Automatic Geolocation of the Scientific Knowledge: Geolocarti.

Geolocalización Automática del Conocimiento Científico: Geolocarti.

Iván Darío Ramos Vacca*, Víctor Andrés Bucheli Guerrero†
Escuela de Ingeniería de Sistemas y Computación
Universidad del Valle
Cali, Colombia.
Email:{*ivan.ramos,†victor.bucheli}@correounivalle.edu.co

***Abstract*—This work shows the state of the art of the methods for geographical location and analysis of the scientific knowledge production. The methods described are *Natural Language Processing* and *Named Entity Recognition*; *Metadata Processing*; *Bibliography and Citations Analysis*; *Scientific Collaboration Networks*; and the *Information Aggregators*. This article presents the importance of the contribution of geolocation on the analysis of knowledge creation and transference systems at hyperlocal, local and regional levels. An application that performs each one of the methods for the analysis of the scientific production based on the geographic information in Colombia is proposed in this paper.**

***Resumen*—Este trabajo muestra el estado actual de los métodos para la localización y análisis geográfico de la literatura científica. Los métodos son el *Procesamiento de Lenguaje Natural* y el *Reconocimiento de Entidades Nombradas*; el *Procesamiento de Metadatos*; el *Análisis de Citas Bibliográficas*; las *Redes de Colaboración Científica*; y los *Agregadores de información*. Se presenta la importancia de la contribución de la geolocalización en el análisis de sistemas de creación y transferencia del conocimiento en entornos hiperlocales, locales o regionales. Se propone un aplicativo que implemente cada uno de los métodos para el análisis de la información geográfica de la producción científica en Colombia.**

***Keywords*—Clasificación, Producción de Conocimiento, Sistemas de Información Geográfica, Análisis Geográfico de Información, Transferencia del Conocimiento, Procesamiento de información, Clústeres Geográficos, Conocimiento Científico.**

I. Introducción

Este documento presenta un proyecto en desarrollo, para el que se propone investigar las formas de extraer información geográfica del conocimiento científico producido en un espacio geográfico delimitado.

La geolocalización es la posibilidad que se tiene de ubicar la información en espacios geográficos *determinados* para tratarla y procesarla. Surge el interrogante ¿cómo desarrollar un sistema para entender las dinámicas de producción y transferencia del conocimiento en las regiones geográficas?. En este trabajo se aborda esta pregunta y se describen los métodos computacionales para el procesamiento de los datos que permitan obtener información partiendo de datos geolocalizados de la producción del conocimiento.

Los usos de la geolocalización automática de la producción y el conocimiento científico, han conllevado un creciente interés por utilizar la información —y el conocimiento obtenidos— como insumo en la elaboración de políticas públicas para el análisis de clústeres geográficos científicos, regionales y locales, o para evaluar la distribución y el impacto de las investigaciones realizadas [1], [2], [3]. El problema ha sido abordado desde numerosas disciplinas [4], no sólo por sus aplicaciones sino por las posibilidades escalables que brindan los trabajos —conjuntos— interdisciplinarios.

Actualmente, la geolocalización es uno de los campos con mayor impacto y crecimiento dentro de las ciencias de la computación [5], [6]. Para el caso de la producción de conocimiento científico localizado se pretende desarrollar herramientas de apoyo para las políticas públicas, que enlacen eficientemente la ciencia y el desarrollo empresarial o de nuevas tecnologías, en un espacio geográfico determinado.

La importancia de la **geolocalización del conocimiento** radica, principalmente, en la influencia que tiene en las comunidades locales la información que se produce y se recoge en éstas [7], [3], [8], las que han sido el motor de desarrollo —social, económico, empresarial y académico— [9], y las generadoras de dinámicas que permiten la interacción y configuración de planes a gran escala.

La construcción colectiva en el paradigma actual nos presenta una actividad científica descentralizada [10], donde, a pesar de la dinámica que genera Internet con las redes de colaboración para la producción científica, las barreras geográficas no se han desdibujado con los procesos globales [11]. Si bien Internet

permite que los centros de investigación alrededor del mundo colaboren hacia fines comunes, el entorno geográfico sigue siendo un factor importante que afecta la dinámica de la ciencia [12].

El proceso de diagramar la información geolocalizada, conlleva retos computacionales para el análisis de entornos sociales, políticos y académicos [13].

Con los desarrollos actuales es posible integrar en un modelo las variables que permiten un análisis transversal de las dinámicas de transferencia de conocimiento, además de observar las relaciones y vínculos para entender cada aspecto de la actividad científica y su desarrollo.

Con el desarrollo de Internet y la Web 2.0 las formas de creación y transferencia de la información y el conocimiento han cambiado notablemente. El rápido crecimiento de las comunidades virtuales, los portales de información y los medios sociales, desde principios de la década del 2000 [14], han extendiendo la noción de ubicuidad en el campo del conocimiento. Esto ha traído consigo una mayor importancia en el análisis enmarcándolo en zonas geográficas delimitadas, lo que nos obliga a entender las nuevas formas de interacción humana, los modos de comunicación, y la dinámica de producción y transferencia del conocimiento [15].

A la par con el desarrollo de las tecnologías de la información y la comunicación, han surgido varios sistemas para la búsqueda, el tratamiento y la presentación de las contribuciones científicas. El modelo contemporáneo permite almacenar la producción científica, los datos asociados a su origen, fuentes de información, áreas de estudio, fuentes de financiación, consumidores, usos y, de manera general, los *metadatos* o la información contenida en el texto que puede ser extraída para su análisis.

Estamos en un estadio de desarrollo de *hiperconexión*[16] que nos permite conocer, con inmediatez, qué está pasando en cualquier momento, en cualquier lugar y qué percepción tienen las personas de ello. El desarrollo mismo de las redes virtuales, y su crecimiento acelerado, son una muestra del interés existente, a nivel global, por los espacios de interacción en el paradigma virtual. Sin embargo, no es suficiente estar siempre conectado a los espacios de interacción, pues tenemos preferencias mediadas por el contexto geográfico local, e hiperlocal [17]. Incluso al usar Internet como herramienta de construcción colectiva, median nuestras preferencias.

