Buscador Geoespacial para Honduras

Dirk Ahlers
UNITEC – Universidad Tecnológica Centroamericana
Sistemas Computacionales, Facultad de Ingeniería
Tegucigalpa, Honduras
ahlers@dhere.de

Resumen – La búsqueda local vía web se ha convertido en un importante y popular medio para la recopilación de información relacionada geográficamente. Las búsquedas locales no sólo necesitan de algoritmos poderosos de los Sistemas de Recuperación Geográfica, sino también de una amplia disponibilidad de data localmente referenciada en la web y en otras fuentes. Mientras los sistemas comerciales disponibles funcionan satisfactoriamente en países desarrollados, otras regiones del mundo poseen desafíos en todas las áreas de búsqueda geoespacial. Este artículo examina el potencial para buscadores locales en Honduras. Los retos, entre otros, son la escasa cobertura de la web, pocas bases de datos públicas o comerciales, retos debido a la estructura informal de direcciones, incertidumbre y datos difusos para localización confiables, así como aspectos de privacidad y seguridad. El objetivo de este artículo es el de discutir esos retos y proveer de soluciones dentro del proyecto hacia el desarrollo de un motor de búsqueda geoespacial para Honduras.

Términos de indexación: Buscadores geoespacial vía web, motores de búsqueda, recuperación de información, extracción de información, estudio de factibilidad.

I. Introducción

En conjunto con el gran crecimiento de la World Wide Web, los motores de búsqueda se han desarrollado de igual manera, para indexar la gran cantidad de documentos, con el propósito de hacerlos disponibles a los usuarios. De igual manera, más y más información basada en un lugar es publicada en línea. Para que la información tenga sentido y que sea accesible a la búsqueda, se han realizado muchos esfuerzos de investigación [1, 2, 3, 4] y servicios comerciales tales como Google Maps, Google Earth y Yahoo! Maps han desarrollado aplicaciones para búsquedas basadas en localización y creado índices de información geoespacial [5]. Las propiedades geoespaciales de esta información permiten búsquedas diferentes a las búsquedas comunes en la web.

En vez de la búsqueda a través de palabras claves para recuperar una lista de documentos, la búsqueda geoespacial permite colocar los resultados en un mapa, buscar por distancias o regiones y recupera información más relevante y estructurada acerca de la entidad geo referenciada. En este sentido, la búsqueda local busca a las entidades geográficas en los documentos y no sólo lo modela sino que también modela la entidad real referenciada en los documentos [1].

Para implementar dicho sistema, en vez de la utilización de motores de búsqueda de propósitos generales de la web, se emplea el concepto de buscadores especializados o verticales. Los motores de búsquedas verticales tienen como propósito desbloquear información para tareas o dominios específicos [6]. En las búsquedas verticales, las restricciones especiales son combinadas con procesamiento, indexación y búsquedas de tópicos específicos. Esto permite tomar en cuenta características específicas y metadata potencial, y construir interfaces y aplicaciones a la medida. La búsqueda local es de esta manera una búsqueda vertical ya que requiere características especiales de la data como ser no tener palabras claves, y requiere estructuras especiales de procesamiento e indexación y permite una interacción más allá del puro paradigma de palabras claves.

La búsqueda geoespacial tiene un gran potencial porque la localización es un poderoso principio de organización en la vida de la gente debido a su fuerte correlación con el mundo real. Sin embargo, la mayoría de los esfuerzos de investigación y desarrollo van hacia países desarrollados, específicamente a Estados Unidos de América y Europa en correspondencia con las sedes de las compañías de investigación y su gran impacto en las economías. El mercado de motores de búsqueda en Honduras está dominado por motores de búsqueda estadounidenses sin ningún actor independiente. Algunos motores de búsqueda están desactualizados y sin mantenimiento o iniciando a ser implementados de manera que no tienen suficiente cobertura. Pero la búsqueda web es universal y de esta manera, motores de búsquedas de propósitos

generales pueden fácilmente reunir información de pequeños países, lo que de hecho hacen y son de esta manera suficientemente útiles. El ejemplo de google.hn provee un servicio local que prefiere páginas que se refieren a Honduras, pero aun así siempre accesa a otras páginas en su índice. Obviamente, esto ya usa algún método para georeferenciar documentos con una granularidad gruesa, aparentemente basado en nombres de dominios, idioma y palabras claves en su mayoría.

La situación se vuelve más fuerte cuando se explora la búsqueda local para Honduras. Los grandes motores de búsqueda proveen data de mapas, algunas veces con buena calidad y también proveen alguna búsqueda local basada en mapas. Sin embargo, comparada con otras regiones del mundo, hay muy poca información disponible y su profundidad es poca, a menudo ofreciendo nada más que un nombre y una localización aproximada. En vez de esperar a que otros actores se posicionen en el mercado, la situación actual ofrece un potencial único para desarrollar un motor de búsqueda geoespacial para Honduras. Mientras la cobertura de la web es aún baja y el esquema de direccionamiento hace que una localización exacta sea extremadamente difícil, la utilización de la web se incrementa, fuentes de datos potenciales existen, y la gente empiece a usar los servicios de búsqueda basados en localización lo cual crea suficiente demanda y apoyo.

