

EXPLORANDO INICIATIVAS PARA LA GEORREFERENCIACIÓN AUTOMÁTICA DE TEXTOS DE LA WEB. DE LO TEXTUAL A LO ESPACIAL

Exploring initiatives for Automatic Geo-referencing of web text.
From textual to spatial

Andrés Mauricio Mesa y Luz Ángela Rocha

Universidad Distrital Francisco José de Caldas, Bogotá – Colombia

Correspondencia: andres.mesa@meridiangroups.com; lrocha@udistrital.edu.co

Recibido: 15 de mayo de 2012 Aceptado: 24 de agosto de 2012

Resumen

El presente artículo es el resultado de la etapa de diagnóstico de la investigación, en la que se exploran iniciativas de extracción de información de textos en la web posibles de ser espacializados, mostrando los principales ejemplos implementados y los componentes que hacen viable su funcionamiento, realizando un contraste entre dichas iniciativas a fin de distinguir los componentes más importantes. Igualmente, se examinan los elementos sustanciales a tener en cuenta para el proceso de recuperación de información espacial y georreferenciación automática, mostrando algunas técnicas que se pueden implementar en cada uno de estos aspectos característicos de la extracción de datos espaciales a partir de documentos web.

Palabras clave: recuperación de información espacial, ambigüedad espacial, georreferenciación automática.

Abstract

This article is the result of the first step of the research which explores initiatives for extracting information from text on the web, possible to be spatialized, showing the existing examples and the main components that make it a viable operation, creating a contrast between these initiatives in order to distinguish the most important components. Also examines the significant elements to be considered in the process of spatial information retrieval and automatic geo-referencing, showing some techniques that can be implemented in each of these aspects of spatial data extraction from web documents.

Keywords: spatial data retrieval, spatial ambiguity, automatic geo-referencing.

Introducción

Antes de pensar en cómo convertir referencias espaciales expresadas en textos de la web en información georreferenciada es necesario preguntarse por qué hacerlo y para qué utilizar capacidad computacional, sofisticados algoritmos de reconocimiento de textos y bases de datos espaciales con el fin de espacializar información textual automáti-

camente, y si aparentemente consideramos que los datos georreferenciados o espaciales o no brindan una utilidad comparada con la utilidad de la información no espacial.

La respuesta para algunos no parece ser tan obvia, por lo cual es necesario mostrar ciertos puntos de vista del porqué son necesarios los datos espaciales o georreferenciados. El primer argumento a exponer lo brinda Tobler. Recorde-

mos la primera ley de la geografía que nos brinda una de las más notables diferencias de datos espaciales contra los datos no espaciales:

Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes[1].

Claramente la cercanía espacial es una propiedad que solo puede analizarse a través de métricas [7] conceptualizadas en el espacio; por el contrario, resultaría imposible realizar un análisis espacial a partir de información textual (aunque en el análisis de cadenas y textos sí es posible establecer distancias y cercanías como la de Levenshtein [8], solo que estas distancias nada tienen que ver con lo geográfico).

Visto desde un enfoque algo filosófico del porqué la información espacial posee propiedades diferentes de la información textual, tenemos las observaciones de Kovouras [9] donde controvierte el enfoque generalista que degrada los conceptos geográficos y propone que no son diferentes a otros conceptos y que no hay nada especial acerca de lo geoespacial. Por el contrario, Kovouras resalta el hecho de que la mayor parte de la realidad tiene una referencia geográfica geoespacial, por lo menos lo físico (vehículos, personas, edificios, municipios) y lo no físico, como los conceptos que adquieren validez, solo bajo un marco de referencia espacial (por ejemplo, la carga de tráfico, el ingreso promedio, la densidad de población, etc.), siendo en este mundo casi todo geográfico y espacializable.