Uno de las argumentos para sustentar esta propuesta de geolocalización en la producción científica, es el peso que tienen las barreras geográficas, incluso en el paradigma virtual, no se han desdibujado, como se creía [3], [18], [19]. Entonces, cobra validez una herramienta de análisis de contextos geográficos determinados, que permite extraer conocimiento relacionado con los centros de investigación y enseñanza así como de sus investigadores y docentes que aportan en la construcción del conocimiento.

El presente artículo se organiza de la siguiente manera: presenta los antecedentes y la importancia contextual conceptual del **análisis geográfico de la producción científica**, luego se revisan los modelos para la localización de artículos: En primer lugar, i) el proceso de Reconocimiento de Entidades Nombradas, mediante técnicas de Procesamiento de Lenguaje Natural, desde dos enfoques, basado en Reglas Gramaticales, y basado en Aprendizaje Automático. ii) El Procesamiento de Metadatos Anotados con un enfoque Simple y otro Compuesto. iii) Análisis de Citas bibliográficas, iv) la extracción de información de las Redes de Colaboración Científica y, finalmente, v) el trabajo realizado por los llamados *Agregadores de Información* —científica—. Por último, están las conclusiones y discusiones, donde se presenta el desarrollo actual de la investigación, para la que se ha desarrollado y adaptado algoritmos para extracción de información geográfica aplicados a SCOPUS.

II. Antecedentes y Contexto

La importancia de la localización geográfica de la producción del conocimiento y la investigación ha radicado, históricamente, en varios factores que han ido cambiando con el tiempo y han evolucionado para adaptarse a las interacciones económicas, políticas, sociales y académicas. Algo común en la época moderna, desde los siglos XVI y XVII, cuando en occidente la producción científica era exclusiva en Europa, hasta finales del siglo XX, cuando las fronteras de transmisión del conocimiento empezaron a expandirse abiertamente, ha sido la imperante relación del fomento a la investigación con el impulso al desarrollo económico e industrial regional y local [20].

Actualmente, la importancia radica en el "estrecho vínculo que existe entre la producción científica, y la innovación y los activos intelectuales como motor del crecimiento y de la competitividad en el largo plazo" [21].

El problema de localizar geográficamente la producción científica cobra especial importancia para comunidades académicas incipientes con gran proyección como el caso colombiano, permite ver el estado actual de la producción de conocimiento en cualquier momento, puede indicar qué áreas del conocimiento aportan en mayor o menor medida al desarrollo regional y local, siendo éste un factor decisivo para adoptar políticas públicas o relaciones interinstitucionales entre universidades y organizaciones para fomentar el desarrollo en áreas específicas.

Después de hacer la revisión de la literatura, los modelos en cada trabajo desarrollado siempre presentan información geográfica, aunque varíen en su arquitectura, en las herramientas para la extracción de información, en los métodos computacionales utilizados, y en las tecnologías y técnicas. Se evidencia la utilidad de una herramienta que permita hacer el proceso de forma automática.

III. Métodos computacionales para la localización geográfica de conocimiento

Una de las formas de abordar el problema de la localización de la información para su análisis es mediante **técnicas de Procesamiento de Lenguaje Natural y el *Reconocimiento de Entidades Nombradas* (NER** —por sus siglas en inglés—), enfocándose en extraer la información geográfica que contienen los artículos en forma de texto (*no estructurado*). Para el Reconocimiento de Entidades Nombradas el problema se

Cuadro I
CLASIFICACIÓN DE MÉTODOS COMPUTACIONALES PARA LA LOCALIZACIÓN GEOGRÁFICA DE ARTÍCULOS CIENTÍFICOS

Tecnología	Desarrollo	Descripción	Usos de la Tecnología
Procesamiento de Lenguaje Natural	Reconocimiento de Entidades Nombradas	Permite reconocer entidades y el uso de técnicas para desambiguarlas encontrando aquellas que se refieren a locaciones. Para el caso de la locación de literatura científica, permite extraer metadatos que no hayan sido anotados y que estén dentro del texto para usarlos como insumo en el análisis geoespacial y en la transmisión del conocimiento.	Sus usos principales son para extraer información contenida en documentos sin estructura, para extraer los datos geográficos que sean reconocidos en el procesamiento del texto y que pueda estar consignado sin ningún tipo de demarcaciones.
Anotación Semántica	Procesamiento de Metadatos Anotados	Permite hacer el procesamiento de los metadatos para reconocer ontologías, especialmente nombres de lugares desde donde se desarrollan las investigaciones para poder analizar la actividad por zonas geográficas.	Es útil para extraer la información geográfica contenida en los metadatos, los que tienen relación con la actividad investigativa con los que, comúnmente, son Universidades, Centros de Investigación o Entidades del Gobierno.
Clusterización	Análisis de Citas	Permite conocer la relación entre áreas del conocimiento y areas geográficas para el análisis bibliométrico y cienciométrico.	Correlación entre las zonas geográficas donde se producen las citas, la actividad científica y la inversión en I+D.
Clusterización	Redes de Colaboración Científica	Permiten la interacción bidireccional entre escritores y lectores para formar redes interdimensionales que pueden ubicarse geoespacialmente. Las relaciones se representan con los enlaces entre los nodos. Permite el análisis local, de un solo nodo geográfico.	Clusterización geográfica de I+D y la transferencia del conocimiento.
Reconocimiento de Patrones	Agregación de Artículos	Permite agrupar la información por algunos criterios que el usuario introduzca como parámetros para su presentación. Hay algunos agregadores de información que permiten agrupamiento por regiones geográficas. Su uso fue inicialmente en otras disciplinas y no para la segmentación geográfica, por eso es más común su implementación para agrupar artículos por áreas del conocimiento. Normalmente se hace sobre los metadatos.	La agregación se usa principalmente para presentar artículos agrupando varios clusters o clasificaciones. Funcionan con el mismo principio que los recomendadores de información y son particularmente útiles para hacer seguimiento en tiempo real de la producción del conocimiento.

enfoca, comúnmente, desde dos perspectivas: el aprendizaje estadístico, e.g., aprendizaje automático, y basados en reglas, e.g., reglas gramaticales definidas. Es normal también encontrar enfoques híbridos [22].

