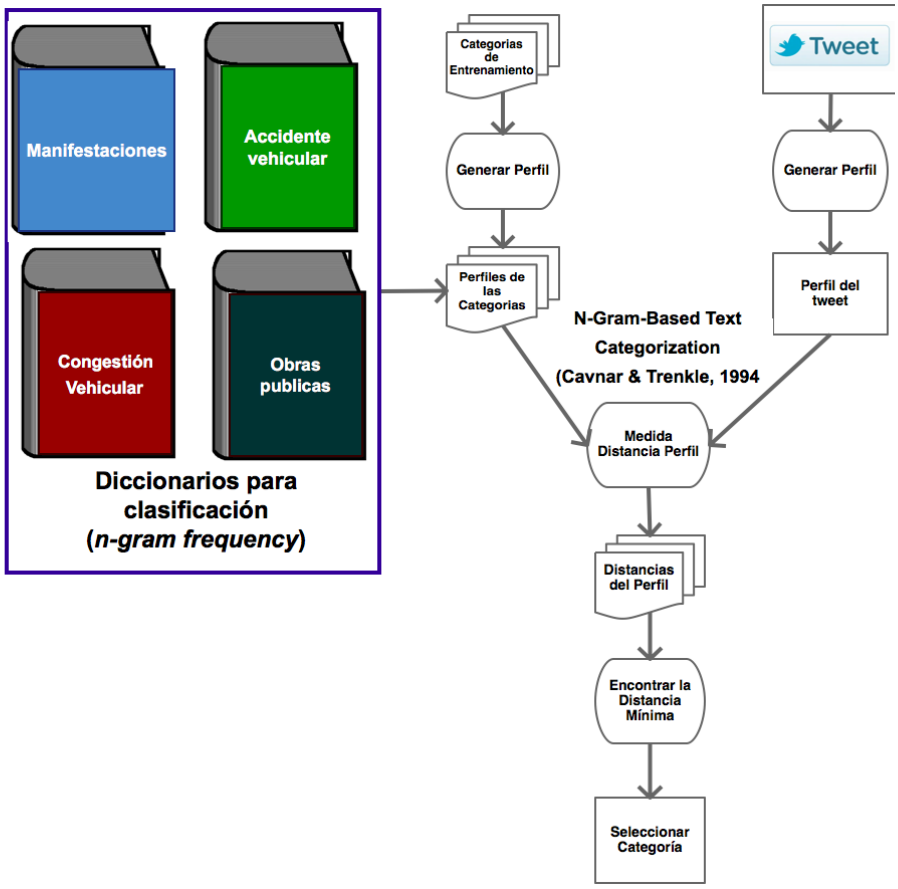

Figura 4.4a. *N-Gram-Based Text Categorization*.

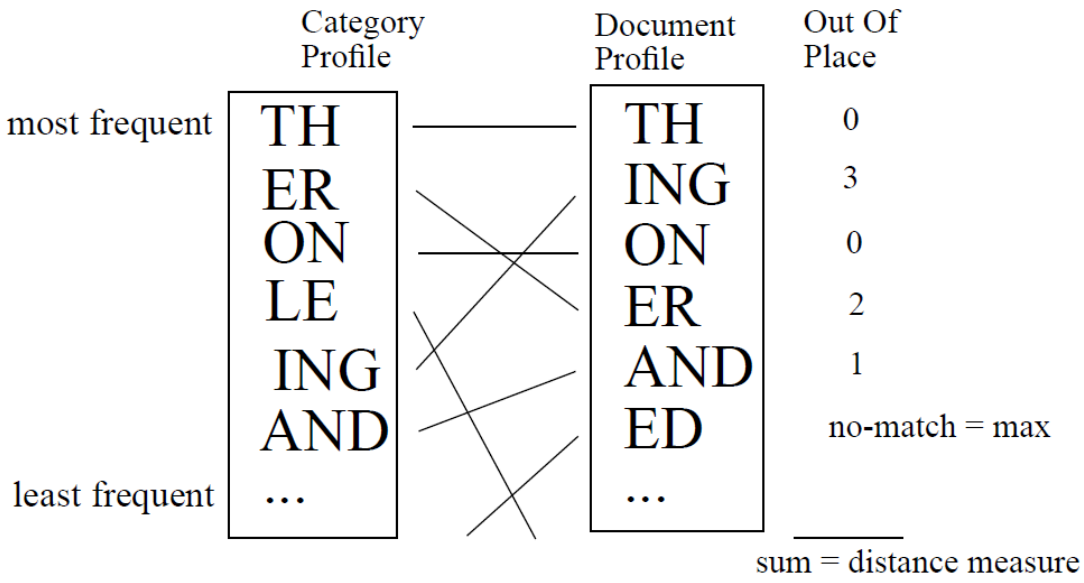

Figura 4.4b. *N-Gram-Based Text Categorization*.

4.4 Geocodificación Semántica

Las direcciones contienen múltiples componentes (es decir, número de casa, nombre de la calle, tipo de calle, dirección, ciudad, estado y código postal), utilizados para expresar una ubicación en la superficie de la Tierra. En un GIS, las direcciones se convierten en entidades en un mapa a través del proceso de geocodificación. Este proceso implica una serie de pasos, mediante el cual se asignan coordenadas a una dirección, comparando los elementos de la dirección con los de los datos de referencia (Hart & Zandbergen, 2013).

En el caso de *Twitter*, hemos identificado que los *tweets* publicados son ricos en información geográfica, como nombres de calles, nombres de sitios de interés y referencias a lugares donde ocurren cierto tipo de eventos. Esta es la razón principal por la cual consideramos los *tweets* como fuente de entrada a nuestro algoritmo de geocodificación.

En la Figura 4.5, se proponen un conjunto de pasos correspondientes a la etapa de geocodificación semántica.

1. Identificación de lugares en los *tweets*, usando *Named Entity Recognition* (NER).

2. Obtención de las Relaciones Espaciales en los *tweets*, usando un diccionario de Expresiones Regulares (ER) que representan Relaciones Espaciales.
3. Construcción de la consulta para recuperar los geo-objetos y de la consulta para recuperar la operación espacial.
4. Construcción de la consulta para recuperar las vialidades afectadas por el evento.

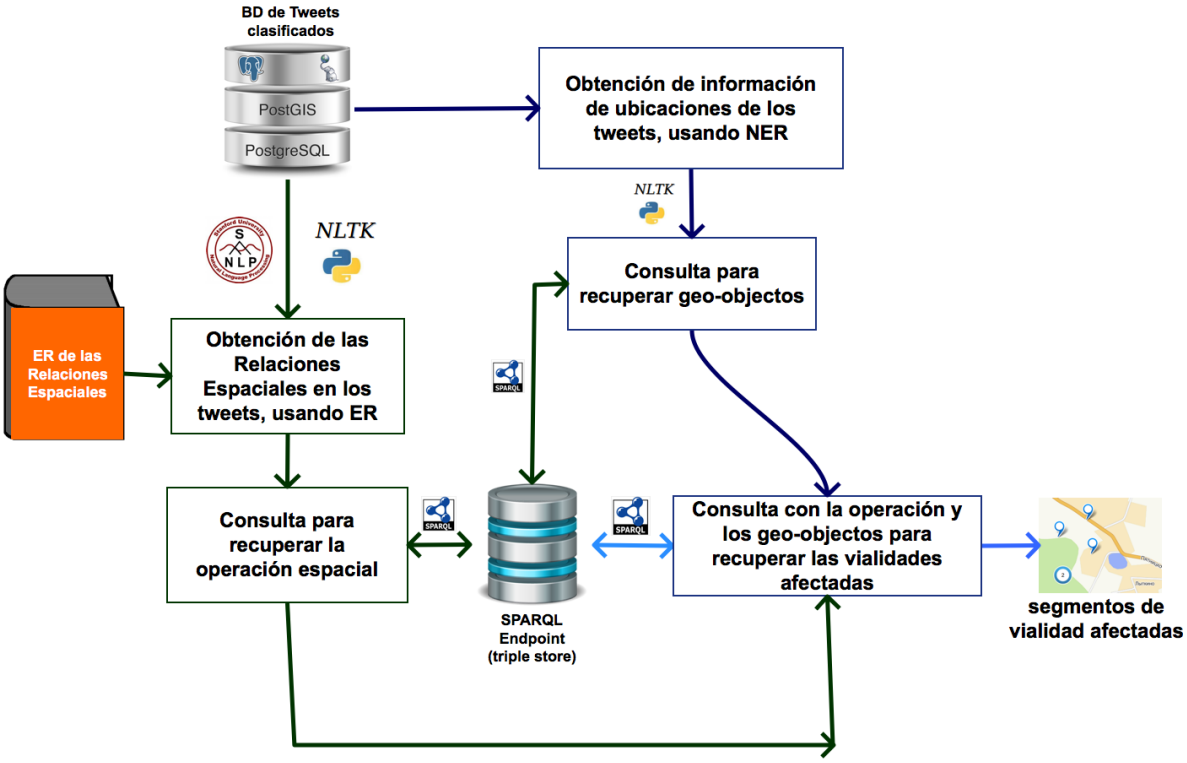

Figura 4.5. Geocodificación Semántica.

4.4.1 Identificación de lugares en los *tweets*, usando NER

Named Entity Recognition (*NER*), también conocido como *Entity Identification* (EI) y *Entity Extraction*, involucra la identificación de nombres propios en un texto, y su clasificación en un conjunto de categorías predefinidas. Las tres categorías universalmente aceptadas son: Personas, Lugares y Organizaciones. Aunque existen otras categorías frecuentes como el reconocimiento de expresiones de fecha y hora, direcciones de correo y medidas (porcentajes, dinero, peso, etc.) (Grishman & Sundheim, 1996).