El resto de esta investigación explorará los retos individuales enfrentados en Honduras y ofrece soluciones potenciales dentro del proyecto de investigación y desarrollo. El propósito de la investigación es el de usar información de localización para construir un motor de búsqueda geoespacial con el objetivo de utilizarlo en tecnología móvil con un proveedor de servicios de redes móviles. El trabajo detalla las etapas iniciales del proyecto, fuentes de datos potenciales, el enfoque y arquitectura propuesta, así como las etapas posteriores y desarrollo.

II. Retos

A. Requerimientos para búsqueda y movilidad basados en localización.

El primer paso para crear un motor de búsqueda geoespacial es lograr una visión de cuales aplicaciones y servicios podrían ser interesantes y relevantes. Los posibles usos varían desde análisis de mercados, análisis de viabilidad e investigación de fuentes de datos y también requerimientos de ingeniería para análisis de datos, estudios de usuario respecto a búsquedas y aplicaciones móviles. Estos pueden proveer una visión inicial dentro de las necesidades en los sistemas actualmente disponibles y podrían proveer también requerimientos para los pasos subsecuentes. Una encuesta y estudio de usuarios está siendo preparada para lograr conocimiento sobre el uso de datos de localización por la gente y como ellos buscan y averiguan acerca de información local, cuales son los modos preferidos y que fuentes de información utilizan para la búsqueda y como esto podría ser incorporado dentro del motor de búsqueda. La investigación tratará con descubrimiento de la búsqueda y aspectos de recuperación de la misma, así como aspectos tecnológicos, equipo y conocimiento. Como resultado preliminar, se puede decir que el modo preferido de búsqueda sigue siendo de boca a boca o conocimiento previo de localización, en conjunto con un círculo de conocimiento. Los motores de búsqueda se utilizan, pero no tanto para obtener información local.

B. Desarrollo de un motor de búsqueda geográfica para Honduras.

Como se discutió anteriormente, existe la necesidad de búsquedas locales en Honduras. Pero no puede ser satisfecha por ninguno de los servicios existentes. Por lo tanto, un motor de búsqueda geoespacial a nivel de país tendrá que ser desarrollado. Debido a la situación de las fuentes de datos, se tiene que seguir un enfoque híbrido tanto para la búsqueda en la web con documentos georeferenciados, así como para bases de datos adicionales y unión para fuentes de datos específicos. Esto es explorado con más detalle en Sección IV. Este sistema es por mucho el más grande reto a enfrentar, pero los posteriores son dependientes del mismo y necesitan ser solventados mientras se discuten los prerrequisitos.

C. Análisis de fuentes locales de datos de Honduras y el grafo de la web hondureña.

Los primeros pasos del análisis de mercado han sido terminados sin encontrar resultados significantes tal como se discutió anteriormente. Debido a que no se encontraron motores de búsqueda para el mercado, la investigación se movió a una búsqueda de fuentes de datos potenciales para la búsqueda geoespacial. El motor de búsqueda debe ser primariamente un sistema de recuperación de información automático y sólo, secundariamente, si aún caso, ser asistido por contenido contribuido por el usuario, la identificación y el análisis de fuentes de datos utilizables es lo de mayor importancia. Un gran número de fuentes de datos han sido identificadas que pueden proveer datos y ser el objetivo para una entidad de recuperación de datos aparte de un motor de búsqueda general. Esto es detallado en Sección III. Para el desarrollo de búsqueda local basado en la misma web, los grafos de la web y el análisis de la estructura de enlaces pueden proveer puntos de vista dentro de la data dispo-

nible de la web hondureña, dar un método de clasificación y ayudar a recolectar información más eficientemente.

D. Explotando data local

Una característica desafiante de referencias locales de Honduras es sobre la exacta localización en la forma de direcciones respecto a numeración de las casas, el esquema de alta granularidad generalmente no son dadas. Esto impide seriamente un enfoque de alta granularidad que podría mapear la información a edificios individuales. [7, 8, 9]. Hay algunas áreas o pequeñas ciudades donde existe un patrón rectangular de calles, que generalmente ayudan a tener un mejor esquema de direccionamiento. Sin embargo, en la mayoría de regiones, las referencias de dirección son dadas por nombre de ciudad, distritos y algunas veces por nombre de calle. Se han desarrollado otras formas de direccionamiento que permiten la localización de un edificio individual. A menudo, esas descripciones son dadas junto con información adicional tales como lugares de interés o edificios cercanos conocidos. Algunas veces las direcciones son acompañadas de un pseudo mapa, llamados croquis para ayudar a la localización. La baja granularidad encontrada en las referencias – en páginas comunes de la web así como en bases de datos – presenta un problema particular para el análisis geográfico, el cual es la extracción de referencias locales de un texto [10]. El análisis geográfico es un tema ampliamente examinado en recuperación de información geográfica y ha sido investigado extensivamente. Por lo tanto, sería posible ejecutar el análisis geográfico basado en regiones. Para mejorar en este aspecto, la Sección IV explorará las posibilidades para lograr alta granularidad aún con referencias de direcciones no tan específicas.