El análisis del comportamiento geográfico puede hacerse también **sobre los metadatos** de los artículos. Aunque no sea información geográfica directa, puede extraerse el contexto espacial y geográficos de universidades, centros de investigación, organizaciones o centros de I+D —Investigación y Desarrollo— y las entidades gubernamentales donde se desarrollan las investigaciones.

Una manera transversal de analizar la producción y transmisión del conocimiento científico es mediante **los patrones geográficos de las citas en los artículos** [12]. Es posible procesar las citas entrantes y las citas salientes: las primeras son aquellas que se hacen al construir un artículo y se citan como referencias; las segundas son aquellas que usan el artículo como un conocimiento que puede ser citado varias veces; un artículo puede contener muchas referencias diferentes, y un artículo puede ser referenciado en varios artículos [1], i.e., que citamos al hacer el artículo son las entrantes, los artículos que usan nuestra producción como referencia son las salientes.

Otros métodos se han enfocado en revisar la interacción que se genera en la transferencia del conocimiento mediante modelos de ***redes científicas de colaboración***[23], revisando los datos de los diferentes nodos que las componen y extrayendo la información geográfica más relevante en la formación de *clusteres* alrededor de los desarrollos.

Otra manera que ha sido ampliamente utilizada son los llamados agregadores de información, los que se han desarrollado para permitirnos filtrar de manera amplia, rápida y "segura"[24] las fuentes de información para presentarla agregada en algún orden.

En el Cuadro I se presentan los modelos estudiados dentro de este trabajo, los que comprenden los métodos computacionales y sus principales usos como localizadores de información. Se hace referencia a las tecnologías que se aplican en su implementación y sobre qué dominios aplica. Así como

una descripción que permite abordar de forma general cada aspecto.

IV. Reconocimiento de Entidades Nombradas

El Procesamiento de Lenguaje Natural es uno de los campos más desarrollados para la extracción de información geográfica contenida en textos. Considerando que más del 80 % de la información en Internet está de forma no estrucutrada y en crecimiento, los procesadores semánticos son uno de los campos más áctivos y que mayor crecimiento preven [25].

La información que contienen los artículos puede verse como un tipo de información semi-estructurada [26], que con el modelo apropiado, puede ser tratada para realizar la extracción de los datos geográficos contenidos [27].

Las técnicas para Reconocer Entidades tienen por tarea la Extracción de la Información, específicamente reconociendo *entidades*, que son *sujetos con significado*, que contienen nombres de personas, organizaciones, locaciones, tiempos y cantidades. Empero, la tarea no es del todo fácil y para ello se emplean técnicas supervisadas y no supervisadas; se busca distinguir y desambiguar claramente las entidades reconocidas dentro de los textos. Es común que éstas tengan más de un posible significado, el principal reto es anotar satisfactoriamente a qué entidad específica se refiere.

El problema de la desambiguación ha sido ampliamente abordado, e.g., un texto puede contener, en el mismo párrafo una entidad que se refiera a una persona y otra a una locación, sea el caso de *Magdalena*; el reto es poder distinguir claramente cuál de los dos se refiere al departamento de Magdalena y cuál se refiere al nombre Magdalena, para hacerlo, se han propuesto varios enfoques. Un enfoque es el uso de ontologías, otros enfoques son con técnicas de aprendizaje automático, y otros más con reglas gramaticales. No obstante, no es fácil probar un sistema que sea eficiente en más de un idioma o en más de un contexto. Es por esto que las plataformas más eficaces para NER tienen grandes volúmenes de datos que permiten hacer las clasificaciones con menos errores, paradójicamente, son menos eficientes por tener que hacer más comparaciones. Se tiene entonces que para el Procesamiento de Lenguaje Natural y Reconocimiento de Entidades Nombradas, enfocadas a la extracción de información geográfica, hay 3 disciplinas que son las más relevantes. a) basadas en reglas, b) modelos estadísticos de aprendizaje, y c) basadas en ontologías.

IV-A. Proceso Genérico para Resolver Locaciones

El proceso genérico de resolver las entidades geográficas, recibe varios nombres, para este documento se utiliza el *reconocimiento de toponimias*, que es la extracción de datos geográficos en textos, y la *resolución de toponimias*, asignar una locación a alguna toponimia reconocida [28], [29], [30].

En el proceso de identificar las locaciones asociadas a un texto uno de los problemas que presentan mayor dificultad, es el de resolver las ambigüedades. La primera es para saber, de las entidades identificadas, cuáles no representan y cuáles representan, efectivamente, locaciones. Éste es el problema conocido como *ambiguedad geo/non-geo*.

La segunda ambigüedad es saber si se presenta *aliasing*: cuando para referirse a un mismo sitio se usan varios nombres, por ejemplo cuando se refieren a la ciudad de Cali, y, al mismo tiempo, a la Sultana del Valle.

Por último, la ambigüedad más compleja de resolver es la llamada polisemia: se puede presentar al nombrar dos espacios geográficos distintos, con diferentes locaciones, pero con el mismo nombre, e.g., Madrid, España, y Madrid, Colombia.

