Análisis semántico del lenguaje natural para expresiones geotemporales

Willington Siabato[1], Alberto Fernández Wyttenbach[1], Bruno Martins[2], Miguel Ángel Bernabé[1], Mabel Alvarez[1]

[1] Universidad Politécnica de Madrid - LatinGEO
Autovía de Valencia Km 7.5 Campus Sur UPM, 28031, Madrid, España
wsiabato@acm.org a.fernandez@topografia.upm.es
ma.bernabe@upm.es mablop@speedy.com.ar

[2] Universidad Técnica de Lisboa - Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
bruno.g.martins@ist.utl.pt

Resumen

En el contexto de bibliotecas y cartotecas digitales, los recursos son generalmente descritos en registros de metadatos que definen su objeto, ubicación, período de tiempo, formato y palabras clave. En lo que se refiere a ubicación y periodos de tiempo, los registros son generalmente insuficientes o proporcionan la información de una forma que no es comprensible por los sistemas informáticos (descripciones textuales). Este artículo presenta técnicas para la extracción de información geotemporal de colecciones de texto, utilizando técnicas básicas de minería de datos apoyadas en servicios de *Gazetteer*. El objetivo es partir de referencias textuales geotemporales descritas por humanos, identificar las entidades geográficas y temporales y expresarlas en un lenguaje comprensible y procesable por un sistema informático. Un prototipo es desarrollado para poner a prueba los conceptos presentados. Resultados experimentales demuestran la eficiencia y la exactitud de los enfoques propuestos.

Palabras clave: Análisis semántico, *geo-parser*, *gazetteer*, servicios geográficos, DIGMAP.

1 Introducción

El análisis semántico y la correcta interpretación por parte de los sistemas informáticos del lenguaje natural, es un tema de creciente estudio en la actualidad

que se remonta a décadas de arduo trabajo. En este contexto, el adecuado análisis de las expresiones geotemporales es un tópico que debido a su complejidad computacional debe ser analizado de forma independiente y en el que al estar implicadas sencillas expresiones que son utilizadas por los seres humanos habitual y cotidianamente, resultan de fácil entendimiento para ellos pero no para los sistemas informáticos.

El conocimiento de contextos, periodos, fechas puntuales, hechos de referencia o extensión de una cultura, son algunos de los elementos que nos permiten (a los humanos) realizar una lectura rápida de un antes y un después, con un alto grado de comprensión y exactitud al comunicarnos. Sin embargo, el sistema computacional no cuenta con este raciocinio vinculante, y por tanto no dispone de la capacidad de interpretar el significado real de este tipo de expresiones. El hecho de que de forma implícita el oyente cambie la escala temporal al escuchar frases como: "*En los inicios del Pleistoceno"* (y en la que se entiende que se está hablando de un periodo comprendido entre 50.000 o 100.000 años tras el inicio de esta época del periodo Cuaternario) o "*En los inicios del Románico"* (referido a un entorno de 50 años), lleva a introducir bases de datos de conocimiento (o algún otro método informático) que acoten la duración y geolocalización de los periodos o hechos históricos, dentro del propio razonamiento de la máquina.

Este trabajo presenta métodos automáticos para la extracción de información geotemporal de textos en lenguaje natural, usando para ello técnicas como la minería de datos apoyada sobre servicios de *gazetteer* [22]. Las técnicas que se proponen son evaluadas a través de comparaciones con una colección de textos, cuyos ítems contienen un contexto geotemporal determinado por múltiples perfiles y usuarios. Las colecciones evaluadas provienen de los registros de metadatos accesibles desde de la cartoteca digital del proyecto DIGMAP[1], que proporciona un rico acceso a anotaciones temporales y geográficas consignadas en su mayoría por bibliotecarios y cartotecarios expertos.

Se expone cómo apoyándose en diferentes proyectos como WordNet[2], GeoNames[3] y elementos teóricos como el Algebra de Allen [1], es posible ir descomponiendo la expresión introducida por el usuario, clasificarla en subconjuntos hasta llegar a encontrar el significado de cada término y transformarlo en lenguaje de máquina.