E. Tema de granularidad en aplicaciones móviles.

El tema mencionado anteriormente sobre la falta de datos de direcciones exactas no solamente dificulta la extracción de información, sino que también su uso. Datos sobre localización del usuario son utilizados en muchas aplicaciones como el dato inicial y más importante para la recuperación de información. Si la información es desplegada en un mapa, la localización es marcada. Pero sin un mapa como interface, dar una descripción y una dirección exacta del lugar es difícil. Si una descripción de navegación estuviera disponible, entonces podría ser dada como parte de los resultados. Pero en otros casos, una dirección tendría que ser generada basada en edificaciones cercanas, calles principales, etc., una vez esas estén disponibles [11]; similar a navegación con sitios de interés [12], [13].

F. Seguridad y privacidad en geolocalización.

El tema de revelar la ubicación de alguien podría permitirles a los atacantes a inferir la ubicación de la casa, ruta o lugares favoritos. Por el contrario, usar servicios de localización para aplicaciones tales como conocer la ubicación de los amigos o colegas tiene múltiples beneficios. Sin embargo, además del tema de privacidad, los hondureños necesitan considerar las preocupaciones reales de seguridad. Debido al alto nivel de criminalidad, mucha gente prefiere mantener su información personal, especialmente su ubicación, muy privada. Pero muchas entradas en servicios de intercambio de información explícitamente concierne a la propia casa de la gente (“Mi casa”, “My house”). En esos casos, la funcionalidad parece anular los problemas de seguridad. Sin embargo, los servicios que están siendo desarrollados tendrán que ser cuidadosos al implementar técnicas y procesos que tomen en cuenta los requerimientos de seguridad y privacidad de la ubicación.

III Fuentes de Datos Potenciales

Una investigación de fuentes de datos disponibles para Honduras fue realizada para lograr una vista a vuelo de pájaro de datos existentes tal como se discutió en Sección II-C. La investigación se empezó a través de fuentes de datos para otros países, de fuentes globales o por medio de motores de búsqueda. Expandiendo la investigación por medio de búsquedas generales en la web y en motores especializados de búsqueda, se encontraron más fuentes. Esto condujo a una gran cantidad de fuentes de datos potenciales. El resto del artículo considera fuentes de datos especializadas y la Deep Web [14, 15] y subsecuentemente discute la web en general.

A. Directorios y bases de datos geoespaciales.

Un número de fuentes fueron identificadas para obtener entidades de información estructurada. Una clasificación de las fuentes fue ejecutada y dada seguimiento, un examen inicial fue realizado para conocer la profundidad, densidad y otras propiedades de la información. Una revisión breve de la cantidad de información se muestra en la Fig. 1 en una escala logarítmica. La cantidad de entidades fue estimada a través de una consulta de ubicaciones o palabra clave Honduras, donde fuera aplicable. En algunos casos, el enfoque más directo no fue factible y no habían sido desarrolladas ninguna implementación de conectores para las fuentes de datos, lo cual explica algunas barras vacías. Las fuentes son extremadamente heterogéneas. Mientras algunas fuentes muestran información muy detallada, otras sólo muestran un nombre y una dirección parcial. Esto influye en el uso de los datos que pueden variar desde data que se refiere únicamente

a información de soporte como para descartar la data en algunos casos. Existen varios servicios de mapeo que no solamente proveen buenas imágenes satelitales sino que también redes de carreteras.

Aparte de entidades de información reales tales como Open-StreetMap, agencias de viajes, etc., hay alguna información de diccionarios geográficos que serán utilizados para ubicaciones terrestres en el motor de búsqueda propuesto. La más completa es geonames, y otros que proveen solamente una fracción de su cobertura. Sin embargo, un hallazgo interesante es que geonames tiene la cantidad más alta de lugares. Esto significa que los demás servicios tienen mucho menos cobertura que simplemente todos los nombres de lugares en Honduras. Cuando se remueven los ríos, montañas, etc., y reteniendo nombres de lugares poblados, geonames se mantiene arriba que las demás fuentes, lo cual indica que los demás no tienen al menos una entidad por ciudad o pueblo, dejando mucho trabajo por hacer.

Este es un trabajo en proceso para aprender más acerca de las fuentes de datos y extraer data estructurada de entidades.

B. Estimados esperados para cobertura Web

Una buscador geoespacial debería no sólo tomar data de otros directorios, sino que realmente tomar páginas web que tratan con la ubicación y encontrar los documentos sin estructura pero geo-relevantes y analizarlos. Para ello, un primer paso es asegurar donde y como se pueden encontrar las páginas que tratan sobre Honduras. Por lo tanto, se está haciendo una estimación del número de dominios potenciales con información Web que tratan sobre Honduras.