Referencias espaciales en Internet

Una vez establecida la importancia de generar información georreferenciada es necesario preguntarse: ¿cuánta información de la web es potencialmente espacial? Varios autores han tratado el tema de encontrar referencias espaciales en textos de la web [5] y se ha estimado que existe una fracción de las páginas disponibles en la web donde hay un contexto geográfico reconocible. McCurley encontró en 2001 que aproximadamente un 4,5% contienen referencias válidas de direcciones (*Zip codes*) en Estados Unidos; 8,5% contienen número telefónico (al cual se le puede asociar una posición) y un 9,5% de páginas que pueden tener al menos una de estas, sin contar las referencias espaciales expresadas por medio de topónimos, como nombres de ciudades, vías, plazas parques, etc. En resumen, dichas cifras

revelan un potencial para extraer información espacial de textos en la web.

Ahora bien, conociendo la gran cantidad de referencias espaciales en la web, nace la pregunta de ¿cómo es posible georreferenciar datos a partir de información textual? En los siguientes ítems se presentan algunas de las fuentes de información espacial para extraer coordenadas a partir de textos de la web [10].

Una de las principales fuentes de información espacial son los diferentes tipos de textos que hacen relación a una ubicación geográfica y que se encuentran en párrafos de la web, en ellos existen técnicas (ver [5]) con los que se explora la existencia de patrones [11] como direcciones, números telefónicos (a los cuales es posible asignarles posición), topónimos y puntos de interés, en los textos y párrafos de los recursos web. Sin embargo, el problema es la estandarización de esta información y la ambigüedad que se presenta al interpretar algunas palabras.

Asimismo, es posible asociar la localización de una página web (ejemplo página de contacto), si se asume que la localización es la misma y los datos de contacto son concordantes con la localización geográfica donde está publicada la página.

Como lo señaló Buyukokkten [12], un enfoque aceptado para determinar el contexto geográfico de un recurso web, es enfocarse en la localización del servicio de *hosting* de dicho recurso web. Este análisis es posible revisando el registro del dominio de Internet para dicha página web. Sin embargo, este enfoque tiene varios problemas; la heurística no funciona en todos los casos, ya que muchos sitios web son publicados en servidores en diferente país. El segundo problema con respecto a los dominios web es que no siempre corresponden con la localización de donde son publicados.

Alternativamente Buyukokkten [12] trabajó en la identificación y extracción de referencias espaciales explorando los *links* relacionados con las páginas para encontrar las referencias de los sitios web que contienen direcciones y códigos postales.

Otro recurso importante para recuperar información relacionada con la posición geográfica del usuario son las búsquedas locales en los motores como Google, Yahoo,

MSN, donde la información de contexto [7] como la localización obtenida a partir de la dirección IP del usuario permite asistir las búsquedas limitándolas a resultados solo locales, permitiendo que los resultados sean más cercanos al usuario. Varios autores [13] han estudiado diferentes algoritmos de búsqueda local comparando su efectividad, precisión y rendimiento.

Por otra parte, Wang mostró que muchas consultas realizadas por usuarios en la web son implícitamente espaciales; la palabra que acompaña una referencia espacial, como *cinemas*, *pizza*, *restaurante*, hacen que sean servidos por una búsqueda espacial [14], en vez de una búsqueda normal; esta estrategia es bastante usada, tanto que el uso de búsquedas espaciales tiende a ser muy frecuente. Sanderson [15], usando un análisis de los *logs* de las consultas de buscadores, encontró que alrededor de un 20% de búsquedas realizadas en la web contienen referencias espaciales.

Si bien investigaciones académicas en textos geográficos han confirmado el éxito en el uso de búsquedas con palabras clave de referencias espaciales [16], no se puede asumir que los servicios de búsquedas espaciales que ofrecen los más exitosos motores de búsqueda cubren las necesidades de todos los posibles usuarios [5].

Por otra parte, el avance en los sensores GPS¹ y la reducción en sus precios han permitido que los dispositivos comúnmente usados como cámaras fotográficas digitales, celulares, generen fotografías georreferenciadas proporcionando información de localización y facilitando la creación de modelos virtuales de las ciudades [17].