Se implementa la técnica propuesta para interpretar las expresiones de las consultas realizadas y convertirlas en lenguaje de máquina: salidas en formato XML (a través

---

[1] http:// www.digmap.eu

[2] http://wordnet.princeton.edu

[3] http://www.geonames.org

de las cuales será posible consultar los índices, catálogos y demás componentes del sistema), retornando los elementos que sean coincidentes con la consulta para su posterior visualización. El prototipo informático, desarrollado en el contexto del proyecto DIGMAP, demuestra la efectividad de las técnicas propuestas. Se dejan entrever algunos pasos intermedios del proceso para su mejor comprensión y la salida final de expresiones geotemporales en formato XML.
Los conceptos aquí expuestos pueden ser extrapolados a las Infraestructuras de Datos Espaciales –IDE– desde múltiples puntos de vista: consultas de catálogos, consulta de datos en diferentes fuentes y repositorios, consultas correlacionadas por significado y relaciones espaciales. Cada uno puede ser tratado de forma independiente enriqueciendo el espacio de trabajo de las IDE con herramientas que hagan más efectivos los métodos de consulta, facilitando a los usuarios el acceso a datos y ofreciendo mayor riqueza desde el punto de vista semántico.

2 Conceptos y trabajos relacionados

Múltiples estudios [3] [6] [15] muestran como los componentes temporal y geográfico desempeñan un importante rol al filtrar, agrupar y dar prioridad a recursos de información; motivando de este modo la investigación en métodos para transformar referencias geotemporales descritas por humanos en formas y representaciones comprensibles por los ordenadores y otros sistemas de procesamiento digital de información.
La extracción automática de información geotemporal está relacionada con los procesos de (i) identificación y análisis del texto, (ii) la búsqueda de elementos y referencias geográficas y temporales y su precisa ubicación sobre la superficie y el espacio, y (iii) la combinación de estas referencias en recopilaciones semánticas significativas, tales como el contexto de búsqueda indicado. El texto analizado puede provenir de páginas Web, de contenidos registrados en sistemas de gestión de contenidos, o de metadatos previamente estructurados.
Aunque el problema presentado ha sido tratado con relativo éxito en comunidades como la *Natural Language Processing* [17] y la *Geographical Information Retrieval* [15] [16], el asunto discutido en este artículo difiere del NER –*Named Entity Recognition*– convencional en los siguientes elementos:

- Los tipos de las entidades analizadas (ciudades, villas, provincias) son más precisas que los tipos generales que son considerados habitualmente (persona, organización o ubicación).
- Los documentos analizados están en múltiples idiomas y es posible que se deba solventar problemas con aquellos lenguajes para los que las anotaciones en el corpus del documento resultan insuficientes. En el caso

de minería de datos aplicada a NER en la mayoría de los casos ha sido realizada teniendo en cuenta el Inglés como único idioma

- El propio reconocimiento de las entidades no implica un significado en sí mismo, por lo que es necesario ajustarlas explícitamente a una ubicación espacial y temporal (usando por ejemplo registros de los *gazetteers*). Al extender los procesos NER con asociaciones a un *gazetteer* implica nuevos problemas y retos que van más allá del simple reconocimiento de elementos. [23]
- La manipulación de grandes colecciones requiere procesar las fuentes de información individuales en un tiempo razonable, haciéndose necesaria la selección de técnicas y heurísticas apropiadas. El desempeño de las aplicaciones ha sido frecuentemente dejado de lado en los estudios previos.
- Las entidades a reconocer en un texto pueden ser vistas como unidades individuales o como parte de un contexto semántico específico. Éstas deben ser combinadas en agrupaciones semánticas significativas (el alcance geotemporal de cada aplicación) teniendo en cuenta las posibles relaciones que existan entre ellas.

Los sistemas NER tradicionales combinan fuentes léxicas (como los *gazetteers*) con operaciones de procesamiento secundarias, en las que se incluyen por lo menos una semilla, un diccionario especializado y un conjunto de reglas de extracción. Las reglas para el reconocimiento de entidades son el núcleo del sistema, es posible por ejemplo combinar los nombres de elementos en el diccionario con características como las mayúsculas y textos contiguos o circundantes. Estas reglas pueden ser generadas manual o automáticamente, aplicando técnicas de inteligencia artificial. Los métodos convencionales confían en expertos, mientras que los más recientes infieren reglas a partir de anotaciones manuales de entrenamiento.

El mejor sistema de aprendizaje desarrollado alcanza resultados superiores al 90% en columnas de noticias, sin embargo requiere de textos debidamente balanceados y con un corpus representativo [27]. En las pruebas realizadas, se observó que se presenten cuellos de botella cuando los datos no están fácilmente disponibles, este caso es muy común con idiomas diferentes al inglés o en tareas muy específicas como el reconocimiento de referencias geotemporales altamente detalladas.