El primer paso es examinar el espacio de direcciones IP y los nombres de dominio de Honduras. Datos de la base de datos de bloques de IP[1] muestra alrededor de 129,000 IPs en el bloque asignado a Honduras. Este número sólo consta de bloques de direcciones, pero no de direcciones IP asignadas o servidores reales, de tal manera que el número real de servidores es mucho menor. El Centro Coordinar de Redes (LACNIC) tiene porcentajes de asignación para todo el espacio de direcciones manejado[2]. Para esos bloques en los cuales Honduras (a través de nic.hn) tiene una parte, el promedio de asignación yace cerca del 90% y podría estimarse cerca de 116,000 servidores. Hay que hacer notar que no todas las

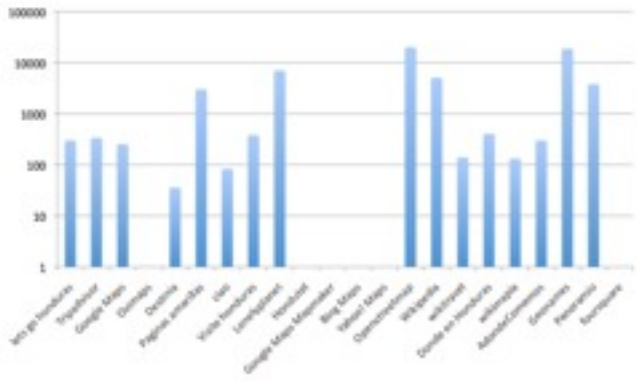

Fig. 1. Overview of information amount for examined sources

domain name	amount
.hn	4260
.com.hn	1253
.gob.hn	85
.org.hn	84
.net.hn	56
.edu.hn	41
.mil.hn	1
sum	5780

Fig. 2. Assigned Honduran domain names (as of June 2011)

direcciones IP están realmente asignadas, y no todas las que están asignadas están activas, y no todas las activas corren un servidor web. De esta manera, el número esperado de servidores web reales es mucho menor.

Datos más detallados de los nombres de dominios se pueden obtener del Registro Nacional de Nombres de Dominios[3]. Por motivos de protección de datos, el sistema de nombres de dominio (DNS por sus siglas en inglés) no permite la búsqueda por dirección IP; las búsquedas inversas de DNS no siempre son confiables, tampoco se pueden obtener todos los nombres de dominios que tiene disponibles. Esto no es específico para Honduras sino que para todos los demás países. Sin embargo, algunas estadísticas están disponibles acerca de los nombres de dominio asignados. Honduras utiliza un sistema estructurado de nombres de espacio de segundo nivel, pero no es obligatorio y un uso directo de un nombre de dominio raíz también es posible. La Tabla 2 muestra que hay cerca de 5780 nombres de dominio registrados en Honduras. Además de consultas inversas al DNS, es posible consultar cada una de esas direcciones IP para averiguar si hay un servidor web corriendo en el puerto por defecto.

Sin embargo, en muchos casos los servidores utilizan tecnología de virtualización para correr múltiples servi-

1 http://www.countryipblocks.net/country-blocks/select-formats/

2 http://www.lacnic.net/en/registro/espacio-disponible-ipv4.html

3 http://nic.hn/english/index.html

dores en un solo equipo. En este caso, el nombre de dominio debe ser dado, de otra manera ninguna información estará disponible. Por ejemplo, la dirección IP de Unitec (200.107.212.133) provee múltiples nombres de dominio, no solamente para unitec.edu, sino también para www.unitec.edu, www.ceutec.unitec.edu, etc. También, un dominio .hn ccTLD podría ser registrado en Honduras, pero el servidor físicamente podrían estar en algún otro lado y de esta manera no tener una IP asignada a Honduras. Este es un problema general al tratar de obtener contenido de ubicaciones usando solamente la infraestructura, el cual a su vez puede ser impreciso [16], [17].

Un problema relacionado, y potencialmente el más grande, es que bastante información acerca de Honduras es almacenada en páginas web fuera del dominio .hn ccTLD. Por ejemplo, unitec.edu tiene un dominio perteneciente al sector educativo de EUA. Sin embargo, está físicamente localizado en Honduras. En este caso sólo un análisis del contenido de la página web podría darnos una indicación que esta institución está realmente en Honduras y debería ser incluida en la búsqueda. Un problema adicional al usar sólo páginas web hondureñas, es que muchas páginas web internacionales podrían quedar fuera aun cuando tengan muy buena cobertura, especialmente los destinos turísticos y hoteles, o páginas generales relacionadas al área. Adicionalmente, especialmente información turística, tiene más disponibilidad en inglés que en español. El inglés, como un lenguaje regional reconocido, es hablado en Islas de la Bahía en el Caribe, un gran destino turístico. Esto no es normalmente un problema, pero en el caso de Honduras, a menudo mencionan al unico ciertos lugares.