De manera similar, los comentarios de las fotografías muchas veces hacen posible inferir la localización geográfica donde fueron tomadas. Esta técnica es empleada por Rattenbury [18] al estudiar formalmente algoritmos para evitar ambigüedad de los *tags*.

Recuperación de datos espaciales y modelos de georreferenciación automática

Después de haber explorado las diversas fuentes y técnicas de extracción de referencias espaciales y de haber expuesto algunas cifras que reflejan la gran cantidad de textos con

referencia espacial en Internet, a continuación se explorarán algunos prototipos de software que realizan la tarea de extraer información espacial en textos y los componentes que definen su funcionamiento.

A pesar de que la extracción de información espacial ha sido un tema explorado ampliamente, las experiencias en implementación de modelos y prototipos arquitectónicos o computacionales son pocas. Tampoco se hace énfasis en el tiempo de construcción de dichos sistemas, esquema de pruebas, comparación y revisión de estos [19]. Sin embargo, a continuación se expone un resumen de la exploración de estas iniciativas implementadas, resaltando los componentes o procesos que son necesarios para llevar al cabo la georreferenciación automática.

Uno de los ejemplos más completos de modelos de búsqueda de referencias espaciales es el proyecto Spirit (Spatially Aware Information Retrieval on the Internet) [20], el cual propone ampliar las funciones de búsqueda de información espacial en Internet dando importancia a las relaciones espaciales implícitas en las búsquedas con el fin de recuperar información aparentemente más acertada. Sin embargo, dicho proyecto no expone la validación de su modelo, lo cual hace que dicha propuesta sea débil en el momento de su aplicación. A pesar de ello es notable la propuesta en cuanto a los componentes que gestionan o manejan los diferentes ámbitos necesarios para cumplir con la georreferenciación, dividiendo los aspectos más importantes de la extracción de referencias espaciales y georreferenciación de textos, en los siguientes módulos:

- Ontologías que modelan la terminología geográfica.
- Algoritmos para gestionar las consultas y algoritmos para ordenar por relevancia basados en ontologías geográficas.
- Algoritmos de aprendizaje automático para la extracción del contexto geográfico en documentos web y para generar metadatos que definen el contexto geográfico.
- Interfaces multimodales que proveen formas de interacción del usuario para plantear las búsquedas o establecer palabras clave.
- Índices espaciales para categorizar y ordenar colecciones de documentos.

Otro ejemplo notable lo expone Markowertz [21] al proponer la implementación de un prototipo de motor de búsqueda geográfico que lleva a cabo la extracción masiva

¹ GPS (Global Positioning System o sistema de posicionamiento global).

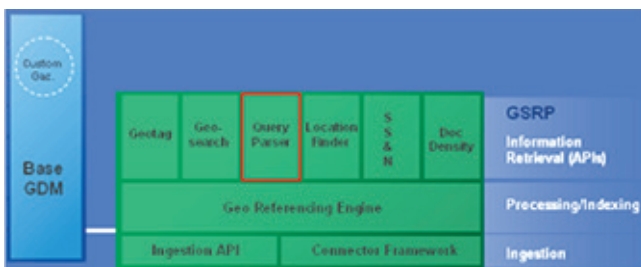
de elementos geográficos de los datos de la web añadiendo una referencia cruzada o enlace a la página de la que se extrajo la información. Los aspectos de la recuperación automática de información espacial que propone dicho autor son: [22].

- Geocodificación.
- Búsqueda de patrones espaciales de páginas web.
- Integrar bases de datos de topónimos, direcciones y puntos de interés.
- Geoextracción.
- Geomatching.
- Geopropagación.
- Ejecución básica de las consultas espaciales.
- Búsqueda geográfica.
- Ranking geográfico.
- Procesamiento eficiente de consultas geográficas.