La tarea de identificación de lugares en los *tweets*, usando NER, consiste en encontrar los nombres de entidades tales como, nombres de calles y avenidas, estaciones de metro y puntos de interés, usando un modelo basado en las reglas del lenguaje para encontrar las palabras que hagan referencia a estas entidades.

Para obtener los nombres propios de lugares contenidos en el texto de los *tweets*, se desarrolló un *script* en *python,* que emplea el módulo de *Named Entity Classification* incluido en *FreeLing 4.0*[10]. El cuál coloca una etiqueta, dependiendo de la clasificación de la entidad, distinguiendo cuatro clases: Persona (etiqueta NP00SP0), Lugar (NP00G00), Organización (NP00O00) y Otros (NP00V00).

Por ejemplo, supongamos que el siguiente *tweet* es analizado por nuestro script:

“llueve en Avenida Insurgentes y La Raza extreme precauciones”.

En este caso, se identifican “*Avenida Insurgentes*” y “*La Raza*” como *location* (ver Figura 4.6), lo cual indica que fueron reconocidas lugares, resaltando que es necesario un paso adicional que sería la desambiguación, pero este paso se lleva a cabo al momento de realizar la consulta para obtener los geo-objetos.

```
1 llueve              llover              VMIP3S0 VMI
2 en                  en                  SP      SP
3 Avenida_Insurgentes avenida_insurgentes NP00G00 NP
4 y                   y                   CC      CC
5 La_Raza             la_raza             NP00G00 NP
6 extreme             extremar            VMSP3S0 VMS
7 precauciones        precaución          NCFP000 NC

pos=verb|type=main|mood=indicative|tense=present|person=3|num=singular
pos=adposition|type=preposition
pos=noun|type=proper|neclass=location
pos=conjunction|type=coordinating
pos=noun|type=proper|neclass=location
pos=verb|type=main|mood=subjunctive|tense=present|person=3|num=singular
pos=noun|type=common|gen=feminine|num=plural
```

Figura 4.6. Dependencias y NER.

[10] http://nlp.lsi.upc.edu/freeling/

4.4.2 Obtención de las Relaciones Espaciales en los *tweets*, usando un diccionario de Expresiones Regulares (ER)

Esta etapa, consiste en identificar las Relaciones Espaciales contenidas en el texto de los *tweets*, empleando Expresiones Regulares.

Se realizó un script en *python*, que obtiene la relación espacial descrita en el *tweet*, buscando la expresión regular que le corresponda en el diccionario que fue generado en la primera etapa de la metodología.

Por ejemplo, tomando el *tweet* anterior para ser analizado, el resultado sería:

```
“llueve en Avenida Insurgentes y La Raza extreme precauciones”
La relación es: en_y
La ER es: .*(en)+?.*(y)+?.*
```

Que indica cual es la relación que se encuentra en el mensaje y la expresión regular con la que fue identificada; en este caso estos datos serán relevantes para elaborar la consulta para obtener la operación espacial a realizar.

Existen varias relaciones que se componen de elementos similares, pero su representación es completamente diferente, tal es el caso de las relaciones que involucran a la preposición “sobre”, por ejemplo, en el siguiente *tweet*:

“transito lento sobre Viaducto Rio de la Piedad desde Circuito Interior hasta calzada San Antonio Abad”.

Tenemos una relación de tipo “`sobre_desde_hasta`”, que involucra tres entidades y dos operaciones espaciales, además de una validación de más elementos para la expresión regular.

Por otro lado, en el *tweet*: “accidente sobre Eje Central Lazaro Cardenas a la altura de Fray Servando Teresa”. Se tiene una relación tipo “`sobre_a la altura`”, que involucra solamente dos entidades, pero dos operaciones espaciales.

Por último en el caso del *tweet*: "corte de circulacion por retiro de espectacular sobre Avenida General Francisco".

Se tiene una relación de tipo “sobre”, que involucra una sola entidad y una operación espacial, además de una relación que tiene menor complejidad.

Para este tipo de relaciones el script ocupa condicionales que validan los tres casos y definen cual es la relación indicada y la expresión regular correspondiente.

```
patron1=re.compile(".*(sobre)+?.*")
patron2=re.compile(".*(desde)+?.*(hasta)+?.*")
patron3=re.compile(".*(a la altura)+?.*")
if patron1.search(texto) and patron2.search(texto):
     relacion = “sobre_desde_hasta”
       ER= “.*(sobre)+?.*+ .*(desde)+?.*(hasta)+?.*")”
        print ("La relación es: sobre_desde_hasta")
        print("La ER es: .*(sobre)+?.*+
.*(desde)+?.*(hasta)+?.*")
else:
    if patron1.search(texto) and patron3.search(texto):
      relacion = “sobre_a la altura”)
       ER=”.*(sobre)+?.*+ .*(a la altura)+?.*")”
        print ("La relación es: sobre_a la altura")
        print("La ER es: .*(sobre)+?.*+ .*(a la altura)+?.*")
    else:
        if patron1.search(texto):
            relacion=”sobre”
            ER= “.*(sobre)+?.*”
             print("La relación es: sobre")
             print("La ER es:.*(sobre)+?.*")
```

El nombre recuperado de la variable relación, será el que se utilice para buscar en el *Endpoint*, la o las operaciones espaciales que se deben realizar.

4.4.3 *Endpoint* y Red de ontologías

Antes de comenzar a describir las consultas para obtener los geo-objetos y las operaciones espaciales, se describirá el tratamiento que se le dio a los datos geográficos iniciales, lo que es un *Endpoint* y como está compuesta nuestra red de ontologías.

Para este trabajo se crea un *SPARQL Endpoint,* el cual almacenará la red de ontologías, que detallaremos más adelante. En este trabajo se decidió usar como *triple store* a *Parliament* [11], puesto que es posible realizar consultas con el estándar de *GeoSPARQL*.

4.4.3.1 Tratamiento de los datos geográficos iniciales

El tratamiento y procesamiento de los datos geográficos originales correspondientes a la Infraestructura vial de la Ciudad de México y los sitios de interés de la misma, los cuales se encuentran en una Base de Datos Espacial. Primeramente fue necesario estandarizar los nombres de las vialidades con los diccionarios obtenidos en la primera etapa, para cambiar las abreviaturas por palabras completas, tal como es el caso de “Av.” por “Avenida” y “Cto.” por “Circuito”, de igual forma se remplazaron las palabras con acentos y se eliminaron filas en blanco.

Una vez estandarizada la información y almacenada en la en la Base de Datos Espacial, se emplea el *plugin* desarrollado en el trabajo de Rivera et al (2015) para transformar los datos contenidos en las tablas a un archivo *RDF*, éste contendrá tripletas que describan a cada instancia y también la relación entre cada instancia y su componente geográfica, generando una *URI* con esta información. El *RDF* generado además contendrá los prefijos necesarios para el uso del estándar de *GeoSPARQL* y se definen nuevos prefijos relacionados a la red de ontologías, los cuales contendrán la *URI* del concepto o clase a la que pertenecen en la ontología (ver Figura 4.7) .

[11] http://www.parliament.semwebcentral.org/

De tal forma que al llevar a cabo el almacenamiento de los archivos *RDF* de la Red de Infraestructura vial y de los puntos de interés en el *triple store*, estos comiencen a poblar la ontología de manera automática, debido a la relación generada por las *URIS* únicas utilizadas al momento de transformar los datos.

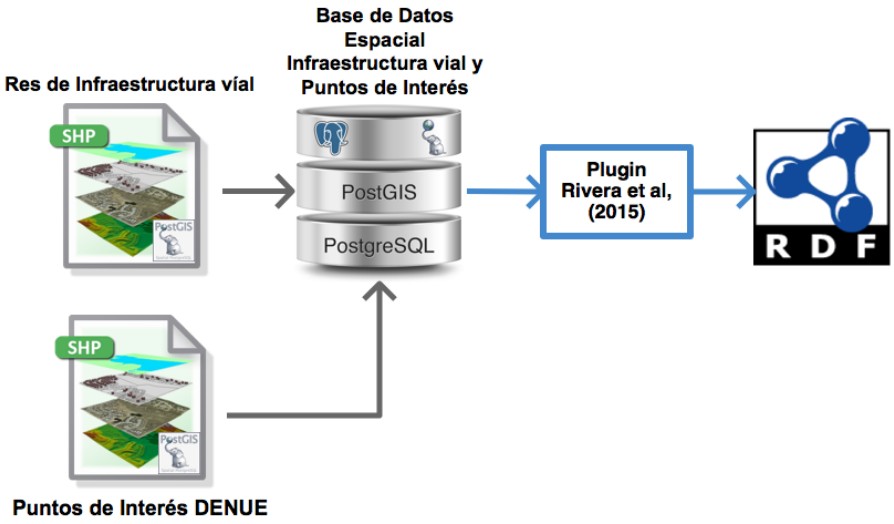

Figura 4.7. Procesamiento de *shapefiles* a *RDF*.

4.4.3.2 Red de ontologías

Una red de ontologías puede ser definida como la colección de ontologías relacionadas entre una variedad de diferentes relaciones tales como mapeo, modularización y control de versiones entre otras (Allocca, C. Et al., 2009). Para este trabajo, la red de ontologías se encuentra en formato *RDF*, con el objetivo de ser almacenado en un *triple store*, las ontologías únicamente contienen los conceptos , las relaciones y los tipos de datos, pero no contienen ninguna instancia. En la Figura 4.8 se muestran las ontologías que conforman nuestra red de ontologías.