Una fuente adicional para la búsqueda es el Open Directory DMOZ. Sin embargo, tiene muy poca cobertura para Honduras, pero se confirma lo escrito anteriormente. En su jerarquía en inglés, contiene solamente 421 entradas, con 10 del dominio .hn (2.5%) y 411 otros; en español, hay únicamente 96 entradas, pero 46 son del dominio .hn (48%), y 50 de otros dominios. La sección en inglés contiene principalmente sitios de viajes y descripciones generales, en cambio, la sección en español contiene páginas locales. Una primera aproximación estimada de los sitios que están disponibles, los links del DMOZ están correlacionadas con los 5780 dominios disponibles para .hn. Esto nos da un número estimado entre 6200 dominios a una sobreestimada cantidad de 225000 dominios, con su media geométrica debajo de 40000 dominios.

Finalmente, los motores de búsquedas generales tales como Google, Yahoo o Bing pueden ser utilizados para recuperar páginas individuales, pero ello limita la cantidad de resultados que se obtienen, de tal manera que sólo proporciona una visión limitada. La búsqueda de dominios únicos no es posible. Los resultados estimados de 60 hasta 500 millones de páginas son una sobreestimación, pero ayuda a estimar el potencial de páginas del país. En general, esto significa que existen bastantes problemas que nos impiden todos los nombres de dominio para Honduras. Una búsqueda educada pondría inicialmente el número de servidores web alrededor de 40,000, un número bien manejable para los arañas web modernos [18, 19]. El reto yace en el descubrimiento y selección de dominios relevantes y documentos realizado por una estrategia de araña focalizado [20].

Los pasos descritos hasta este punto son básicamente los mismos que se necesitan para construir un motor de búsqueda web a nivel de país. Un proyecto similar ha sido descrito en Chile [21]. Por lo tanto, el trabajo realizado en el proyecto podría sentar las bases para un motor de búsqueda general para los sitios web de Honduras si esos sitios que están relacionados con Honduras son identificados, sin extracción de identidades y validación.

C. Granularidad de las referencias para localización

Adicionalmente, ejemplos de direccionamiento y esquemas de direccionamiento están siendo recopilados tal como aparecen en la web para obtener datos de prueba para desarrollar subsecuentes extractores de localización orientados a entidades. Un tema difícil es que el sistema de direccionamiento es muy informal tal como se discutió en la sección II-D. El formato dado por la Organización Postal Universal (UPO, por sus siglas en inglés) es prácticamente desconocido [22]. Las direcciones son generalmente incompletas y sólo mencionan el distrito o lugares conocidos, por ejemplo, “Zona Jacaleapa, frente a Colonia Honduras, Tegucigalpa”, “final de Bulevar Morazán”, “3ª Calle, Tegucigalpa, Honduras”. Esto presenta grandes problemas para el reconocimiento de direcciones, resolución basada en toponimias, y parseo geográfico [23]. En general, la granularidad de las referencias a localidades es bastante baja y se deben emplear otros métodos para extraer y geocodificar localidades con gran precisión ya que los métodos previos [24] son de limitada aplicabilidad.

IV. Propuesta

Las secciones previas han mostrado que existen fuentes de datos disponibles para Honduras en la web. Sin embargo, las fuentes de datos estructuradas son escasas y las páginas web sobre Honduras están muy dispersas, con varios dominios .hn fuera. Esta sección describe la propuesta general y metodología hacia la implantación de un motor de búsqueda local para Honduras. Las dos

fuentes de datos principales serán páginas disponibles gratuitamente tal como se describió anteriormente, la otra fuente son las fuentes de datos estructurados.

Adicionalmente, las aplicaciones futuras podrían también integrar datos proporcionados por los usuarios. Esto permitiría crear un sistema más dinámico que podría ser capaz de proveer datos que no están disponibles aún en otras fuentes y podrían ser utilizados para realizar correcciones.

El esquema general de un sistema de recuperación de información se muestra en Fig. 3. Describe como las fuentes de datos son reunidas y procesadas dentro de un documento indexado de tal manera que los usuarios puedan hacer consultas al motor de búsqueda. La solución propuesta para la búsqueda local está basada en esto, pero empleará una arquitectura más específica que será descrita posteriormente. La primera parte se centra en un motor de búsqueda para reunir páginas relacionadas con el país. Contendrá componentes de arañas, indexación y búsqueda de páginas web sobre Honduras, y también la extracción de información específica adecuada al tipo de búsqueda deseada, consultas interesantes, y las características de Honduras. Como se discutió anteriormente, las páginas web sobre Honduras se encuentran muy diseminadas. El requerimiento es que todas deberían ser encontradas, pero sin tener que rastrear gran parte de la web. Para esto, la técnica de araño focalizado [20] será utilizada. Debido a que los aspectos geográficos no están bien atendidos en búsquedas puramente textuales, se necesitarán adaptaciones a través de la arquitectura del motor de búsqueda, que son discutidas posteriormente.