Hay investigaciones que se centran en resolver las ambigüedades, como se presenta la metodología completa en [31]. Para el caso de *aliasing* [28] lo implementa para RSS de noticias. Para resolver polisemias [29] expone un trabajo que se encarga de resolver polisemias desde el Procesamiento de Lenguaje Natural. La tercera ambiguedad, *geo/non-geo*, la abordan [32] y [33], cada uno desde un enfoque diferente.

IV-B. Reconocimiento de Entidades Nombradas, Basado en Reglas Gramaticales

El enfoque basado en reglas gramaticales tiene limitaciones en términos de idioma, pues para cada idioma deben definirse las reglas para reconocer los *tokens* en los textos[34].

El *NER* basado en reglas requiere de un exhaustivo trabajo de programación y produce algoritmos, por lo general, demandantes de capacidad de procesamiento. Al procesar los textos, se extraen las toponimias que están dentro de estructuras gramaticales que pueden ser locaciones, luego se descartan aquellas que definitivamente no lo son, finalmente se busca en los *gazetteers* [35] para saber si es necesario desambiguar las entidades resultantes y analizarlas dentro de un contexto, lo que se hace con técnicas basadas en heurísticas, donde principalmente se analizan todas las toponimias encontradas en el documento y se buscan sus ubcaciones geográficas, para finalmente [33], con el contexto de las locaciones, definir a qué lugar geográfico pertenecen.

Una fuerte limitación con el enfoque basado en reglas es que los algoritmos deben crecer cada vez que se agregan nuevas restricciones, lo que lo hace inviable para prácticas en gran escala [32].

Este enfoque es usado para varios *datasets*, entre ellos: TUD-Loc-2013, CLIR-WSD, el desarrollado por Wing & Baldridge; y, para el análisis de noticias, que en esta disciplina es mucho más desarrollado, están GLocal y NewsStand.

IV-C. Reconocimiento de Entidades Nombradas, Basado en Aprendizaje Automático

A diferencia del enfoque anterior éste recibe documentos preprocesados anotados manualmente y con entidades desambiguadas, con los que se entrena. Tiene ciertas características que permiten clasificar las entidades de acuerdo a su cercanía con la opción más probable de locación. Los aplicativos usan, normalmente, el entrenamiento en conjunto con *gazetteers*—catálogos de nombres geográficos con información asociada escalable—los que se van modificando para agregar nuevas locaciones.

Están diseñados de forma que sólo las toponimias anotadas despues de reconocer las entidades, puedan ser consideradas realmente como toponimias, es decir, que aquellas que no sean anotadas, se consideran como ejemplos negativos para futuras iteraciones o documentos que se analicen.

Una de las diferencias con el enfoque basado en reglas es que el contexto lo extrae de otras entidades que, si no están en el *gazetteer*, se buscan en alguna base de datos de ontologías, e.g., Dbpedia Spotlight [36]. De esta manera, se van formando las instancias. [cambiarlos a referencias]

Aunque el aprendizaje automático tiene un comportamiento "opuesto" al que se presenta en el enfoque por reglas —el rendimiento y consumo de recursos tiende a disminuir entre más datos estadísticos contenga el modelo de clasificación, teniendo un punto de rendimiento máximo, como una campana de Gauss—, es necesario saber qué factores afectan la eficiencia, pues el uso del aprendizaje automático en distintos campos del conocimiento tiene desempeños diferentes.

V. Procesamiento de Metadatos

Los metadatos son los **datos** que acompañan un archivo y lo describen, es decir, son los datos que describen cada artículo científico, e.g., autores, centros de investigación, universidades, ciudades, keywords, entre otros. La importancia del Procesamiento de Metadados es que proveen la información más relevante y tienen una estructura fácil de trabajar. Sin embargo, el uso de metadatos geográficos para extraer información de los registros que contienen las publicaciones científicas no ha sido extensamente aplicado [37], lo que se ha hecho para el procesamiento de noticias.

La mayoría de metadatos no contiene información geográfica directamente sino unos datos sin geolocalizar [7]. Los metadatos contienen nombres de universidades, centros de investigación, nombres de conferencias o *keywords*, datos que deben ser procesados y localizados geográficamente para permitir su análisis, un claro ejemplo son los usos que se le han dado a JournalMap [38], un servicio web eficaz para visualizar literatura científica geolocalizada.

VI. Análisis de Citas

Partiendo de lo anterior, el análisis geográfico de las citas se ha convertido en una de las herramientas más fuertes para comprender el desarrollo geográfico contextual de la producción científica.

El comportamiento de las citaciones puede representarse gráficamente con algunos nodos y conexiones entre estos, por ejemplo, entre ciudades; para observar desde dónde se hace la transferencia del conocimiento. Pueden verse fenómenos sociales y económicos de la ciencia, tales como: el presentado en [12], donde el grado de impacto total de la inversión en investigación y desarrollo de un país crece linealmente con la cantidad de fondos dispuestos. El estudio concluye que la producción científica de un país puede alcanzar un impacto mayor que el promedio mundial sólo si el país invierte anualmente más de USD 100 000 por investigador.

Una de las plataformas más usadas para el análisis de citas es Scopus, enlaza artículos de investigación con la información de las citaciones que esté disponible —tiene en su base de datos alrededor de 55 000 000 de resgistros—, especialmente la que permite conocer en qué disciplinas o áreas del conocimiento es más citado cada artículo, incluso se da la tarea de reconocer los índices de propagación de las citas.

Los resultados de la investigación en [39], [40], muestran la importancia de la geolocalización de citas que permitan el uso de herramientas geográficas para entender la transferencia de la información y el conocimiento, i.e., los espacios geográficos como contenedores de medios para la transferencia del conocimiento, clusterizando la información grográfica. Sin embargo, el mayor valor agregado es el análisis de las redes que se forman entre autores de distintas zonas geográficas y que estudian una misma disciplina [3]. Incluso, conocer cómo fluye la información entre zonas geográficas.