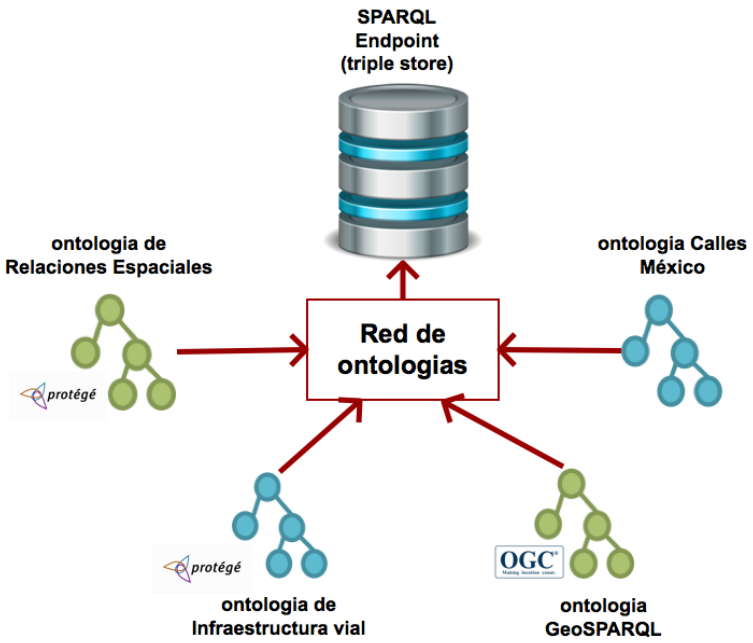

Figura 4.8. Red de ontologías.

- **Ontología de Relaciones Espaciales.** Esta ontología está basada en las relaciones definidas en el artículo de Yecheng, (2011) y en la definición de *GeoSPARQL* y *PostGIS*, adaptando los conceptos al caso de estudio, la ontología está diseñada según la metodología de *METHONTOLOGY* y la herramienta *Protégé* (ver Figura 4.9).

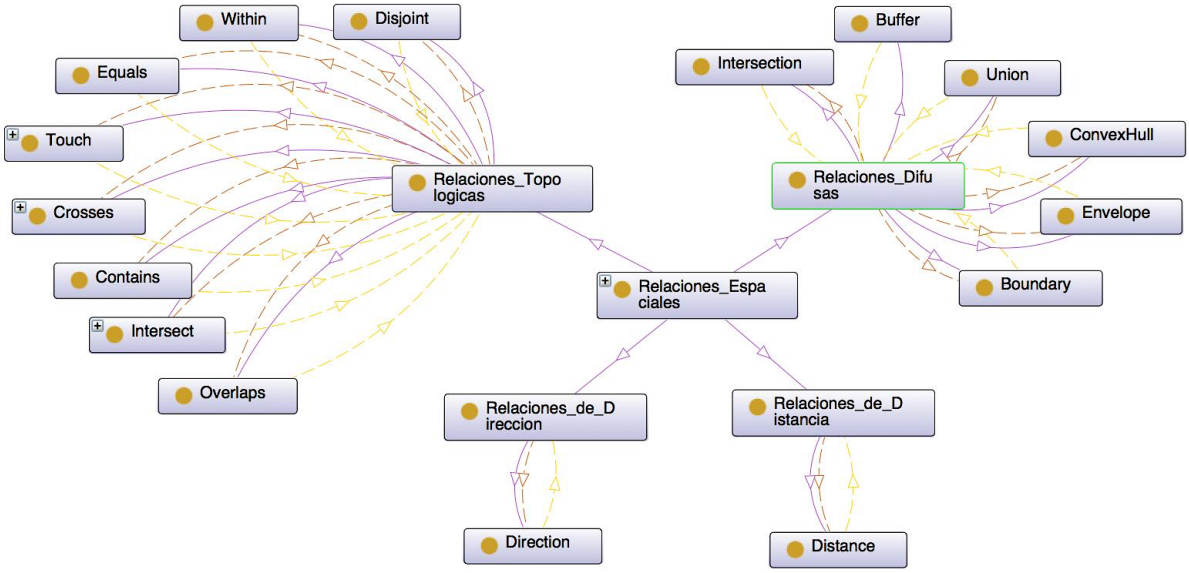

Figura 4.9. Ontología de Relaciones Espaciales.

- **Ontología de Infraestructura vial y puntos de interés (*POIs*).** Esta ontología está basada en la cartografía de la Red de Infraestructura vial de la Ciudad de México, y en los puntos de interés obtenidos del *DENUE* (Directorio Estadístico Nacional de Unidades Económicas-INEGI). Contiene información de las estaciones de los diferentes sistemas de transporte de la Ciudad de México. Esta ontología también está diseñada según la metodología de *METHONTOLOGY* y la herramienta *Protégé* (ver Figura 4.10).

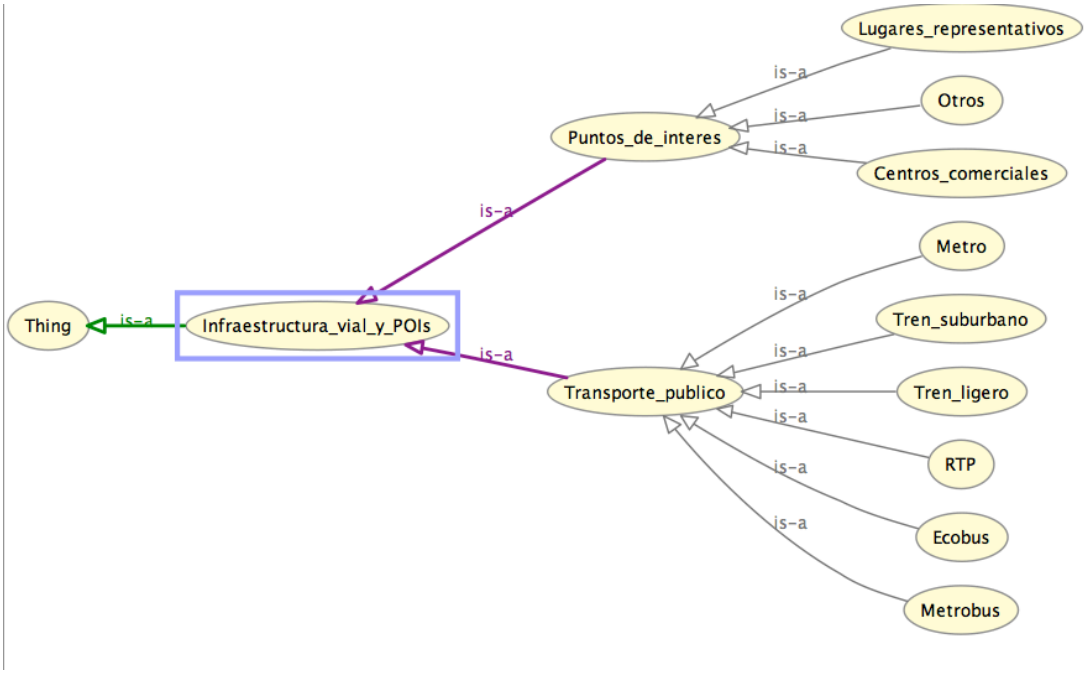

Figura 4.10. Ontología de Infraestructura vial y POIs.

- **Ontología *GeoSPARQL*.** La ontología *GeoSPARQL*, fue desarrollada por el *Open Geospatial Consortium*, esta ontología proporciona a las instancias de la ontología de Infraestructura vial y puntos de interés una componente geográfica, la cual será usada para realizar las consultas espaciales necesarias.
- **Ontología Calles México.** Esta ontología está basada en el trabajo de Rivera et al (2015), pero es enriquecida con una red vial mucho mayor, que fue estandarizada y transformada a *RDF*.

Estas ontologías se encuentran almacenadas en una red de ontologías y actúan como una sola ontología que se encuentra en el *SPARQL Endpoint*, esperando ser poblada.

4.4.4 Construcción de la consulta para recuperar los geo-objetos y de la consulta para recuperar la operación espacial

Con base en los resultados de la identificación de lugares y la obtención de la relación espacial obtenidos anteriormente para el *tweet*; “llueve en Avenida Insurgentes y La Raza extreme precauciones”. En esta sección, se detallará como se realiza la construcción de las consultas para obtener los geo-objetos y la operación espacial, estas consultas serán ejecutadas posteriormente en el *triple store*.

4.4.4.1 Construcción de la consulta para recuperar los geo-objetos

Esta consulta, no solo recuperará los objetos geográficos relacionados con los nombres identificados en el texto, sino que también servirá para desambiguarlos. Por ejemplo, en el caso del *tweet*, “llueve en Avenida Insurgentes y La Raza extreme precauciones”, el método de *NER* aplicado anteriormente, recuperó dos cadenas “Avenida Insurgentes” y “La Raza”, que fueron clasificados como *location*, estas cadenas serán la entrada de la consulta y si la salida es un objeto geográfico se considera que pertenece a un lugar, de lo contrario el resultado será vacío, lo que significa que la cadena no obtuvo correspondencia con ninguna ubicación.

En este caso, utilizamos dos de las ontologías de nuestra red, la ontología de infraestructura vial y POIs, que contiene objetos de tipo punto, puesto que todas las estaciones de transporte público que contiene se encuentran en esta representación, al igual que los puntos de interés. Por otro lado utilizamos la ontología enriquecida de las Calles de la Ciudad de México, que contiene segmentos de vialidad y son objetos de tipo línea.

Es necesario validar los nombres obtenidos en estas dos ontologías e identificar en donde se encuentran ubicados, para posteriormente hacer la consulta junto con la operación espacial. Como indica el siguiente algoritmo.

```
for(i=0;i<n;i++){
 if(query_infraestructuravialypois(obj[i])!=NULL){
      posicioni= query_infraestructuravialypois(obj[i]);
      URIi= query_infraestructuravialypois(obj[i]);}
 if (query_callesMexico(obj[i])!= NULL) {
      posicionc= query_callesMexico(obj[i]);
      URIc= query_callesMexico(obj[i]);}
 if ((posicioni && posicionc) == NULL){
      print ("No se tiene ningún lugar que corresponda con ese
 nombre");
      }
 }
```

Para todos los objetos encontrados en la etapa de *NER*, se realiza la consulta al *Endpoint*, primero se consulta a la ontología de infraestructura vial y *POIs*, si el objeto es válido, la consulta regresa la posición del objeto y la *URI*, que nos sirve para identificar la ontología que lo contiene, posteriormente se realiza la consulta apuntando a la ontología de las Calles de México y de igual forma se recupera su posición y la URI. Si el objeto no existe en ninguna de las dos ontologías, es descartado por que no se considera un lugar válido.