Desde un enfoque empresarial, el gigante de producción de información geográfica Metacarta ofrece servicios y componentes de búsqueda geográfica con las siguientes características [23].

- *Geotagging*: analiza el contenido para identificar referencias geográficas utilizando el contenido de etiqueta geográfica en otras aplicaciones.
- *Geoparsing*: software que lee los documentos y páginas web de una manera similar a la humana para identificar términos y referencias geográficas que utilizan el procesamiento del lenguaje natural (NLP).
- Módulos de datos geográficos: acceso a un nomenclátor digital que incluye más de 225 millones de nombres de lugar con la latitud, desambiguados.

Figura 1. Componentes de GIR (Geographic Information Retrieval) ofrecidos por Metacarta,



Fuente: <http://www.metacarta.com/products-platform-queryparser.htm>

La figura 1, tomada de la página de Metacarta, expone los componentes ofrecidos por dicha empresa.

Sin embargo, debido que son iniciativas empresariales, los algoritmos no están disponibles y dichos programas pueden considerarse sistemas cerrados.

Por otra parte, la extracción de elementos y patrones espaciales no sólo es posible a través de textos, sino también usando imágenes. Un ejemplo de un prototipo de recuperación de datos espaciales lo ilustra Zang [24], quien propone formalmente un mecanismo para indexar datos de tipo vectorial usando Ltree, de acuerdo con las relaciones topológicas [25] y características de las formas de los vectores (figura 2).

Los componentes que propone son los siguientes:

- Extracción de invariantes topológicas.
- Componente de análisis independiente para extraer elementos de patrones topológicos de las colecciones de vectores.
- Componente de automarkado.
- Máquina de análisis de vectores difuso (FSVM Fuzzy Support Vector machine).
- Componente de reconocimiento (SVM).

Figura 2. Componentes de GIR del prototipo de Zang.



Figure 2. The case based spatial data retrieval framework

Fuente: Zang [24].

La anterior imagen ilustra la interacción de los componentes para lograr la extracción de patrón espacial e indexado.

La recuperación de información espacial para este caso se realiza reconociendo patrones topológicos y formas en colecciones de objetos vectoriales.

En el ámbito de la biología, Guralnick [2] propone usar la georreferenciación automática para el mapeo de presencia territorial de especies, pues permite representar cartográficamente las especies nombradas en colecciones de documentos a partir de textos.

Los componentes que propone se listan a continuación:

- Extracción de párrafos con referencias espaciales.
- División de los textos en oraciones con referencias espaciales.
- Interpretar las semánticas de cada oración.
- Mapeo de nombres a posiciones de una gacetera.
- Refinamiento de la posición geográfica a partir de operaciones de análisis espacial.

Similar a la propuesta anterior, Beaman [4] expone la necesidad de la georreferenciación automática en el ámbito de la biogeografía, para realizar el mapeo cartográfico de especies y permitir la realización de análisis biogeográficos. Propone herramientas informáticas al tema de la bioinformática, cuyos elementos notables que componen su sistema se listan a continuación.

- Reprocesamiento del lenguaje local eliminando anomalías (estandarización).
- Análisis de frases.
- *Parsing* de texto y ajuste de patrones usando expresiones regulares con tipos predefinidos junto con nombres de lugares y sus relaciones.
- Cálculo de coordenadas geográficas.
- Retorno de los resultados.

Según los modelos de georreferenciación explorados, se tienen elementos transversales en común, a manera de componentes fundamentales que necesitan un software para extraer referencias espaciales y georreferenciar información textual. La tabla 1 resume los aspectos característicos de cada prototipo:

Finalmente, para entender el cuadro comparativo mostrado en la tabla 1, se hace necesario explicar en detalle, algunos de los elementos que componen un prototipo de software que genere información georreferenciada automáticamente, ya que cada aspecto, como sinonimia, ambigüedad, relevancia textual, relevancia espacial, geocodificación, poseen desafíos en cuanto a su diseño, implementación y funcionamiento. Incluso algunos de ellos son materia de investigación en la actualidad, por lo cual haremos una breve exploración de algunos de estos aspectos.