El grado en el que los *gazetteers* ayudan en la identificación de entidades también varía, mientras algunos estudios concluyen que su uso no implica una mejora del rendimiento [21], otros muestran mejoras significativas [9]. En lo que a entidades geográficas se refiere, Mikheev *et al.* muestran como un sistema NER sin el uso de un lexicón puede tener resultados aceptables para la mayor parte de los criterios pero no para lugares [24]. El mismo estudio demuestra que el *gazetteer* que debe

utilizarse no debe ser muy avanzado para obtener resultados y rendimiento razonables.

Una importante conclusión de la CoNNL-2003[4] es que la ambigüedad en referencias geográficas es bidireccional. El mismo nombre puede ser utilizado para más de una ubicación (ambigüedad referente), y la misma ubicación puede tener uno o más nombres (ambigüedad referencial). Además, es posible que el mismo nombre sea utilizado para ubicaciones y otras clases de entidades, tales como personas o nombres de compañías (ambigüedad de clase referente). Garbin [12] estima que más del 67% de las referencias de lugar en un texto son ambiguas. Harpring [13] por su parte indica que el porcentaje de los nombres de lugares que son utilizados por más de un lugar varían entre el 16.6% para Europa y 57.1% para América del Norte y Centro.

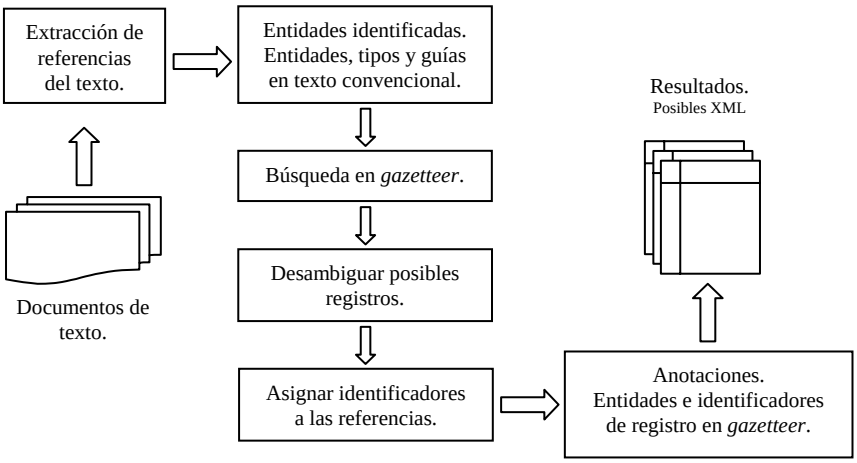

Figura 1. Aproximación típica para el análisis semántico de un texto.

Los trabajos presentados en el marco de [17] abrieron el camino hacia el desarrollo de técnicas para la exploración de referencias sobre la ubicación de sitios y lugares en los textos, centrándose en tareas más complejas que el simple reconocimiento. Algunos de los sistemas presentados orientaron sus esfuerzos para desambiguar plenamente las referencias de lugares (como el análisis semántico) aunque sólo experimentos básicos fueron presentados. La arquitectura habitual para estos sistemas es heredada de los NER, adicionando los elementos necesarios para

[4] Conference on Computational Natural Language Learning

direccionar el aparejamiento de entidades con los registros de un *gazetteer*. Es posible presentar una aproximación para el análisis semántico como el mostrado en la *Figura 1*.

Con el objetivo de encontrar el correcto sentido de una entidad y asignar una interpretación válida a las entidades geográficas, los sistemas generalmente hacen uso de heurísticas, las tenidas en cuenta en este trabajo son [20]:

- Un referente por discurso: una referencia geográfica ambigua probablemente significará uno y sólo uno de sus sentidos cuando es usada múltiples veces dentro del contexto de un discurso (ej. el mismo documento). Esta heurística es similar la propuesta para desambiguar el sentido de las palabras: un sentido por discurso [11].
- Referentes relacionados por discurso: múltiples referencias geográficas que aparecen en el mismo contexto de un discurso tienden a indicar ubicaciones cercanas.
- Sentido predeterminado: un sentido por defecto puede ser asignado a referencias ambiguas teniendo en cuenta la importancia del lugar. Un lugar importante es más sensible de ser referenciado que otro. (ej. Es más probable que Bogotá se refiera a la ciudad y no a una calle)

La investigación con analizadores semánticos geográficos (*geo-parsers*) está iniciando. Un buen ejemplo de ello es el presentado en [19], pero si se compara con la bibliografía disponible en NER resulta insuficiente. Diferentes combinaciones de las tres heurísticas arriba mencionadas han sido probadas [12] [19], pero los resultados son no equiparables, los sistemas presentados varían en tipos de clasificación y en el rendimiento a la hora de desambiguar las entidades; las fuentes para la evaluación no son tampoco comparables ni consistentes [8] [19].