Por ejemplo, la consulta para el objeto "Avenida Insurgentes" en la ontología de infraestructura vial sería:

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://infraestructuravial.mx/ontologia/>
```

```
PREFIX geometria: <http:// infraestructuravial.mx /recurso/geometry/>
PREFIX recurso: <http:// infraestructuravial.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 geometria:hasGeometry ?aGeom.
  ?obj1 rdfs:label "AVENIDA INSURGENTES"@es.
  ?aGeom geo:asWKT ?pos1.
}
```

Pero en este caso, la salida es NULL por que el objeto "AVENIDA INSURGENTES", no se encuentra en las instancias de esta ontología, ahora se realiza la consulta a la ontología de Calles de México.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 geometria:hasGeometry ?aGeom.
  ?obj1 rdfs:label "AVENIDA INSURGENTES"@es.
  ?aGeom geo:asWKT ?pos1.

}
```

En este caso, la consulta retorna la posición de "AVENIDA INSURGENTES" y la URI que es: <http://datos.callesmexico.mx/ontologia/> para conocer a donde se hará la consulta con la operación espacial. Este proceso se realiza con todos los objetos obtenidos del NER, en este caso para el *tweet* de prueba, se debe hacer el mismo procedimiento para el objeto "LA RAZA". Si existiera un lugar que se encontrara en ambas ontologías, se almacena la posición y URI de los dos lugares para su posterior desambiguación.

4.4.4.2 Construcción de la consulta para recuperar la operación espacial

Para esta consulta se emplea la salida del proceso de la identificación de las relaciones espaciales, que fue:

```
“llueve en Avenida Insurgentes y La Raza extreme precauciones”
La relación es: en_y
La ER es: .*(en)+?.*(y)+?.*
```

Y se emplea la etiqueta asignada a la expresión regular que corresponde a la relación, en este caso es “en_y” para realizar la consulta, que preguntará al *Endpoint*, y en particular a la ontología de relaciones espaciales.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://SpatialOntology.ime.mx/ontologia/>
PREFIX recurso: <_http://SpatialOntology.ime.mx/recurso/>

SELECT * WHERE{
  ?s rdfs:label "en_y"@es.
  ?s ?o ?p.
  ?p ?q owl:Class.
 filter(?p != onto:spatialrelationships).
}
```

El resultado de esta consulta es:

```
RelacionEspacial “en_y”
GeoSparqlQuery “geof:sfIntersects”
ExpresionRegular “.*(en)+?.*(y)+?.*”
OperacionDB “ST_Intersects”
```

Con este resultado, ya se tiene la posición de lo objetos geográficos y la operación espacial que se debe realizar.

4.4.5 Consulta para recuperar las vialidades afectadas por el evento

Empleando los resultados de las consultas anteriores para el *tweet*, "llueve en Avenida Insurgentes y La Raza extreme precauciones" , se tiene que los dos objetos evaluados son válidos y tienen como URI "<http://datos.callesmexico.mx/ontologia/>", por otro lado, se identificó que la relación espacial es "en_y" y la operación correspondiente es "geof:sfIntersects", entonces la consulta quedaría estructurada de la siguiente manera:

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 ?obj2 WHERE{
  ?obj1 geometria:hasGeometry ?aGeom.
  ?obj1 rdfs:label "AVENIDA INSURGENTES"@es.
  ?aGeom geo:asWKT ?pos1.

  ?obj2 geometria:hasGeometry ?aGeom2.
  ?obj2 rdfs:label "LA RAZA"@es.
  ?aGeom2 geo:asWKT ?pos2.
  FILTER(geof:sfIntersects(?pos1,?pos2)).
}
```

El resultado de esta consulta son los segmentos de vialidad que cumplan con estas características. Esta consulta también valida que la relación obtenida sea válida para los objetos descritos en el texto, si el resultado de la consulta es NULL, significa que no es posible efectuar la operación espacial con los objetos encontrados.

4.5 Visualización y Evaluación

Finalmente, en esta etapa los segmentos de vialidad obtenidos son visualizados en un mapa, además de asignarles una etiqueta con el *tweet* original y un color dependiendo de su clasificación. Para este proceso empleamos como servidor geográfico *GeoServer*[12] y para visualización *OpenLayers*[13].

El mapa permite visualizar veinte *tweets* que recupera de la Base de Datos y va refrescando esos mensajes con los siguientes veinte *tweets* en la lista.

En esta etapa, también se realiza el proceso de evaluación, que considera los siguiente criterios:

1. Es seleccionado un conjunto de *tweets* para experimentos y son geocodificados por un usuario. Para este trabajo, los *tweets* geocodificados serán considerados nuestro *gold standard*.
2. El conjunto de datos seleccionados, se geocodifica empleando el método que proporciona *Google Maps*.
3. Por último, el conjunto de datos es evaluado con las medidas de *precision*, *recall* y medida *F* (Paradesi, 2011).

[12] http://geoserver.org/

[13] http://openlayers.org/

5. Experimentos y Resultados

En este capítulo, se describen los experimentos realizados para validar la metodología propuesta. Para realizar dichas pruebas se realizó un sistema Web que visualiza veinte *tweets* en un mapa y se actualiza cada 5 minutos con los siguiente veinte mensajes en cola.

5.1 Conjunto de Datos

Desde Noviembre del 2013 a Enero del 2014 se realizó la recuperación de *tweets* en español referentes a eventos viales ocurridos en la Ciudad de México, empleando el API de *Twitter* para python "*tweepy*", después de realizar la depuración de la información, el conjunto de datos final contiene 368,933 *tweets* generados por 57,382 usuarios distintos, de los cuales empleamos 100,000 mensajes para los experimentos aquí presentados.

5.2 Experimentos propuestos

En esta sección, se proponen cuatro experimentos, que demuestran el funcionamiento de la metodología con diferentes tipos de evento y de *tweet*. El primero, consiste en un *tweet* de la categoría de "Manifestaciones", pero que solo contiene un lugar en la descripción proporcionada en el mensaje. El segundo experimento, consiste de un mensaje de la categoría "Congestión vehicular" y una descripción que involucra a tres lugares. El tercer experimento, se enfoca en geocodificar un *tweet* de la categoría "Accidente vehicular" que involucra dos elementos en la descripción. Por último, el cuarto experimento describe un evento de la categoría "Obras públicas" con tres elementos involucrados.

En la Tabla 2, se colocan las características de los usuarios que publican los mensajes utilizados para las pruebas y su *Nivel de Confiabilidad del Usuario (NCU)*, que es calculado con las formulas presentadas en la sección de Metodología, de igual forma en la Tabla 3, se muestran las características de los mensajes incluyendo su clasificación, que se obtuvo utilizando *N-gram-based Text Categorization* y los diccionarios presentados en la sección de Estandarización y Clasificación de la Metodología para estandarizar los mensajes empleados.

Tabla 2. Características de los usuarios, tweets de prueba.

Id_usuario	Cuenta	Ubicación	Fecha creación	# seguidores	#amigos	#*tweets*	NCU
166594238	@OVIALCDMX	CIUDAD DE MEXICO	14-07-2010	763,372	205	462,795	0.85
205339755	@072AvialCDMX	México DF	20-10-2010	124,083	298	1,149,979	0.74
58847335	@ApoyoVial	México DF	21-07-2009	617,757	39,138	349,636	0.65
884604278	@alertiuxmxdf	México D.F.	16-10-2012	35,695	25,467	79,485	0.53

Tabla 3. Características de los tweets de prueba.

#	Mensaje	Usuario	RT	#RT	Clasificación	Fecha y hora
1	RT @TraficoReporte: Continuan manifestantes en Ciudad Universitaria: precaución.	@ApoyoVial	Si	@TraficoReporte	"Manifestaciones"	19-11-2013 16:55:03
2	#TránsitoLento sobre Viaducto Tlalpan desde Anillo Periférico hasta Av. Insurgentes.	@072AvialCDMX	No	----------	"Congestión vehicular"	01-12-2013 08:00:56
3	Accidente sobre Eje Central a la altura de Tacuba.	@alertiuxmxdf	Si	@072AvialCDMX	"Accidente vehicular"	25-11-2013 14:03:48
4	RT @OrientadorVial: Afectado un carril por #obras sobre Masaryk desde Arquímedes hasta Tennyson.	@OVIALCDMX	Si	@OrientadorVial	"Obras públicas"	09-12-2013 01:04:00

5.2.1 Experimento 1

Para este experimento, se considera el *tweet* número 1 de la Tabla 2, que contiene el texto "RT @TraficoReporte: Continuan manifestantes en Ciudad Universitaria: precaución..", primeramente el *tweet* es estandarizado, lo que significa que se remplaza RT y @TraficoReporte: por espacios en blanco, se eliminan los acentos y puntos y la primera letra se coloca en minúsculas. Posteriormente se ejecuta el *script* para obtener los nombres propios del texto, el resultado se muestra en la Figura 5.1:

```
1 continuan            continuan            VMIP3P0 VMI
2 manifestantes        manifestante         NCCP000 NC
3 en                   en                   SP      SP
4 Ciudad_Universitaria ciudad_universitaria NP00G00 NP
5 precaucion           precaucion           NCFS000 NC

pos=verb|type=main|mood=indicative|tense=present|person=3|num=plural
pos=noun|type=common|gen=common|num=plural
pos=adposition|type=preposition
pos=noun|type=proper|neclass=location
pos=noun|type=common|gen=feminine|num=singular
```

Figura 5.1. NER, Experimento1.