La extracción de las referencias a las localidades se realiza a través de un parseador geográfico [23, 7]. No solamente logra palabras claves que ocurren en el documento, sino también logra asignar semántica a las referencias. Esto además, se hace por medio de la ayuda de un diccionario geográfico, un tesauro geográfico de nombres de lugares conocidos y su relación, para ayudar a identificación y desambiguación de los nombres de los lugares. El diccionario geográfico más completo fue geonames. Debido a problemas de baja granularidad (sección III-C), el parseador geográfico no sólo tendrá el propósito de extraer referencias textuales de los lugares, sino que también explotar lugares sin dirección. Para resolver la baja granularidad, las páginas a menudo hacen uso de coordenadas, direcciones o ayudas de navegación, croquis, o mapas anidados. Hasta cierto punto estos pueden ser extraídos automáticamente.

Un motor de búsqueda web perdería cierta información que está escondida en las bases de datos especializadas mencionadas, las llamadas Deep-Web. Mientras el

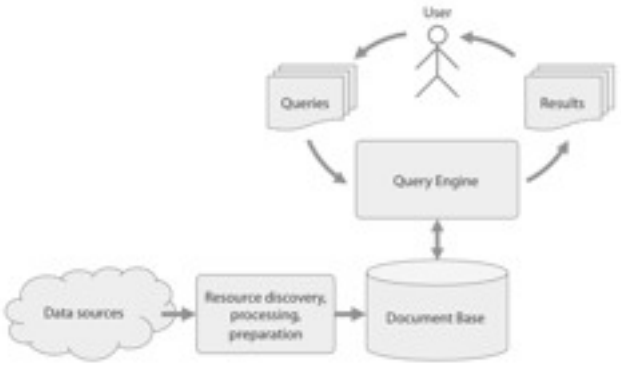

Fig. 3. Information Retrieval model (extended from [25])

contenido de esas bases de datos está disponible en la web para usuarios humanos, es difícil recuperarla a través de una araña web. Sin embargo, algunas de las fuentes son accesibles para conectores a través de APIs, otros podrían ser rastreados por las arañas web y tratados como data no estructurada o ser analizada por un programa específico. Las implementaciones iníciales de extracción ya están siendo desarrolladas para fuentes específicas para convertirse en parte de una arquitectura de búsqueda completa que pueda ser adaptada para este propósito. Alguno motores de búsqueda específicos tales como los agregadores usan un enfoque similar pera adicionalmente explotan estructuras específicas de dominio para recuperar metadata asociada or páginas como contenido estructurado [25]. A través de esto, desambigüedad en la extracción son mantenidas al mínimo y la data necesita solamente ser normalizada para ser usada dentro del sistema.

Naturalmente, el rastreador web y los conectores acumularán bastante información de entidades duplicadas. Esto implica que los sistemas deben fusionar la data través de reconciliación y mezcla. Un método potencial para corrección y para propuesta inicial hacia la solución, se propone un nuevo sistema inteligente de mezcla de datos. Va a tomar ambos, tipos de datos y así, los resultados de la web podrían ser validados y afinados por las fuentes de datos estructurados. Las características por las cuales se debe hacer la comparación son a través de localización, nombre, códigos telefónicos. etc. La fusión de los datos permitiría la búsqueda espacial de la web de Honduras haciendo uso de la fusión de esas fuentes de datos distintas a través de procesos mayormente automatizados y reunir muchos de los datos ya disponibles.

Debido a que muchas entidades no serán descritas de la misma manera, existe la necesidad de coincidencias no exactas. Además, el sistema de direccionamiento permite variaciones en las descripciones de las localidades para el mismo lugar, el cual es otro reto en el desarrollo de la mezcla de datos. Sin embargo, las localidades pueden

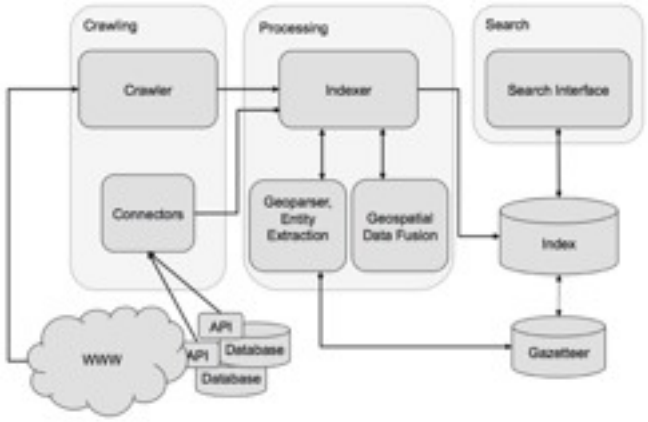

Fig. 4. Architecture of the geospatial search engine (adapted from [1])

ayudar en la mezcla de entidades nombradas de varias maneras. Incluso diferentes lugares descritos podrían ser identificados. En la mayoría de los casos, lo que el codificador geográfico o los programas extraigan son realmente referencias textuales, no las coordenadas exactas que podrían ser dibujadas en un mapa. Estas coordenadas podrían ser dibujadas en un mapa. Para lograr este paso, un codificador geográfico debe ser utilizado.