Para el efecto de este trabajo se revisarán las herramientas para el manejo de la información geográfica que permitan ver "en forma de gráficas o cuadros" los años de las citas, las fuentes, los autores, su afiliación, el origen geográfico, el tipo de documento y la rama de conocimiento de la revista en que fue publicado cada artículo. Si bien son herramientas de análisis, no tienen un mayor grado de personalización para la interacción entre diferentes lugares.

VII. Redes de Colaboración Científica

Para las *redes de* ***colaboración*** *científica* se han incorporado tecnologías y herramientas que permiten un mayor alcance en su construcción, análisis y para la interacción entre autores y lectores [15].

Actualmente, las redes de colaboración científica están teniendo cambios en la forma de empoderamiento de los autores y sus lectores en el proceso de construcción de los artículos: hay un diálogo más abierto y permanente en el desarrollo de las investigaciones. Incluso, algunas instituciones, editoriales y laboratorios, están cambiando para permitir un modo de producción del conocimiento que permita aprovechar las nuevas formas de interacción e intercambio de información.

En esta construcción colectiva uno de los campos que mayor impacto ha tenido es el de las ciencias geográficas aplicadas a la formación de redes de autores y lectores de literatura científica. Cada vez es más común encontrar servicios que fomentan la interacción entre autores, y entre autores y lectores, e.g., ResearchGate, Research Connection.

La localización geográfica en redes aprovecha las técnicas más avanzadas [37] para decodificar la *Web Semántica Geoespacial*. Se construyen nodos que integran los sistemas de transferencia y sus relaciones basándose en el intercambio que se genera entre los actores, e.g., el trabajo en [41] desarrolla los análidis con la ciudad de Andalusia como un nodo. Aunque las redes de colaboración tienden a enfocarse en la interacción de autores y lectores, los estudios se han desarrollado, en mayor medida, para el análisis de las relaciones entre autores—tal vez por la mayor disponibilidad de datos—, un ejemplo es la representación gráfica que hace Olivier H. Beauchesne con la

adaptación de Scimago Labs a la base de datos de Scopus con los papers publicados entre 2008 y 2012.

El análisis geográfico basado en redes lo desarrolla [12] de manera amplia, donde explica su método. Para explicarlo por medio de un ejemplo, tomó 18 199 nodos (ciudades) y obtuvo 9 494 012 enlaces, entre los que se incluían 14 447 en la misma ciudad. Luego se calculan los valores totales absolutos que representa cada nodo frente a los demás, para lo que se diferencian los enlaces entrantes, los enlaces salientes, cuáles son de investigadores y cuáles de lectores.

A partir de este enfoque, nos parece interesante plantear un sistema que permita analizar los datos que corresponden a alguna ciudad o un territorio, para visualizar cómo se comporta dicha ciudad sin tener en cuenta la transferencia de conocimiento fuera de la frontera geográfica definida, también para poder extraer patrones hiperlocales dentro de una región geográfica específica. De esta forma, se puede saber si el desarrollo de conocimiento en una ciudad es por la colaboración interna o por la colaboración con otros nodos, hasta ahora, la mayoría de los estudios demuestran una correlación positiva entre las áreas en desarrollo y la colaboración local para transferir el conocimiento.

Cuadro II
PREGUNTAS

Pregunta	Método de Resolución	Modelo de Implementación
¿Qué vínculos virtuales hay entre autores de una misma región?	Localización geográfica de autores y Correlación entre citas.	Dinamizar la creación o transferencia del conocimiento hacia fines comunes.
¿Cómo descubrir qué autores en un entorno geográfico influye en la creación de nuevo conocimiento?	Redes de colaboración y Agregación de artículos.	Definir políticas de apoyo a centros de investigación o universidades.
¿Cómo conocer qué entidades financian los proyectos?	Procesamiento de metadatos.	Búsqueda de autores o centros para colaborar en investigaciones.
¿Qué correlación hay entre clústeres geográficos?	Procesamiento de metadatos y Clusterización.	Colaboración entre clústeres distantes que investiguen temas en común.
¿Qué patrones en espacios geográficos a través del tiempo se pueden identificar?	Procesamiento de metadatos y Agregación de artículos.	Medición del impacto de políticas públicas.

VIII. AGREGACIÓN DE ARTÍCULOS

Los *agregadores* son, generalmente, SaaS que toman información de múltiples fuentes para mostrarla en un solo lugar. Es un concepto simple en la teoría, en la práctica se han hecho múltiples implementaciones [42]. En todo caso, los agregadores de información se caracterizan por permitir al usuario escoger qué información quiere recibir, normalmente organizada por fuente, por tema, por disciplina o por área del conocimiento. Ahora están los llamados *specialty aggregators* para las noticias en línea, que permiten agrupar la información por locaciones geográficas; enfocados hacia la literatura científica, están: Ingenta Connect, y Information Express [1].

Inicialmente, la agregación de información se hizo en el campo de las noticias, el que actualmente presenta un mayor desarrollo[43], [44], [45]. Con el dinamismo que trajo el paradigma de la Web 2.0 [46], [15] —y el que está trayendo la Web 3.0— y las nuevas formas de compartir información y transferir el conocimiento, se desarrollaron también nuevos servicios y nuevas tecnologías adaptadas al campo de la literatura científica, e.g. EBSCOhost. Algunos agregadores permiten filtrar la información por regiones geográficas, e.g., *Web of Science* permite el uso de un servicio que toma la *dirección de publicación* para agregar los artículos por regiones geográficas [47]. Un trabajo parecido, desarrolla [48] para analizar el contenido de microblogs en redes sociales.

IX. DISCUSIÓN

Este trabajo presenta una revisión de literatura y de modelos propuestos para economías fuertes en I+D, que permiten la transferencia y creación de conocimiento, innovación y desarrollo entre sus actores [49], [50]. Para el caso colombiano, donde el sistema de Ciencia y Tecnología está en una etapa de consolidación y fortalecimiento; se evidencian cifras de inversión menores frente a países de la región como Brasil, México o Argentina [21].