Del *script* se obtiene como entidad “Ciudad Universitaria”, es identificada como *location*, aunque para validar será desambiguada posteriormente.

Ahora, es necesario analizar el texto para recuperar la relación espacial. Para esto se usa el diccionario de Expresiones Regulares, buscando la expresión que describa el mensaje, cuando existe una coincidencia, entonces el *script* regresa; el texto, el nombre de la relación y su expresión regular.

```
“continuan manifestantes en Ciudad Universitaria precaucion'”
La relación es: en
La ER es: .*(\sen\s)+?.*
```

Con esta información se realiza la primera consulta al *Endpoint*, la cual debe retornar la *URI* y la posición que corresponde a las entidades identificadas en el proceso de *NER*. En este caso, la *URI* corresponde a la ontología de infraestructura vial y *POIs*.

```
PREFIX onto: <http://infraestructuravial.mx/ontologia/>
PREFIX geometria: <http:// infraestructuravial.mx /recurso/geometry/>
PREFIX recurso: <http:// infraestructuravial.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "CIUDAD UNIVERSITARIA"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
}
```

Posteriormente, se utiliza la etiqueta correspondiente a la relación espacial identificada, que en este caso es “en” y se realiza otra consulta al *Endpoint*, para recuperar la operación espacial que será aplicada.

```
PREFIX onto: <http://SpatialOntology.ime.mx/ontologia/>
PREFIX recurso: < http://SpatialOntology.ime.mx/recurso/>

SELECT * WHERE{
  ?s rdfs:label "en"@es.
  ?s ?o ?p.
  ?p ?q owl:Class.
 filter(?p != onto:spatialrelationships).
}
```

```
RelacionEspacial “en”
GeoSparqlQuery “geof:buffer”
ExpresionRegular “.*(\sen\s)+?.*”
OperacionDB “ST_Buffer”
```

Por último, se realiza la consulta para recuperar las vialidades afectadas por el evento, en el caso de un objeto tipo punto, como el del ejemplo, es necesario hacer un *buffer* para saber cuales vialidades son cercanas y sufren de afectación por el evento, para este trabajo se considera un *buffer* de 100 metros, puesto que solo nos interesa obtener las vialidades inmediatas.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
PREFIX invi: http://infraestructuravial.mx/ontologia/
PREFIX geometriai: <http:// infraestructuravial.mx /recurso/geometry/>
PREFIX recursoi: <http:// infraestructuravial.mx /recurso/>
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label " CIUDAD UNIVERSITARIA "@es.
  ?obj1 geometriai:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
BIND (geof:buffer(?pos1,mtr,100)AS ?bufferi)
```

```
?calles geometria:hasGeometry ?aGeom2.
  ?aGeom2 geo:asWKT ?pos2.
  FILTER(geof:sfContains (?bufferi,?pos2)).
}
```

Finalmente, en la Figura 5.2 se puede visualizar un marcador, de color verde, en el lugar mencionado en el *tweet* y con líneas azules las vialidades que se encuentran dentro del buffer y existen como instancias de la ontología.

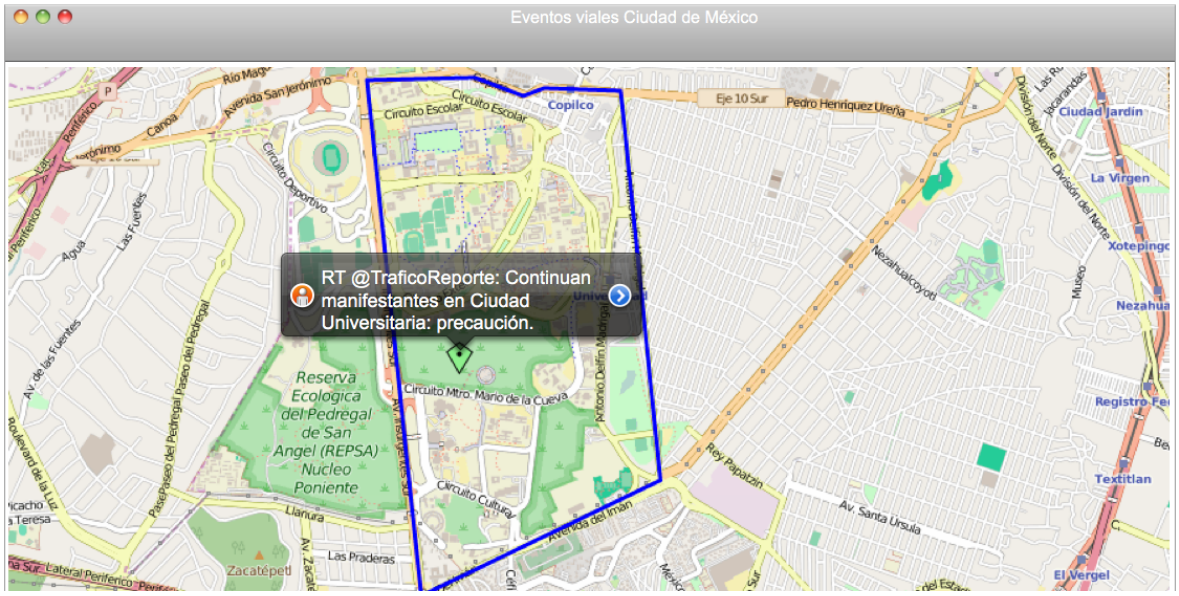

Figura 5.2. Visualización del *tweet* del Experimento1.

5.2.2 Experimento 2

Este experimento geocodifica el *tweet*: "#TránsitoLento sobre Viaducto Tlalpan desde Anillo Periférico hasta Av. Insurgentes.", en este caso después de pasar por la etapa de estandarización el *tweet* queda de la siguiente forma, para su procesamiento:

"transito lento sobre Viaducto Tlalpan desde Anillo Periferico hasta Avenida Insurgentes"

Como se puede observar, se remplazó el *hashtag* "#TránsitoLento" por "transito lento", empleando el diccionario de "*Hashtags*" y la palabra "Av." por "Avenida" empleando el diccionario de "Abreviaturas", generados en la segunda etapa de la metodología. Con el texto estandarizado, se ejecuta el script para obtener los nombres propios del *tweet.*

```
1 transito            transitar           VMIP1S0 VMI
2 lento               lento               AQ0MS00 AQ
3 sobre               sobre               SP      SP
4 Viaducto_Tlalpan    viaducto_tlalpan    NP00SP0 NP
5 desde               desde               SP      SP
6 Anillo_Periferico   anillo_periferico   NP00G00 NP
7 hasta               hasta               SP      SP
8 Avenida_Insurgentes avenida_insurgentes NP00G00 NP

pos=verb|type=main|mood=indicative|tense=present|person=1|num=singular
pos=adjective|type=qualificative|gen=masculine|num=singular
pos=adposition|type=preposition
pos=noun|type=proper|neclass=person
pos=adposition|type=preposition
pos=noun|type=proper|neclass=location
pos=adposition|type=preposition
pos=noun|type=proper|neclass=location
```

Figura 5.3. NER, Experimento2.

Como se observa en la Figura 5.3, se identifican los nombres “Viaducto Tlalpan”, “Anillo Periferico” y “Avenida Insurgentes”, el primero se clasifica como persona y los otros dos como lugares, para este trabajo no es de relevancia esta etiqueta, solo tener la entidad.

Posteriormente, se realiza la búsqueda de la relación espacial en el mensaje, en este ejemplo tenemos lo que nosotros consideramos una relación compuesta, por que contiene la raíz “sobre” pero está acompañada de “desde_hasta”, lo que significa que se realizarán más de una operación.

“transito lento sobre Viaducto Tlalpan desde Anillo Periferico hasta Avenida Insurgentes”

La relación es: sobre_desde_hasta

La ER es: .*(sobre)+?.*+ .*(desde)+?.*(hasta)+?.*

Se deben validar todos los nombres propios identificados en el *script* de *NER*, en este caso son tres elementos: “VIADUCTO TLALPAN”, “ANILLO PERIFERICO” y “AVENIDA INSURGENTES”, para obtener la URI y saber que tipo de objeto geográfico es y su posición.

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>
```

```
SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "VIADUCTO TLALPAN"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
}

PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "ANILLO PERIFERICO"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
}

PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "AVENIDA INSURGENTES"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
}
```

Ahora, se realiza la consulta para recuperar la operación espacial que debe realizar el *Endpoint* con los objetos geográficos.

```
PREFIX onto: <http://SpatialOntology.ime.mx/ontologia/>
PREFIX recurso: < http://SpatialOntology.ime.mx/recurso/>

SELECT * WHERE{
  ?s rdfs:label "sobre_desde_hasta"@es.
  ?s ?o ?p.
  ?p ?q owl:Class.
 filter(?p != onto:spatialrelationships).
}
```

```
RelacionEspacial “sobre_desde_hasta”
GeoSparqlQuery “geof:sfIntersects”
ExpresionRegular “.*(sobre)+?.*+ .*(desde)+?.*(hasta)+?.*”
OperacionDB “ST_Intersects”
```

Finalmente, se ejecuta la consulta para obtener los segmentos de vialidad afectados por el evento, en esta consulta primero se obtiene la posición de los tres elementos y posteriormente se evalúa la posición del objeto 1 contra la del objeto 2, y la del objeto 1 con la del objeto 3, para al final obtener el segmento de vialidades afectadas realmente, que se pueden visualizar en la Figura 5.4.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?p1 ?p2 ?p3 WHERE{
  ?p1 rdfs:label " VIADUCTO TLALPAN "@es.
  ?p1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

  ?p2 rdfs:label " ANILLO PERIFERICO "@es.
  ?p2 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos2.