Existen algunos servicios de codificación geográfica y de diccionarios de nombres geográficos, pero a ellos también les hace falta la granularidad, lo cual solamente coloca las cosas a nivel de ciudad, algunas veces. Este problema de inexactitud en las coordenadas puede ser mejorado posteriormente usando servicios de fuentes que provean dichas coordenadas. En algunos casos, las direcciones ya son descritas por mapas integrados, los cuales permiten extraer coordenadas exactas. Para algunos otros casos, se pueden usar otras fuentes no sólo como fuentes de datos, sino como diccionarios y codificadores geográficos. Una investigación en OpenStreetMap[4] muestra que existe bastante data para Tegucigalpa y Honduras. Especialmente, la data que yace en el mapa contiene mucha más información, a menudo solamente como un punto y un nombre, pero con una coordenada geográfica bien definida. La idea entonces es usar OpenStreetMap y fuentes similares para propósitos de afinamiento y decodificación. Para nombres en otras fuentes, una medida similar puede identificar similitudes en la base de datos y entonces las coordenadas pueden ser extraídas y ser usadas para el afinamiento de entidades. Con los componentes principales definidos, la arquitectura del motor de búsqueda propuesto está graficada en la Figura 4. Toma la estructura general de un motor de búsqueda de la web, específicamente adaptado a búsqueda geoespacial y además incluye un componente para conectarse a las fuentes de datos. Se presentan los componentes principales de una araña web, conectores, parseadores geográficos y extracción de entidades, codificador geográfico y fusión de la data geoespacial; un índice y una interface terminan la arquitectura.

V. Conclusión

El motor de búsqueda geoespacial para Honduras presenta un reto bastante interesante. El primer análisis ha mostrado que existe suficiente data disponible para ser utilizada. Las pruebas iníciales y las implementaciones han sido exitosas. El trabajo realizado será continuado para mejorar el análisis detallado de las fuentes de datos y sus traslapes. Los módulos de extracción serán implementados para accesar las fuentes de datos directamente. Un motor de búsqueda será instalado con un conjunto inicial de dominios identificados para Honduras. La identificación de dominios fuera del .hn será implementado al ejecutar una araña enfocada en obtener esa información. Los análisis grafos subsecuentes de la web deberían ser capaces de recuperar patronos útiles. La investigación posterior será orientada a parseo geográfico, especialmente de descripciones y direcciones. Esto puede ser usado para proporcionar referencias de localidades a los usuarios para navegación. Un método de fusión de datos para combinar datos de diferentes fuentes será desarrollada. Nuevas metodologías serán desarrolladas para combinar texto e identidades, unido a métodos con técnicas de fusión geográfica para mezclar con confianza entidades geoespaciales. Además, diferentes interfaces para estos datos tiene que ser desarrollada. Esto será influenciado especialmente por ideas de productos e impulsos de investigación dados por el socio de la industria. Este artículo no hace frente a asuntos acerca de anchos de banda [26] o a la limitada disponibilidad de computadoras [27]. Sin embargo, bajos anchos de banda móvil o soluciones SMS [28] son considerados como interfaces hacia el motor de búsqueda.

Sobre todo, Honduras provee un campo ideal para generar un prototipo de investigación debido a los numerosos retos que requerirán la combinación de varios campos diferentes de investigación y motores de búsqueda con recuperación de información geográfica. Además, debido al tamaño pequeño del país, inclusive el prototipo de investigación será capaz de cubrir a gran fracción de la web hondureña, de esta manera se construiría un índice completo. Esto también facilitará un movimiento fácil hacia el servicio público, mejorando la cobertura de motores de búsqueda local.

4 http://www.openstreetmap.org

Agradecimientos

Este artículo es una versión traducida de un artículo previamente publicado en Inglés [29]. Un agradecimiento especial a Jorge Reyes de Unitec para la traducción y gracias a Isaac Martinez y José Matute para las contribuciones.

Referencias

[1] D. Ahlers and S. Boll, "Location-based Web search," in The Geospatial Web. How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society, A. Scharl and K. Tochterman, Eds. London: Springer, 2007.

[2] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang, "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet," International Journal of Geographical Information Science, vol. 21, no. 7, pp. 717–745, 2007.

[3] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger, "Design and Implementation of a Geographic Search Engine," in WebDB 2005, A. Doan, F. Neven, R. McCann, and G. J. Bex, Eds., Baltimore, Maryland, USA, 2005, pp. 19–24.

[4] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, A. S. D. Silva, and J. Clodoveu A. Davis, "TheWeb as a Data Source for Spatial Databases," in Anais do V Brazilian Symposium on Geoinformatics, 2003.