Tabla 1. Cuadro comparativo de los diferentes prototipos de georreferenciación automática estudiados

	Extracción automática de la web	Uso de algoritmos de extracción de textos	Uso de ontologías	Algoritmos para calcular relevancia	Algoritmos para eliminar ambigüedad	Algoritmos para tratar sinonimia	Uso de <i>parsers</i> o estandarizadores	Uso de geocodificadores y gaceteras	Interfaces multimodales	Técnicas de inteligencia artificial
Spirit										
Metacarta										
Markowitz										
Biogemancer										
Zang										
Beaman										

Convenciones

	Especificado en los <i>papers</i>
	No especificado

Fuente: elaboración propia

Sinonimia

Existen muchas formas diferentes en las cuales una localización puede ser referida en un texto, y estas formas pueden ser explícitas, implícitas y vagas.

La forma explícita se refiere a cuando se encuentra una localización en un texto, ya que sus coordenadas pueden ser adquiridas con el uso de una gran base de datos de nombres y posiciones (Gacetera).

En la web es frecuente encontrar referencias espaciales en textos con nombres alternativos o no tan comunes de algún lugar. Por ejemplo, en vez de Bogotá podemos encontrar Distrito Capital, o Santafé de Bogotá; en estos casos, aunque estamos tratando de la misma ciudad tenemos diferente nombre, o topónimo; otro ejemplo lo brinda Buscaldi [26] cuando expone el topónimo "Saint Petersburg" y "Leningrad",

que aunque son dos topónimos diferentes indican la misma ciudad. Esos son casos de sinonimia en la recuperación de datos espaciales en textos.

En algunos casos es posible evitar la dificultad de la sinonimia por medio de tablas con nombres alternativos, como lo expone Denshan [27]; se consulta en primera instancia la coordenada del nombre a buscar y en caso de no encontrarla se consulta un nombre alternativo que potencialmente pueda tener coordenadas válidas.

Ambigüedad

En general las palabras son ambiguas cuando estas pueden tener más de un significado [10]. El ejemplo clásico de ambigüedad lo tenemos con la palabra banco (bank en inglés), que en español lo podemos asociar con dos significados: la entidad financiera o el lado del cuerpo de agua.

Desafortunadamente la ambigüedad se extiende también a topónimos Smith y Mann [28] mostraron que particularmente en ciudades que fueron colonizadas se reutilizaron nombres, por ejemplo existe un Springfield en casi cada estado de Estados Unidos.

Desafortunadamente también es frecuente encontrar ambigüedades con nombres de personas, apellidos o cosas.

- Nombres de personas, ejemplo: estado de Victoria en Australia, ciudad de Estados Unidos. En Colombia los ejemplos son comunes, como localidad Rafael Uribe, San Andrés islas, barrio Galán, etc.
- Organizaciones que toman el nombre de algún lugar, por ejemplo: clubes de fútbol como Real Madrid, AC Milán, Club América de Cali, etcétera.
- Nombres usados metafóricamente como Hollywood, que son usados más para referirse a la industria del cine que a la ciudad de California.
- Palabras empleadas en diferentes lenguajes teniendo significados muy diferentes, como por ejemplo, una ciudad en Alemania llamada The, o también en Alemania una ciudad llamada Fucking [29] o calle la Independencia en Cali.
- Lugares con significados de fechas, por ejemplo, barrio 20 de julio, barrio 12 de octubre.
- Lugares con ambigüedades a nivel de escala, como por ejemplo en Bogotá la llamada avenida Ciudad de Cali.

- Palabras utilizadas que significan diferentes cosas, como por ejemplo el código postal de Estados Unidos, que es expresado mediante cinco dígitos, que a la vez pueden ser confundidos como cinco dígitos de un número en cualquier contexto.