  ?p3 rdfs:label " AVENIDA INSURGENTES "@es.
  ?p3 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos3.

  geof:sfIntersects(?pos1,?pos2)AS ?punto1)
  geof:sfIntersects(?pos1,?pos3)AS ?punto2)
  FILTER(geof:sfIntersects(?punto1,?punto2))
}
```

Figura 5.4. Visualización del *tweet* del Experimento2.

5.2.3 Experimento 3

Para este experimento empleamos el *tweet*: “Accidente sobre Eje Central a la altura de Tacuba.”, que pertenece a la categoría de “Accidente vehicular”, el proceso que se realiza es el mismo que en los experimentos anteriores, la única diferencia radica, en la cantidad de objetos involucrados, el tipo de relación recuperada y la operación asociada a la misma.

```
1 accidente   accidente   NCMS000 NC pos=noun|type=common|gen=masculine|num=singular
2 sobre       sobre       SP      SP pos=adposition|type=preposition
3 Eje_Central eje_central NP00SP0 NP pos=noun|type=proper|neclass=person
4 a           a           SP      SP pos=adposition|type=preposition
5 la          el          DA0FS0  DA pos=determiner|type=article|gen=feminine|num=singular
6 altura      altura      NCFS000 NC pos=noun|type=common|gen=feminine|num=singular
7 de          de          SP      SP pos=adposition|type=preposition
8 Tacuba      tacuba      NP00G00 NP pos=noun|type=proper|neclass=location
```

Figura 5.5. NER, Experimento3.

Como se observa en la Figura 5.5, se identifican dos nombres propios del texto, primero “Eje Central” como *person* y “Tacuba” como *location*, que son evaluados al momento de hacer la consulta. Primero, se valida si son un objeto geográfico válido, puesto que si la consulta regresa el valor de NULL, al preguntar a la ontología de infraestructura vial y a la de Calles de México, el nombre propio recuperado no pertenece a una ubicación y es descartado.

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "EJE CENTRAL"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.
}
```

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "TACUBA"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

}
```

Ahora, se recupera la relación espacial, junto con la expresión regular que la representa y se procede a realizar la consulta para obtener la o las operaciones espaciales ligadas a dicha expresión.

```
“accidente sobre Eje Central a la altura de Tacuba”
La relación es: sobre_a la altura
La ER es: .*(sobre)+?.*+ .*(a la altura)+?.*
```

```
PREFIX onto: <http://SpatialOntology.ime.mx/ontologia/>
PREFIX recurso: < http://SpatialOntology.ime.mx/recurso/>

SELECT * WHERE{
  ?s rdfs:label "sobre_a la altura"@es.
  ?s ?o ?p.
  ?p ?q owl:Class.
 filter(?p != onto:spatialrelationships).
}

RelacionEspacial “sobre_a la altura”
GeoSparqlQuery “geof:sfIntersects,geof:sfTouches”
ExpresionRegular “.*(sobre)+?.*+ .*(a la altura)+?.*”
OperacionDB “ST_Intersects, ST_Touches”
```

En este caso, se observa que se recuperan dos operaciones, “geof:sfIntersects,geof:sfTouches”, que son aplicadas en la consulta para recuperar las vialidades afectadas, no en todos los casos es necesario aplicar todas las operaciones que regresa la consulta, solamente si los objetos geográficos lo permiten. El resultado de la consulta es visualizado en la Figura 5.6.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?p1 ?p2 WHERE{
  ?p1 rdfs:label " EJE CENTRAL "@es.
  ?p1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

  ?p2 rdfs:label " TACUBA "@es.
  ?p2 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos2.

  geof:sfIntersects(?pos1,?pos2)AS ?punto1)
  FILTER(geof:sfTouches(?p1,?punto1))
}
```

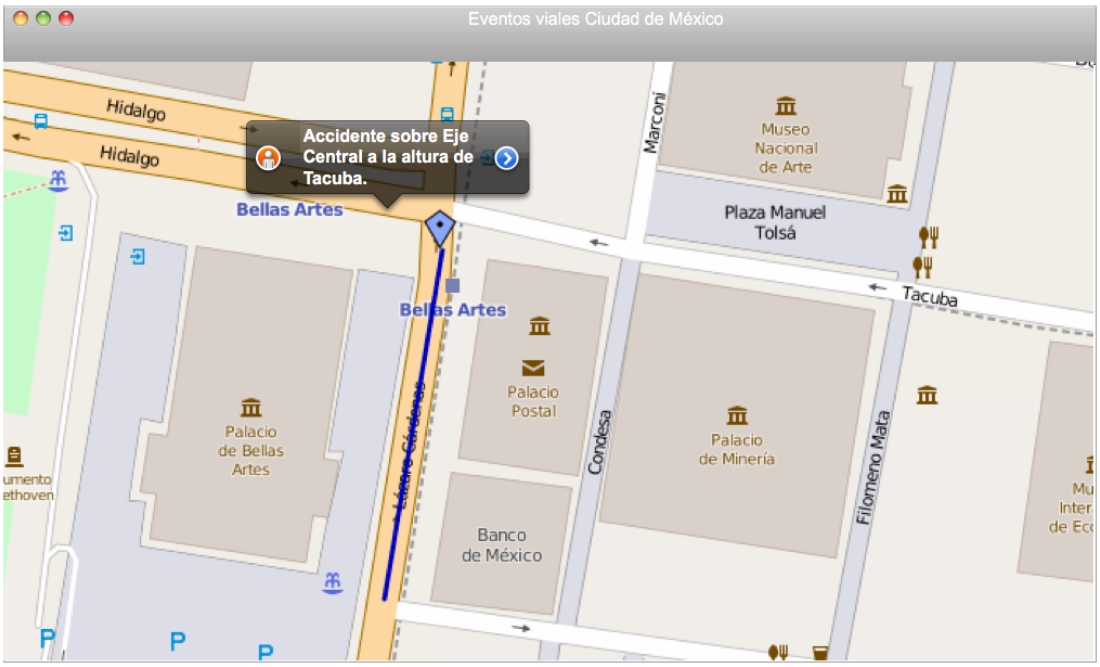

Figura 5.6. Visualización del *tweet* del Experimento3.

5.2.4 Experimento 4

Por último, tenemos una variación de la relación "sobre_desde_hasta", en este *tweet* perteneciente a la categoría de "Obras públicas" y representado por un marcador en color amarillo en la Figura 5.8.

"RT @OrientadorVial: Afectado un carril por #obras sobre Masaryk desde Arquímedes hasta Tennyson."

```
1  afectado    afectar     VMP00SM VMP pos=verb|type=main|mood=participle|num=singular|gen=masculine
2  un          uno         DI0MS0  DI  pos=determiner|type=indefinite|gen=masculine|num=singular
3  carril      carril      NCMS000 NC  pos=noun|type=common|gen=masculine|num=singular
4  por         por         SP      SP  pos=adposition|type=preposition
5  obras       obra        NCFP000 NC  pos=noun|type=common|gen=feminine|num=plural
6  sobre       sobre       SP      SP  pos=adposition|type=preposition
7  Masaryk     masaryk     NP00SP0 NP  pos=noun|type=proper|neclass=person
8  desde       desde       SP      SP  pos=adposition|type=preposition
9  Arquimedes  arquimedes  NP00G00 NP  pos=noun|type=proper|neclass=location
10 hasta       hasta       SP      SP  pos=adposition|type=preposition
11 Tennyson    tennyson    NP00SP0 NP  pos=noun|type=proper|neclass=person
```

Figura 5.7. NER, Experimento4.

Después de generar el árbol sintáctico de dependencias e identificar los nombres propios, se obtienen tres posibles lugares, es importante recordar que hasta que no se efectúe la consulta que recupera la posición del nombre propio, no se puede afirmar que se trate o no de un lugar válido, “Masaryk”, “Arquemedes”, “Tennyson”.

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "MASARYK"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

}
```

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "ARQUIMEDES"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

}
```

```
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>

SELECT DISTINCT ?obj1 WHERE{
  ?obj1 rdfs:label "TENNYSON"@es.
  ?obj1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

}
```

Puesto que las tres consultas regresaron las posición de los objetos y su respectiva URI, entonces se ejecuta el *script* para la obtención de la relación espacial.

```
“afectado un carril por obras sobre Masaryk desde Arquimedes hasta Tennyson”
```

```
La relacion es: sobre_desde_hasta
La ER es:.*(sobre)+?.*+.*(\sdesde\s)+?.*(hasta)+?.*

PREFIX onto: <http://SpatialOntology.ime.mx/ontologia/>
PREFIX recurso: < http://SpatialOntology.ime.mx/recurso/>

SELECT * WHERE{
  ?s rdfs:label "sobre_desde_hasta"@es.
  ?s ?o ?p.
  ?p ?q owl:Class.
 filter(?p != onto:spatialrelationships).
}

RelacionEspacial “sobre_desde_hasta”
GeoSparqlQuery “geof:sfIntersects, geof:sfCrosses, geof:sfTouches”
ExpresionRegular “.*(sobre)+?.*+.*(\sdesde\s)+?.*(hasta)+?.*”
OperacionDB “ST_Intersects, ST_Crosses, ST_Touches”
```

La relación identificada es “sobre_desde_hasta”, y las operaciones espaciales que tiene asociadas son: “geof:sfIntersects, geof:sfCrosses, geof:sfTouches”, esto significa que dependiendo del tipo de dato de los objetos geográficos es la o las operaciones que se aplican en la última consulta, en este caso se emplea la operación “geof:sfIntersects” como en el caso del Experimento 2. El segmento de vialidad afectada es visualizado en la Figura 5.7, además del *tweet* que representa.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX gml: <http://www.opengis.net/ont/gml#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX par: <http://parliament.semwebcentral.org/parliament#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
PREFIX onto: <http://datos.callesmexico.mx/ontologia/>
PREFIX geometria: <http:// datos.callesmexico.mx /recurso/geometry/>
PREFIX recurso: <http:// datos.callesmexico.mx /recurso/>
```

```
SELECT DISTINCT ?p1 ?p2 ?p3 WHERE{
  ?p1 rdfs:label " MASARYK "@es.
  ?p1 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos1.