Una herramienta de geolocalización automática de la producción y la transferencia del conocimiento para Colombia, es algo necesario para la consolidación de un sistema integrado de Ciencia, Tecnología y Transferencia del conocimiento mediante la toma de decisiones informadas. Por ello, es interesante un enfoque que posibilite identificar espacios generadores de conocimiento científico, los que son considerados fuentes de desarrollo intangibles para economías sustentables [51].

Surgen preguntas sobre el proceso de creación y transferencia del conocimiento, y la posibilidad de encontrar respuestas en las representaciones de los datos. Para el caso colombiano, por ejemplo, es válido responder a ¿Cómo descubrir patrones geográficos y cómo influyen en el proceso de transferencia del conocimiento? pues es un país con condiciones distintas a las del resto de la región [21].

En general, es importante el uso que puede darse a la información geográfica. Explorar sus aplicaciones en otros campos como las ciencias políticas, la economía, la gestión del conocimiento, la innovación y el desarrollo, el estudio cienciométrico de la producción científica, o incluso para generar mecanismos de apropiación de técnicas de investigación que permitan un mayor aprovechamiento de los recursos. Aunque hay más preguntas en el Cuadro II, estás son de un alcance mayor al del proyecto.

IX-A. Geolocarti

Geolocarti es el prototipo de un aplicativo que agrupa varios métodos para la extracción de información y conocimiento

geográfico. Es un sistema de Geolocalización de Artículos Científicos.

Geolocarti tiene tres fases. El 1) inicio, es la preparación de un *corpus* con artículos científicos en formato PDF y sus metadatos.

Luego viene la fase de 2) Extracción de Datos y Geolocalización. Mediante PLN y NER es posible extraer los temas y entidades geográficas localizadas. Con el procesamiento de metadatos se recuperan otros datos como el año, autor, ciudad, *keywords*, entre otros. Por último, en esta fase está la construcción de redes de coautoria y colaboración, para lo que deben extraerse las citas y referencias.

Finalmente, está la fase de 3) Extracción de Información y Conocimiento. Se identifican clústeres geográficos y sus dinámicas, se diagraman las redes de autores o el análisis de las citas, se agrega la información por atributos geográficos para su visualización, y se identifican vínculos hiperlocales y redes entre organizaciones en un mismo espacio geográfico.

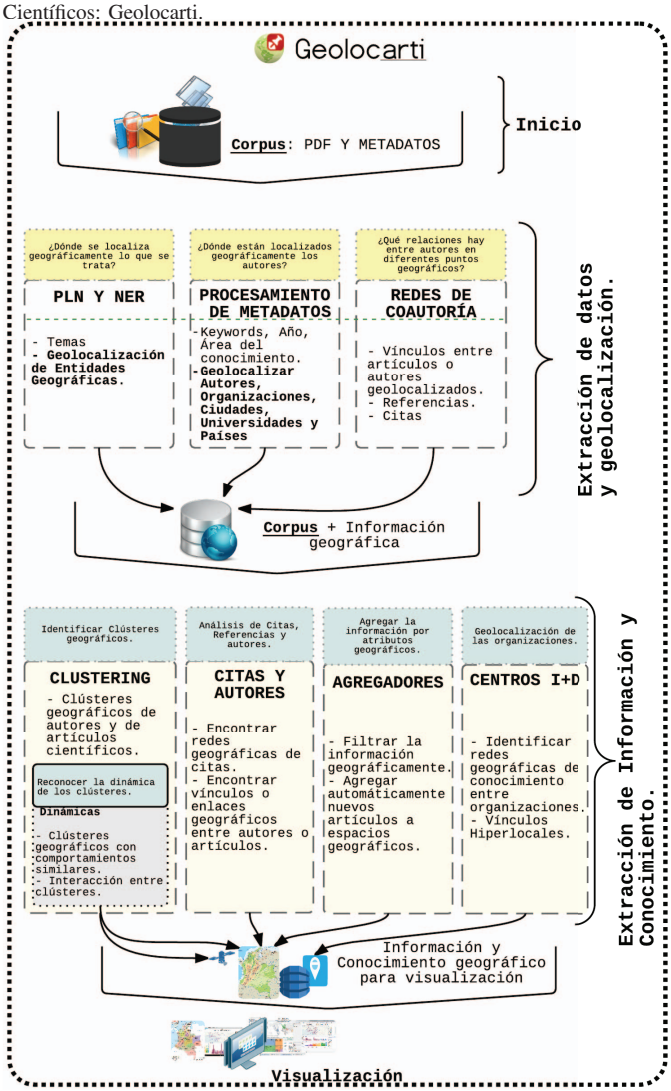
Los datos de la última fase salen procesados para ser visualizados. La visualización permite identificar tendencias de crecimiento en número de autores en espacios geográficos, especialización de las regiones, formación de clústeres, o interacción entre centros de producción de conocimiento dentro de un mismo espacio geográfico delimitado.

X. Conclusión

Al presentar la revisión de literatura sobre los métodos para la extracción de información geográfica relacionada con el quehacer científico. Identificamos una pregunta de investigación que puede resolverse aplicando los métodos revisados. Por ello, se propone una herramienta que pueda extraer, de forma automática, la información geográfica relacionada con la producción de conocimiento y la literatura científica.