La calidad en la identificación de topónimos en textos ha sido estudiado durante muchos años en investigaciones lideradas por las conferencias de MUC (Message Understanding Conferences) en los noventa, también en las investigaciones de Olligschlaeger y Hauptmann [30], quienes trataron el problema de desarrollar una librería videodigital, intentando pintar en el mapa algunas localizaciones mencionadas en noticias. También Smith y Crane [31] y Rauch *et al.* [32] se valieron de heurísticas basadas en el cálculo de centroides, enfocándose en las referencias geográficas en documentos que emplearon datos de población de las ciudades, para determinar a qué ciudad se refieren. En estos casos la ambigüedad se resuelve con la heurística que se fundamenta en la suposición de que es más probable mencionar lugares altamente poblados, que lugares con baja población.

Garvin y Mani [33] fueron un paso más allá al plantear clasificadores estadísticos que fueron entrenados con conjuntos de datos sobre topónimos.

Buscaldi [34] creó un método basado en mapas y análisis espacial, en el que se plantean procedimientos para eliminar la ambigüedad de textos con referencias espaciales.

Más recientemente el trabajo se ha incrementado, pero el tema de identificación automática de topónimos se ha convertido en área de investigación [34].

Relevancia textual

Según Shekar y Xiong [10], la relevancia textual depende de los tipos de documentos que van a ser analizados. Por ejemplo, para colecciones de documentos como archivos de oficina, artículos, periódicos, correos electrónicos, la relevancia textual es calculada buscando palabras que los autores con frecuencia repiten, siendo este un enfoque efectivo para encontrar referencias [35]. Un ejemplo puede ser revisando en diferentes fuentes el reporte de una noticia, el cual es relevante cuando el algoritmo detecta de diferentes fuentes los mismos conjuntos de palabras, siendo valiosa la noticia, ya que diversas fuentes validan la ocurrencia de un mismo incidente.

Otro enfoque para examinar la relevancia en documentos web es revisar documentos que estén relacionados o que tengan enlaces [36], y revisar las referencias espaciales referidas en dichos textos.

Relevancia espacial

Determinar la relevancia espacial de una oración o determinar qué tan importante espacialmente es la oración analizada, comienza con establecer la firma espacial de la consulta y determinar los documentos que son relacionados con el área espacial de la consulta. La firma espacial de una consulta es determinada por la localización de una relación especificada en la consulta, por ejemplo, la búsqueda “concesionarios cerca a Chía” establece la relación de cercanía entre el topónimo Chía y los concesionarios que son las localizaciones a buscar.

Determinar la relevancia espacial de la consulta también implica estimar el tamaño y la forma del topónimo, ya que según el texto de la consulta que contenga una referencia espacial o topónimo, puede traer mucho o pocos resultados según el cubrimiento geográfico que signifique. Por ejemplo, la palabra Unicentro (que es un centro comercial) puede traer menos resultados que la palabra Bogotá (que es una ciudad).

De igual manera, si los usuarios buscan restaurantes en una ciudad, los resultados pueden ser muchos, de acuerdo con la cantidad de estos en una ciudad, a diferencia de buscar aeropuerto igualmente en una ciudad (que puede retornar pocos resultados).

Una vez la firma espacial ha sido establecida, documentos textual y espacialmente relevantes son recuperados, lo que implica combinar los puntajes o conteos de los componentes espaciales y textuales, como lo describe Kreven [37]. De la misma manera dicho autor expone algoritmos para calcular la dispersión de una búsqueda y reporta mejoras utilizando motores de búsqueda determinados.

Otros enfoques exploran el uso de métodos de razonamiento espacial donde se emplean relaciones topológicas y proximidad espacial [38] para determinar qué tan relevantes son los resultados de búsquedas que impliquen un criterio espacial.

Por otra parte, otros estudios, como los de Vaid [39] emplean las palabras en la consulta espacial, operadores que

tienen asignados una relevancia en la cual las palabras que poseen más notabilidad son las que implican una relación espacial con el lugar, por ejemplo: cerca, al norte, próximo, son indicadores que el texto es relevante espacialmente.