  ?p2 rdfs:label " ARQUIMEDES "@es.
  ?p2 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos2.

  ?p3 rdfs:label " TENNYSON "@es.
  ?p3 geometria:hasGeometry ?aGeom.
  ?aGeom geo:asWKT ?pos3.

  geof:sfIntersects(?pos1,?pos2)AS ?punto1)
  geof:sfIntersects(?pos1,?pos3)AS ?punto2)
  FILTER(geof:sfIntersects(?punto1,?punto2))
}
```

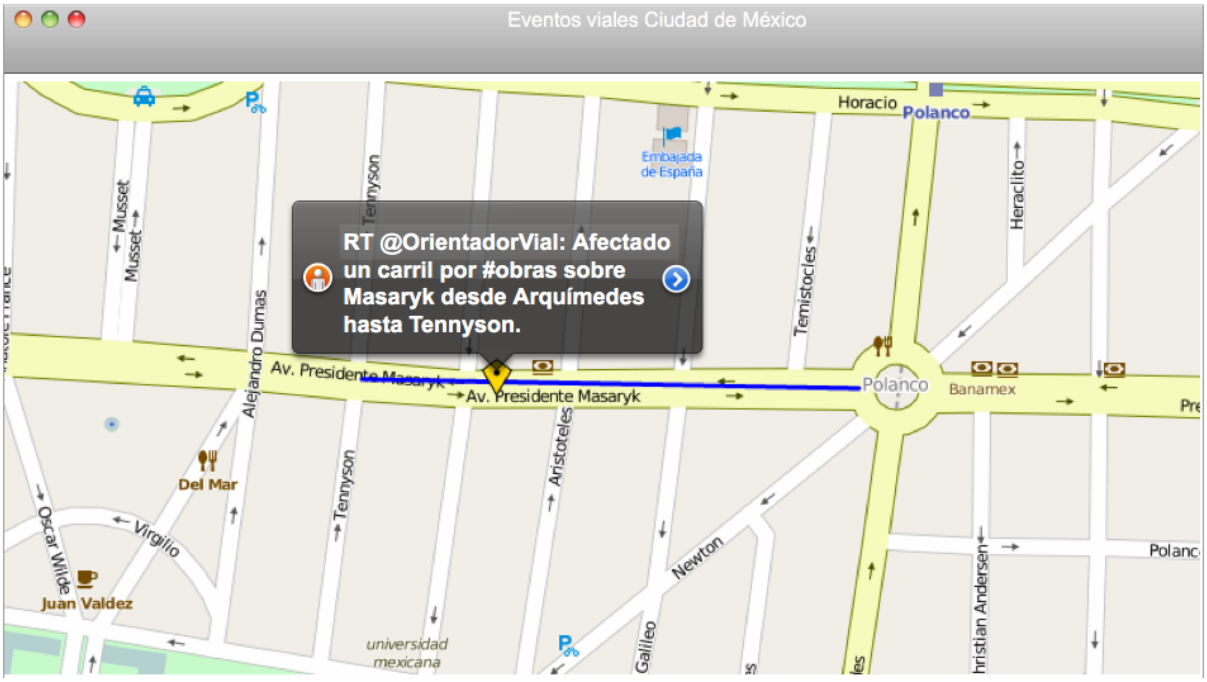

Figura 5.8. Visualización del *tweet* del Experimento4.

5.3 Análisis de Resultados

En esta sección, es presentado el análisis de los resultados obtenidos del conjunto de *tweet* seleccionados para los experimentos. Primeramente son comparados con nuestro *gold standard*, que son las geocodificaciones realizadas por un usuario, posteriormente se comparan con el método de geocodificación proporcionado por Google y finalmente se obtienen las medidas de *Precision*, *Recall* y la medida *F*.

Se consideraron dos características importantes para evaluar que la geocodificación del evento es correcta. Primero, que se hayan identificado correctamente los objetos geográficos contenidos en el *tweet*, estos elementos pueden ser vialidades, estaciones de transporte público o puntos de interés y segundo, que los tramos de vialidad afectados por el evento, sean identificados correctamente en el mapa.

Cuando el sistema, identifica todos los elementos del *tweet*, se considera un acierto. Cuando el sistema identifica alguno de los elementos del *tweet*, se considera un acierto parcial. Los errores ocurren cuando el sistema no identifica ningún elemento del *tweet*. De esta forma, las medidas de *precision* y *recall* , emplean verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, como se muestra en la Ecuación 4.

$$P = \frac{V_p}{V_p + F_p} \; ; R = \frac{V_p}{V_p + F_N} ; mF = 2\frac{P \times R}{P + R} \tag{4}$$

Los verdaderos positivos V_p , son los elementos geográficos que fueron encontrados por la metodología y pertenecen al *gold standard*, los falsos positivos F_p, son elementos geográficos que fueron encontrados por la metodología pero que no pertenecen al *gold standard* y los falsos negativos F_N son aquellos elementos geográficos que pertenecen al *gold standard*, pero no fueron encontrados por la metodología. Las medidas de *precision*, *recall* y medida *F*, fueron calculadas para cada *tweet* y se obtuvo el promedio de cada caso. Los resultados se muestran en la Tabla 4.

Tabla 4. Precision, Recall y medida F de la metodología.

Elementos geográficos	***Precision***	***Recall***	***mF***
Vialidades	0.878	0.870	0.872
POIs	0.880	0.784	0.829
Estaciones de transporte	0.845	0.734	0.785
Eventos	0.945	0.956	0.960

De igual forma en la Tabla 5, se obtienen estas medidas realizando una comparación con el servicio de geocodificación de *Google* contra el nuestro, aunque para estas pruebas se utilizó el *tweet* ya estandarizado en el servicio de *Google*, para obtener resultados diferentes de 0

Tabla 5. Precision, Recall y medida F de la metodología vs Google.

Elementos geográficos	***Precision***		***Recall***		***mF***	
	Google	**Nuestro método**	**Google**	**Nuestro método**	**Google**	**Nuestro método**
Vialidades	0.820	0.874	0.560	0.866	0.665	0.872
POIs	0.720	0.889	0.615	0.784	0.663	0.829
Estaciones de transporte	0.730	0.890	0.527	0.734	0.612	0.785
Eventos	0.10	0.945	0.02	0.956	0.039	0.960

6. Conclusiones y Trabajo futuro

En este trabajo, se propone un enfoque novedoso para geocodificar eventos viales recopilados del *streaming* de *Twitter,* usando una red de ontologías y las relaciones espaciales contenidas en el mensaje, visualizando en un mapa los eventos y las vialidades afectadas.

En el desarrollo de la metodología se implementaron dos ontologías, la primera es la de relaciones espaciales que se emplea para obtener la relación entre una descripción en lenguaje natural y una operación espacial. Y la segunda ontología contiene la infraestructura vial y los Puntos de Interés de la Ciudad de México, adicionalmente se empleó una ontología de Calles de la Ciudad México del trabajo de Rivera et al. (2015) con estas tres ontologías se genera una red ontoló.

Es importante, resaltar el uso de la red de ontologías para la realización de la desambiguación de las entidades recuperadas, mediante el uso de *Named Entity Recognition* y la validación de la operación espacial con los objetos geográficos, contribuyendo a la disminución de Falsos Positivos.

También consideramos como un aporte importante, el diccionario de Expresiones Regulares generado, que representan las Relaciones Espaciales comúnmente usadas en el idioma español y de la misma forma los diccionarios para estandarización que incluyen “Abreviaturas”,” Acrónimos”,”*Hashtags*” y ”*Nicknames*” que se emplean frecuentemente en los mensajes relacionados con eventos viales.

Adicionalmente, se generan cuatro diccionarios más que contienen expresiones recolectadas de nuestro corpus, clasificadas en cada uno de los diccionarios con las temáticas de “Manifestaciones”,”Accidente vehicular”,”Congestión vehicular ” y ”Obras públicas”, estos diccionarios también se emplearon en nuestro *script* de clasificación que hacen más fácil la visualización en un mapa al usuario final.

Con respecto a los resultados, al no considerar los *tweets* de usuarios poco confiables, se proporciona cierto nivel de certeza a los resultados proporcionados por la geocodificación, esto se logra empleando la medida *NCU* propuesta en el trabajo que obtiene el *Nivel de Confiabilidad del Usuario*.

Estos resultados son evaluados empleando las medidas de *precision*, *recall* y la medida F, además de compararlo con la geocodificación realizada por un usuario manualmente y con una geocodificación empleando el servicio de *Google*, obteniendo buenos resultados cuando se realiza la comparación con el método propuesto en este trabajo. Obteniendo una *precision* de 0.945 en la geocodificación de eventos, un *recall* de 0.956 y una medida *F* de 0.960 contra una *precisión* de 0.10, un *recall* de 0.02 y una medida *F* de 0.039 del método de geocodificación de *Google*, obteniendo medidas similares cuando se compara con la geocodificación de un usuario que se considera nuestro *gold standard.*

Como trabajo futuro, sería interesante el poder enriquecer la red de ontologías con fuentes externas de información como DBPedia o alguna otra fuente, esto incrementaría las posibilidades de encontrar los lugares mencionados en los *tweets* y descartar la menor cantidad posible. Otra aportación interesante sería el agregar información de otras redes sociales para enriquecer los datos y la implementación de esta metodología en una aplicación móvil.

7. Referencias

Agarwal, P., Vaithiyanathan, R., Sharma, S., & Shroff, G. (2012). Catching the Long-Tail: Extracting Local News Events from Twitter. In *ICWSM*.

Albuquerque, F. C., Casanova, M. A., Lopes, H., Redlich, L. R., de Macedo, J. A. F., Lemos, M., ... & Renso, C. (2015). A methodology for traffic-related Twitter messages interpretation. *Computers in Industry*.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, *31*(1), 132-164.

Allocca, C.; D'Aquin, M.; Motta, E., (2009) *DOOR - Towards a Formalization of Ontology Relations*. In Jan L. G. Dietz, editor, KEOD, (pp. 13–20)

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Bouillot, F.; Poncelet, P. & Roche, M. (2012): How and why exploit tweet's location information? In: J. Gensel; D. Josselin & D. Vandenbroucke (Hrsg.), *Proceedings of the AGILE'2012 International Conference on Geographic Information Science,Avignon, April, 24-27*. Avignon, France.

Boyd, d. & Ellison, N. (2008): Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13: 210–230.

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1-10). IEEE.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, *8*(12), e83672.

Buriano, L.; Marchetti, M.; Carmagnola, F.; Cena, F.; Gena, C.; Torre, I., (10-12 May 2006) *The Role of Ontologies in Context-Aware Recommender Systems*, Mobile Data Management, 2006. MDM 2006. 7th International Conference on (pp.80).

Candillier, L.; Meyer, F.; Boullé, M.; (2007) *Comparing state-of-the-art collaborative filtering systems*, Lecture Notes in Computer Sciece, 4571, (pp. 548–562).

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor MI*, *48113*(2), 161-175.

Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759-768). ACM.

Corcho, Ó.; Fernández, M.; Gómez, A.; López, A., (2005) *Construcción de ontologías legales con la metodología METHONTOLOGY y la herramienta WebODE*.

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language resources and evaluation*,*47*(1), 217-238.

Davis Jr, C. A., Pappa, G. L., de Oliveira, D. R. R., & de L Arcanjo, F. (2011). Inferring the location of Twitter messages based on user relationships. *Transactions in GIS*, *15*(6), 735-751.