La localización geográfica de artículos y de literatura científica, reviste importancia no sólo para las comunidades académicas, los centros de I+D, los gobiernos, o las empresas. Podrían parecer actores que inciden sobre los factores que determinan hacia dónde enfocar el desarrollo en este campo, sin embargo, es necesario pensar un modelo para integrarlos. Si bien las integraciones a nivel global son complejas, puede pensarse investigar hasta qué nivel es posible integrar zonas geográficas que se enfoquen en desarrollar las mismas áreas del conocimiento, revisando la producción científica que tienen y hacia dónde se enfoca el trabajo futuro. Incluso, puede retomarse una pregunta que ha sido recurrente —para aplicarla en el caso colombiano en vez de hacer a nivel global— y es saber qué tan factible se hace interactuar entre clústeres de desarrollo alrededor de áreas del conocimiento, sólo que enfocándolos a la interacción en áreas geográficas determinadas.

Figura 1. Proceso para Extracción de Información Geográfica con Artículos Científicos: Geolocarti.



Referencias

[1] C. Ni, D. Shaw, S. M. Lind, and Y. Ding, "Journal impact and proximity: An assessment using bibliographic features," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 4, pp. 802–817, 2013.

[2] B. Bozeman, H. Rimes, and J. Youtie, "The evolving state-of-the-art in technology transfer research: Revisiting the contingent effectiveness model," *Res. Policy*, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.respol.2014.06.008

[3] P. Maskell, "Knowledge Creation and Diffusion in Geographic Clusters," *Int. J. Innov. Manag.*, vol. 05, no. 02, pp. 213–237, 2001.

[4] H. Inoue, K. Nakajima, and Y. Saito, "Localization of collaborations in knowledge creation," Research Institute of Economy, Trade and Industry (RIETI), Discussion papers, 2013. [Online]. Available: http://EconPapers.repec.org/RePEc:eti:dpaper:13070

[5] G. Bordogna, G. Ghisalberti, and G. Psaila, "Geographic information retrieval: Modeling uncertainty of user's context," *Fuzzy Sets Syst.*, vol. 196, pp. 105–124, Jun. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165011411001679

[6] L. Bornmann, L. Leydesdorff, C. Walch-Solimena, and C. Ettl, "Mapping excellence in the geography of science: An approach based on Scopus data," *J. Informetr.*, vol. 5, no. 4, pp. 537–546, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.joi.2011.05.005

[7] J. W. Karl, J. E. Herrick, R. S. Unnasch, J. K. Gillan, C. Erle, W. G. Lutters, L. J. Martin, and E. C. Ellis, "Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching," *Bioscience*, vol. 63, no. 8, pp. 674–682, 2013. [Online]. Available: http://www.jstor.org/stable/info/10.1525/bio.2013.63.8.10

[8] M. Acosta, D. Coronado, E. Ferrándiz, and M. D. León, "Regional Scientific Production and Specialization in Europe: The Role of HERD," *Eur. Plan. Stud.*, no. November 2014, pp. 37–41, 2012. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84871245694&partnerID=40&md5=6d83a673fe2a303a8bbb711d3b9074f0

[9] J. R. McNeill and W. H. McNeill, *The Human Web: A Birds–eye View of World History*. New York: W W Norton & Co, 2003.

[10] M. Grossetti, D. Eckert, Y. Gingras, L. Jégou, and V. Larivière, "The Geographical Deconcentration of Scientific Activities (1987-2007)," *Proc. 17th Int. Conf. Sci. Technol. Indic.*, vol. 1, pp. 348–356, 2012. [Online]. Available: http://sticonference.org/Proceedings/vol1/Grossetti_Geographical_348.pdf

[11] S. Hennemann, D. Rybski, and I. Liefner, "The myth of global science collaboration-Collaboration patterns in epistemic communities," *J. Informetr.*, vol. 6, no. 2, pp. 217–225, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.joi.2011.12.002

[12] R. K. Pan, K. Kaski, and S. Fortunato, "World citation and collaboration networks: uncovering the role of geography in science." *Sci. Rep.*, vol. 2, p. 902, 2012. [Online]. Available: http://www.nature.com/srep/2012/121129/srep00902/full/srep00902.html

[13] H. Nowotny, P. Scott, and M. Gibbons, "'Mode 2' Revisited: The New Production of Knowledge," pp. 179–194, 2003.

[14] A. Nurwidyantoro and E. Winarko, "Event detection in social media: A survey," *ICT for Smart Society (ICISS), 2013 International Conference on*, pp. 1–5, 2013.

[15] O. Ahlqvist, F. Harvey, H. Ban, W. Chen, S. Fontanella, M. Guo, and N. Singh, "Making journal articles 'live': Turning academic writing into scientific dialog," *GeoJournal*, vol. 78, no. 1, pp. 61–68, 2013.

[16] J. Q. Anderson and L. Rainie, "Millennials will benefit and suffer due to their hyperconnected lives," *Pew Research Center*, 2012.

[17] M. Vähämaa and M. D. West, "The dilemma of group membership in the internet age: Public knowledge as preferred misinformation," *Javnost*, vol. 21, no. 1, pp. 5–18, 2014.

[18] K. J. Borowiecki, "Geographic clustering and productivity : An instrumental variable approach for classical composers," *Journal of Urban Economics*, vol. 73, no. 1, pp. 94–110, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jue.2012.07.004

[19] G. Buenstorf and A. Schacht, "We need to talk – or do we ? Geographic distance and the commercialization of technologies from public research," *Research Policy*, vol. 42, no. 2, pp. 465–480, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.respol.2012.06.010

[20] M. Davis, *SCIENTIFIC PAPERS AND PRESENTATIONS*, Elsevier, Ed. Academic Press, 2005, vol. 8, no. 1.

[21] RedEmprendia, Universia, and CINDA, "La Transferencia de I+D, La Innovación y el Emprendimiento en las Universidades," Senén Barro. Coordinador, Santiago, Chile, Tech. Rep., 2015.

[22] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, "NewsStand: A new view on news," *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, p. 18, 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1463434.1463458

[23] S. Hennemann, "Evaluating the Performance of Geographical Locations Within Scientific Networks Using an Aggregation—Randomization—Re-Sampling Approach (ARR) Stefan," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 12, pp. 2393–2404, 2012.