De forma similar, los trabajos de Jones [40] combinan distancias y relaciones ontológicas para establecer relevancia espacial, siendo este un caso muy interesante, ya que emplea el enorme potencial que ofrece las ontologías en el caso espacial para calcular relevancia.

Consulta geográfica o no geográfica (*query triague*)

El término *query triague* hace referencia a cómo es la forma más apropiada para realizar una exploración en algún motor de búsqueda, para que dicho motor entienda que se está haciendo referencia a una búsqueda espacial. En otras palabras, dependiendo de la forma como el motor de búsqueda sea usado, él debe identificar si la búsqueda que realiza el usuario hace referencia a un lugar, y así desplegar las capacidades de búsqueda local para que el contexto [7] sea utilizado en realizar la búsqueda. En muchos casos simplemente es necesario examinar en las consultas del usuario la presencia del nombre de un lugar o topónimo, para que sea suficiente determinar si es una consulta espacial o no.

Una vez se ha identificado que es una búsqueda espacial, es necesario establecer los componentes de esta, su propósito y la relación espacial entre el sujeto y la consulta. Reconocer la relación espacial en una consulta es relativamente un aspecto de investigación; la existencia de palabras usadas comúnmente como: dentro, en, al norte, cerca, son indicadores de que puede haber una relación espacial.

Como lo indican Shekar y Xiong [10], una manera muy simple de determinar si una consulta es espacial, es diseñar una interfaz de usuario que capture estos elementos individualmente.

Conclusiones

Existen temas poco explorados en geomática que vale la pena investigar, debido a las nuevas tendencias en las ciencias de la información y las telecomunicaciones. Uno de ellos es la recuperación e interpretación espacial de textos de la web.

En este artículo se presentaron algunas cifras de información georreferenciable en la web; sin embargo, es

necesario conocer actualmente cuántos textos pueden ser georreferenciados, bajo qué niveles de calidad y cuáles componentes de software pueden hacer posible esta tarea automáticamente.

Asimismo, existen todavía temas abiertos a investigar, como son la ambigüedad espacial y la relevancia textual, temática y espacial en recuperación de textos de la web.

Se puede concluir que es posible construir un software que logre realizar automáticamente dicha tarea, el cual puede ser el primer paso para elaborar sistemas avanzados que realicen tareas como las siguientes:

- Representación espacial de datos de la web usando cartografía temática.
- Herramientas para integrar automáticamente información en la web de tráfico vehicular.
- Servicios web que analicen los contenidos de correos y agendas virtuales y brinden sugerencias de rutas de acuerdo con la localización y hora de las reuniones.
- Generación automática de metadatos con referencia espacial, para documentos en español.
- Herramientas para construir mapas de criminalidad de acuerdo con datos de la web.
- Geocodificador semántico.