[5] D. Ahlers, "Local Web Search," in Web Search Engine Research, D. Lewandowski, Ed. Emerald, 2012, to appear.

[6] R. Steele, "Techniques for Specialized Search Engines," in Proceedings of Internet Computing 2001. Las Vegas: CSREA Press, 2001.

[7] D. Ahlers and S. Boll, "Retrieving Address-based Locations from the Web," in GIR '08: Proceedings of the 5th International Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2008.

[8] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and C. A. Davis, Jr., "Discovering Geographic Locations in Web Pages Using Urban Addresses," in GIR'07: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, R. Purves and C. Jones, Eds. New York, NY, USA: ACM, 2007, pp. 31–36.

[9] W. Cai, S. Wang, and Q. Jiang, "Address Extraction: Extraction of Location-Based Information from the Web," in APWeb 2005, ser. LNCS, Y. Zhang, K. Tanaka, J. X. Yu, S. Wang, and M. Li, Eds., vol. 3399. Springer, 2005, pp. 925–937.

[10] S. E. Overell and S. M. Rüger, "Identifying and grounding descriptions of places," in Proceedings of the 3rd ACM Workshop on Geographic Information Retrieval, GIR 2006, R. Purves and C. Jones, Eds. Seattle, WA, USA, Department of Geography, University of Zurich, 2006.

[11] L. A. Souza, C. A. Davis, Jr., K. A. V. Borges, T. M. Delboni, and A. H. F. Laender, "The Role of Gazetteers in Geographic Knowledge Discovery on the Web," in LA-WEB '05: Proceedings of the Third Latin American Web Congress. Washington, DC, USA: IEEE Computer Society, 2005, pp. 157–165.

[12] H. Hile, R. Grzeszczuk, A. Liu, R. Vedantham, J. Košecka, and G. Borriello, "Landmark-Based Pedestrian Navigation with Enhanced Spatial Reasoning," in Proceedings of the 7th International Conference on Pervasive Computing, ser. Pervasive'09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 59–76.

[13] D. Caduff and S. Timpf, "The Landmark Spider – Representing Landmark Knowledge for Wayfinding Tasks," in AAAI 2005 Spring Symposium. AAAI Press, 2005.

[14] D. Mundluru and X. Xia, "Experiences in Crawling Deep Web in the Context of Local Search," in GIR'08: Proceedings of the 5th International Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2008, pp. 35–42

[15] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the Deep Web," Communications of the ACM, vol. 50, no. 5, pp. 94–101, 2007.

[16] M. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic Locality of IP Prefixes," in Internet Measurement Conference (IMC) 2005, Berkeley, CA, October 2005.

[17] K. S. McCurley, "Geospatial Mapping and Navigation of the Web," in WWW '01: Proceedings of the 10th international conference on World Wide Web. New York, NY, USA: ACM Press, 2001, pp. 221–229.

[18] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice, 1st ed. Addison Wesley, February 2009.

[19] M. Najork and A. Heydon, "High-performance web crawling," in Handbook of massive data sets. Norwell, MA, USA: Kluwer Academic Publishers, 2002, pp. 25–45.

[20] D. Ahlers and S. Boll, "Adaptive Geospatially Focused Crawling," in CIKM '09: Proceeding of the 18th ACM Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2009, pp. 445–454.

[21] M. Mendoza, H. Guerrero, and J. Farias, "Inquiro.CL: a New Search Engine in Chile," in WWW in Ibero-America track, 18th International World Wide Web Conference. ACM, 2009.

[22] Universal Postal Union, “Postal addressing systems in member countries – Honduras,” Universal Postal Union, Tech. Rep., 2004.

[23] J. Leidner, Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal Publishers, 2008, PhD thesis.

[24] D. Ahlers and S. Boll, “On the Accuracy of Online Geocoders,” in Geoinformatik 2009, Osnabrück, ser. ifgiprints, W. Reinhardt, A. Krüger, and M. Ehlers, Eds., vol. 35, Münster, 2009, pp. 85–91.

[25] A. Arasu and H. Garcia-Molina, “Extracting Structured Data From Web Pages,” in SIGMOD ’03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2003, pp. 337–348.

[26] J. Chen, L. Subramanian, and J. Li, “RuralCafe: Web Search in the Rural Developing World,” in 18th International World Wide Web Conference, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 411–420.

[27] S. Amershi and M. R. Morris, “Cosearch: a system for co-located collaborative web search,” in Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ser. CHI ’08. New York, NY, USA: ACM, 2008, pp. 1647–1656.

[28] J. Chen, L. Subramanian, and E. Brewer, “SMS-Based Web Search for Low-end Mobile Devices,” in Proceedings of the sixteenth annual international conference on Mobile computing and networking, ser. MobiCom ’10. New York, NY, USA: ACM, 2010, pp. 125–136

[29] D. Ahlers, “Towards Geospatial Search for Honduras,” in Proceedings of the 3rd Latin American Conference on Networked and Electronic Media LACNEM2011. San José, Costa Rica, 2011.