Egenhofer, M. J., & Herring, J. (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases. *The*, *9*, 94-1.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010, October). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1277-1287). Association for Computational Linguistics.

Elwood, S.; Goodchild, M. & Sui, D. (2011): Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*.

Fernández, M.; Gómez, A.; Juristo, N., (1997) *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*, AAAI Symposium on Ontological Engineering, Stanford.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In*Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363-370). Association for Computational Linguistics.

Gelernter, J., & Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, *15*(6), 753-773

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, 1-33.

Girres, J.F. & Touya, G. (2010): Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4): 435–459.

Goodchild, M. F. (2009). Geographic information systems and science: today and tomorrow. *Annals of GIS*, *15*(1), 3-9.

Gordon, E. & de Souza e Silva, A. (2011): *NetLocality. Why Location Matters in a Networked World*. Wiley-Blackwell, West Sussex, UK.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568-578.

Grishman, R., & Sundheim, B. (1996). Message understanding conference – 6: A brief history. In Proceedings of the 16th conference on computational linguistics (COLING'96), Copenhagen, Denmark.

Gruber, T., (1995) *Towards principles for the design of ontologies used for knowledge sharing*, International Journal of Human-Computer Studies, 43(5/6), (pp. 907-928).

Guarino, N., (1995) *Formal ontology, conceptual analysis and knowledge representation*, International Journal of Human-Computer Studies, 43(5- 6), (pp. 625-640).

Gutierrez, C., Figuerias, P., Oliveira, P., Costa, R., & Jardim-Goncalves, R. (2015, July). Twitter mining for traffic events detection. In *Science and Information Conference (SAI), 2015* (pp. 371-378). IEEE.

Guzmán, J. A.; López, M.; Torres, I. D., (Mayo 2012) *Metodologías y métodos para la construcción de ontologías*. Scientia et Technica, [S.l.], 2(50), (pp. 133-140).

Haklay, M. (2010): How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37: 682–703.

Hart, T. C., & Zandbergen, P. A. (2013). Reference data and geocoding quality: Examining completeness and positional accuracy of street geocoded crime incidents. *Policing: An International Journal of Police Strategies & Management*,*36*(2), 263-294.

Honey, C., & Herring, S. C. (2009, January). Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on* (pp. 1-10). IEEE.

Jung, J. J. (2011, July). Towards named entity recognition method for microtexts in online social networks: a case study of Twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on* (pp. 563-564). IEEE.

Karger, D. R., & Quan, D. (2005). What would it mean to blog on the semantic web?. *Web Semantics: Science, Services and Agents on the World Wide Web*, *3*(2), 147-157.

Kinsella, S., Murdock, V., & O'Hare, N. (2011, October). I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (pp. 61-68). ACM.

Laurini, R. (2012, June). Importance of spatial relationships for geographic ontologies. In *Seventh International Conference on Informatics and Urban and Regional Planning INPUT* (pp. 122-134).

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks* (pp. 1-10). ACM.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5).

Liu, X., Zhou, M., Wei, F., Fu, Z., & Zhou, X. (2012, July). Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 526-535). Association for Computational Linguistics.

Liu, Y., Piyawongwisal, P., Handa, S., Yu, L., Xu, Y., & Samuel, A. (2012, December). Going beyond citizen data collection with mapster: a mobile+ cloud real-time citizen science experiment. In *e-Science Workshops (eScienceW), 2012 IEEE Seventh International Conference on* (pp. 1-6). IEEE.

Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, Mauro Gaio. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. ACM. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014), Nov 2014, Dallas, Texas, United States. Proceedings of the 22th ACM SIGSPATIAL International Conference on Advances in Geographic Information.

Neches, R., Fikes, R.E., Finin, T., Gruber, T.R., Patil, R., Senator, T. & Swartouy, W.R., (1991) *Enabling technology for knowledge sharing*, AI Magazine, 12(3), (pp. 16-36).

Neis, P.; Zielstra, D. & Zipf, A. (2012): The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(1): 1–21.

Nielsen, S. S., Nielsen, G. B., Denwood, M. J., Haugegaard, J., & Houe, H. (2015). Comparison of recording of pericarditis and lung disorders at routine meat inspection with findings at systematic health monitoring in Danish finisher pigs. *Acta Veterinaria Scandinavica*, *57*(1), 1.

O'Reilly, T. (2005): What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. URL http://oreilly.com/web2/archive/what-is-web-20.html.

Oussalah, M., Bhat, F., Challis, K., & Schnier, T. (2013). A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems*, *37*, 105-120.

Paradesi, S. M. (2011). Geotagging Tweets Using Their Content. In *FLAIRS Conference*.

Park, Y., & Byrd, R. J. (2001, June). Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing* (pp. 126-133).

Ponte, J. M., & Croft, W. B. (1998, August). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281). ACM.

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *KR*, *92*, 165-176.

Ritter, A., Etzioni, O., & Clark, S. (2012, August). Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104-1112). ACM.

Rivera, L. C., Vilches-Blázquez, L. M., Torres-Ruiz, M., & Ibarra, M. A. M. (2015). Semantic Recommender System for Touristic Context Based on Linked Data. In *Information Fusion and Geographic Information Systems (IF&GIS'2015)* (pp. 77-89). Springer International Publishing.

Robinson, B., Power, R., & Cameron, M. (2013, May). A sensitive Twitter earthquake detector. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 999-1002). International World Wide Web Conferences Steering Committee.

Roche, C., (2003) *Ontology : a survey*, in: 8th Symposium on Automated Systems Based on Human Skill and Knowledge, (pp. 28–41), Gteborg, Sweden.

Roick, O., & Heuser, S. (2013). Location Based Social Networks–Definition, Current State of the Art and Research Agenda. *Transactions in GIS*.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012, July). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1500-1510). Association for Computational Linguistics.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*, *67*(2), 214-237.

Seol, J. W., Jeong, K. Y., & Lee, K. S. (2013). Follower classification through social network analysis in twitter. In *Grid and Pervasive Computing* (pp. 926-931). Springer Berlin Heidelberg.

Suárez, M. C.; Gómez, A.; Fernández, M., (2012). *The NeOn Methodology for Ontology Engineering*. In Ontology Engineering in a Networked World, (pp. 9-34), Springer Berlin Heidelberg.

Sui, D. & Goodchild, M. (2011): The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11): 1737–1748.

Sure,Y.; Staab, S.; Studer.R., (2003*) On-to-knowledge methodology*, Handbook on Ontologies, Series on Handbooks in Information Systems, 6, (pp. 117-132).

Turner, A. (2006): *Introduction to Neogeography*. O'Reilly Media.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73-81.

Twitter (2013b). Twitter Help Center: FAQs about Tweet location, at http://support.Twitter.com/articles/78525-faqs-about-tweet-location, accesed 16 January 2013.

Twitter (2013c). Twitter Help Center: Adding your location to a Tweet, at http://support.Twitter.com/articles/122236-how-to-tweet-with-your-location, accesed 16 January 2013.

Twitter (2009). Twitter Blog: Location, location, location (20 August), at http://blog.Twitter.com/2009/08/location-location-location.html , accesed 16 January 2013.

Uschold, M.; Gruninger,M., (1996) *Ontologies: principles, methods and applications*. The Knowledge Engineering Review, 11, (pp. 93-136).

Van Exel, M.; Dias, E. & Fruijtier, S. (2010): The impact of crowdsourcing on spatial data quality indicators. In: *GIScience 2010*. Zurich, Switzerland.

Van Oort, P. (2006): *Spatial data quality: from description to application*. Dissertation,
Wageningen Universiteit.

Weller, M. (2012). The Digital Scholar: How technology is transforming academic practice. londres: Bloomsbury Academic.

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011, October). Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2541-2544). ACM.

Yuan, Y. (2011, June). Extracting spatial relations from document for geographic information retrieval. In *Geoinformatics, 2011 19th International Conference on*(pp. 1-5). IEEE.

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science*, *2014*(9), 37-70