Referencias

1. Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.
2. Guralnick, R. P. et al. (2006). BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology* 4(11), p. e381.
3. Adriani, M. Paramita, M. L. (2007). *Identifying location in indonesian documents for geographic information retrieval*. ACM.
4. Beaman, R. Conn, B. (2003). *Automated geoparsing and georeferencing of Malesian collection locality data*. *Telopea*, 10(1) 43-52.
5. McCurley, K. S. (2001). *Geospatial mapping and navigation of the web*. ACM.
6. Buscaldi, D. y Rosso, P. (2008). *Geo-wordnet: Automatic georeferencing of wordnet*. Proc. LREC, Marrakech, Morocco.
7. Worboys, M. y Duckham, M. (2004). *GIS: A computing perspective*. CRC.
8. Gilleland, M. (2009). *Levenshtein distance, in three flavors*. Merriam Park Software: <http://www.merriam-park.com/ld.htm>.
9. Kavouras, M. y Kokla, M. (2008). *Theories of geographic concepts: ontological approaches to semantic integration*. CRC.
10. Shekar, S. y Xiong, H. (2008). *Encyclopedia of GIS*.
11. Aho, A.V. y Corasick, M. J. (1975). Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6): 333-340.
12. Buyukokkten, O. et al., (1999). *Exploiting geographical location information of web pages*.
13. Aarts, E. et al., (1994). A computational study of local search algorithms for job shop scheduling. *ORSA Journal on Computing*, 6(2): 118-125.
14. Wang, L. et al. (2005). *Detecting dominant locations from search queries*. ACM.
15. Sanderson, M. y Kohler, J. (2004). *Analyzing geographic queries*.
16. Gey, F. et al. (2006) *GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview*. Accessing Multilingual Information Repositories, 908-919.
17. Snaveley, N.; Seitz, S. M. y Szeliski, R. (2008). Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2) 189-210.
18. Rattenbury, T. y Naaman, M. (2009). *Methods for extracting place semantics from Flickr tags*. ACM Transactions on the Web (TWEB), 3(1): 1.
19. Frigg, R. y Stephan, H. (primavera de 2012). *Models in Science*. Obtenido de: <http://plato.stanford.edu/archives/spr2012/entries/models-science>.
20. Jones, C.B. et al. (2002). *Spatial information retrieval and geographical ontologies an overview of the SPIRIT project*. ACM.
21. Markowetz, A. et al. (2005). *Design and implementation of a geographic search engine*.
22. Markowetz, A.; Brinkhoff, T. y Seeger, B. (2005) *Geographic information retrieval. Next generation geospatial information*, 5-17.
23. Metacarta (2012) *MetaCarta GSRP*. Obtenido de: <http://www.metacarta.com/products-platform-queryparser.htm>
24. Zhang, J.; Pan, H. y Yuan, Z. (2009). *A novel spatial index for case based geographic retrieval*. ACM.
25. Egenhofer, M. J.; Clementini, E. y Felice, P. D. (1994). Topological relations between regions with holes. *International Journal of Geographical Information Systems*, 8(2), 129-142.

26. Buscaldi, D. (s.f.). *Toponym Disambiguation in Natural Language Processing*.
27. Densham, I. y Reid, J. (2003) *A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service*. Association for Computational Linguistics.
28. Smith, D. A. y Mann, G. S. (2003) *Bootstrapping toponym classifiers*. Association for Computational Linguistics.
29. BBC_mundo (2012). Pueblo llamado Fucking quiere cambiar de nombre, *Semana*.
30. Olligschlaeger, A.M. y Hauptmann, A.G. (1999). *Multimodal information systems and GIS: The Informedia digital video library*.
31. Smith, D. y Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Research and Advanced Technology for Digital Libraries*, 127-136.
32. Rauch, E.; Bukatin, M. y Baker, K. (2003). *A confidence-based framework for disambiguating geographic terms*. Association for Computational Linguistics.
33. Garbin, E. y Mani, I. (2005). *Disambiguating toponyms in news*. Association for Computational Linguistics.
34. Buscaldi, D. y Rosso, P. (2008). *Map-based vs. knowledge-based toponym disambiguation*. ACM.
35. Baeza-Yates, R. y Ribeiro-Neto, B. (1999). *Modern information retrieval*, 82.
36. Levene, M. (2006). *An introduction to search engines and Web navigation*. Wiley Online Library.
37. Kreveld, M. *et al.*, (2005). Distributed ranking methods for geographic information retrieval. *Developments in Spatial Data Handling*, 231-243.
38. Andrade, L. y Silva, M. J. (2006). *Relevance ranking for geographic IR*. ACM GIR.
39. Vaid, S. *et al.* (2005). Spatio-textual indexing for geographical search on the web. *Advances in Spatial and Temporal Databases*, 923-923.
40. Jones, C., H. Alani, y Tudhope, D. (2001). Geographical information retrieval with ontologies of place. *Spatial Information Theory*, 322-335.