[24] J. Nagy and P. Pecho, "Social networks security," in *Emerging Security Information, Systems and Technologies, 2009. SECURWARE '09. Third International Conference on*, June 2009, pp. 321–325.

[25] E. Tsui, W. Wang, L. Cai, C. Cheung, and W. Lee, "Knowledge-based extraction of intellectual capital-related information from unstructured data," *Expert systems with Applications*, vol. 41, no. 4, pp. 1315–1325, 2014.

[26] M. Stonebraker and J. Hellertein, "What goes around comes around," *World Sport. Act.*, vol. 8, no. 1, pp. 19–22, 2002.

[27] H. Tveite, "Data Modelling and Database Requirements for Geographical Data," Ph.D. dissertation, Agricultural University of Norway, 1997.

[28] M. D. Lieberman, "Multifaceted Geotagging for Streaming News," p. 258, 2012.

[29] J. L. Leidner and M. D. Lieberman, "Detecting geographical references in the form of place names and associated spatial natural language," *SIGSPATIAL Spec.*, vol. 3, pp. 5–11, 2011.

[30] J. N. E. K. Geoffrey Andogah, Gosse Bouma, "Geographical Scope Resolution." [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.1305

[31] N. Xia, S. Miskovic, M. Baldi, A. Kuzmanovic, and A. Nucci, "GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic," *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 1–8, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6927600

[32] P. Katz, "To Learn or to Rule : Two Approaches for Extracting Geographical Information from Unstructured Text," no. Ml, 2012.

[33] E. Amitay, E. Amitay, N. Har'El, N. Har'El, R. Sivan, R. Sivan, A. Soffer, and A. Soffer, "Web-a-where: geotagging web content," *Proc. SIGIR '04 Conf. Res. Dev. Inf. Retr.*, pp. 273–280, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1009040

[34] E. Klien, "A rule-based strategy for the semantic annotation of geodata," *Trans. GIS*, vol. 11, no. 3, pp. 437–452, 2007.

[35] J. Partyka, P. Parveen, L. Khan, B. Thuraisingham, and S. Shekhar, "Enhanced geographically typed semantic schema matching," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 9, no. 1, pp. 52–70, Mar. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570826810000909

[36] DBPedia, *DBPedia Spotlight*, 2015 (último acceso: 10 de julio de 2015), spotlight.dbpedia.org/.

[37] M. Bidney and K. Clair, "Harnessing the Geospatial Semantic Web: Toward Place-Based Information Organization and Access." *Cat. Classif. Q.*, vol. 52, no. 1, pp. 69–76, 2014.

[38] T. O. Firm, *Journal Map. Research, Reimagined.*, 2013 (último acceso: 13 de agosto de 2015), www.journalmap.org.

[39] L. Bornmann, M. Stefaner, F. D. M. Anegón, and R. Mutz, "Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers : A visualization of results from multi-level models," *Online Inf. Rev.*, vol. 38, no. 1, pp. 43–58, 2013.

[40] M. M. Force and N. J. Robinson, "Encouraging data citation and discovery with the Data Citation Index," *J. Comput. Aided. Mol. Des.*, pp. 1043–1048, 2014.

[41] Z. Chinchilla-Rodríguez, C. López-Illescas, and F. D. Moya-Anegón, "Biomedical scientific publication patterns in the Scopus database: a case study of Andalusia, Spain," *Acimed*, vol. 23, no. 2, pp. 219–237, 2012. [Online]. Available: http://hdl.handle.net/10261/63796

[42] K. Isbell, "The Rise of the News Aggregator: Legal Implications and Best Practices," *SSRN eLibrary*, vol. 7641, 2010. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1670339

[43] A. Kavanaugh, A. Ahuja, S. Gad, S. Neidig, M. A. Pérez-Quiñones, N. Ramakrishnan, and J. Tedesco, "(Hyper) local news aggregation: Designing for social affordances," *Gov. Inf. Q.*, vol. 31, no. 1, pp. 30–41, Jan. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0740624X13001251

[44] M. Personal, R. Archive, S. Siew-ling, A. Mansor, and V. Khim-sen, "Discussion of "Local News Online: Aggregators, Geo-Targeting and the Market for Local News"," 2012.

[45] A. M. Lee and H. I. Chyi, "The Rise of Online News Aggregators: Consumption and Competition," *Int. J. Media Manag.*, no. January, pp. 1–22, 2015. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14241277.2014.997383

[46] P. Gardois, N. Colombi, G. Grillo, and M. C. Villanacci, "Implementation of Web 2.0 services in academic, medical and research libraries: A scoping review," *Health Info. Libr. J.*, vol. 29, no. 2, pp. 90–109, 2012.

[47] L. Leydesdorff and I. Rafols, "Interactive overlays: A new method for generating global journal maps from Web-of-Science data," *J. Informetr.*, vol. 6, no. 2, pp. 318–332, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.joi.2011.11.003

[48] J. W. Crampton, M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. Wilson, and M. Zook, "Beyond the Geotag ? Deconstructing " Big Data " and Leveraging the Potential of the Geoweb," pp. 1–29.

[49] H. Inoue, K. Nakajima, and Y. Saito, "Localization of Collaborations in Knowledge Creation," *RIETI Discussion Paper Series*, no. 13-E-070, 2013.

[50] K. Firlej, *Knowledge transfer and diffusion of innovation*.

[51] V.-R. López-Ruiz, J.-L. Alfaro-Navarro, and D. Nevado-Peña, "Knowledge-city index construction: An intellectual capital perspective,"

Expert Systems with Applications, vol. 41, no. 12, pp. 5560 – 5572, 2014, {EMPIRICAL} {APPROACHES} {IN} {KNOWLEDGE} {CITY} {RESEARCH}. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414000700