Buscando la interoperabilidad de los servicios, el OGC[5] ha propuesto un sencillo servicio Web para *geo-parsers* y el reconocimiento de entidades geográficas. No obstante, este documento está actualmente discontinuado [18]. Aunque identifica un completo conjunto de detalles a tener en cuenta en la interfaz del servicio, no muestra detalles sobre la implementación del servicio. SpatialML[6] es otra reciente propuesta para la interoperabilidad entre sistemas *geo-parsing*, enfatizando la necesidad de una evaluación estándar de fuentes y recursos de información. El prototipo desarrollado en el proyecto DIGMAP ha utilizado un *geo-parser* similar al propuesto por el OGC, con algunas extensiones relativas a elementos temporal y

---

[5] Open Geospatial Consortium - http://www.opengeospatial.org

[6] http://sourceforge.net/projects/spatialml

a la asociación de sitios, lugares y ubicaciones con sus respectivas coordenadas espaciales.

Otros trabajos han centrado esfuerzos en la combinación de referencias a sitios dadas en un texto con el objeto de identificar repeticiones e identificar el foco geográfico que el documento puede contener, analizando el texto como un todo. Uno de ellos es *Web-a-Where* [2], en el que se propone identificar el interés geográfico de páginas Web usando parte de las relaciones descritas en un *gazetteer*, por ejemplo, Lisboa hace parte de Portugal, y los documentos que referencien estos lugares tengan probablemente como alcance Portugal. Iterando sobre un conjunto de referencias geográficas no ambiguas, *Web-a-Where* incluye para cada página analizada el nivel de importancia de los múltiples niveles que componen el árbol de jerarquías del *gazetteer*. Estos niveles taxonómicos son posteriormente organizados por peso eliminando aquello que no se encuentren dentro de un rango determinado, los que cumplan son tenidos en cuenta como el foco geográfico de la página. En páginas Web de la ODP[7] los algoritmos del proyecto descrito han descrito en su orden el continente, país, ciudad y el alcance exacto con una precisión del 96, 93, 32 y 38% respectivamente. Métodos más avanzados también ha sido desarrollados [24], pero sacrificando rendimiento y necesitando más capacidad de procesamiento.

En lo relativo a entidades o referencias temporales, algunos informes se han ocupado de la vinculación de eventos con el tiempo y de su clasificación y orden [4] [12]. Similar a los casos de lugares presentados, existen sistemas adecuados para especificar tiempo y rangos (como calendarios), pero usualmente se hace uso de entidades ambiguas [5]. La ambigüedad en referencias temporales es quizá un reto mayor que el supuesto para lugares y entidades geográficas, especialmente para aquellas aplicaciones que requieren de un alto detalle en anotaciones temporales y geotemporales; por ejemplo, la Pascua implica diferentes fechas para las iglesias católicas y ortodoxas, el invierno depende del hemisferio, etc. El trabajo descrito en [4] presenta una aproximación para el análisis avanzado del tiempo, con la flexibilidad suficiente para satisfacer las necesidades de motores de razonamiento avanzados. Este trabajo se ha apoyado en *TimeML*, un estándar emergente para anotaciones temporales basado en XML que almacena las propiedades y relaciones entre expresiones que implican tiempo [26]. No obstante, el trabajo presentado manipula las expresiones de tiempo a un nivel más simple, teniendo en cuenta nombres, fechas y periodos históricos. Se controlan las ambigüedades existentes, referenciales y referentes, del mismo modo que en las entidades geográficas y aplicando heurísticas. Otros trabajos tenidos en cuenta por su relevancia son el *ECAI TimeMap* [5] y el desarrollo de *gazetteers* geotemporales [25].

---

[7] http://www.dmoz.org

3 Minería de datos geotemporal

Esta sección presenta las técnicas propuestas para la extracción de información geográfica y temporal. Se hace una breve descripción del *gazetteer* utilizado, los algoritmos y métodos de extracción, y la descripción del *geo-parser* que implementa los conceptos mencionados.

3.1 *Gazetteer* geotemporal

Contar con un *gazetteer* fiable, robusto, con información completa y detallada, con un registro de nombres de lugares, periodos históricos y la descripción de sus propiedades (ej. tipos de lugar, coordenadas, intervalos temporales, jerarquías, nombres alternativos, asociaciones semánticas) es un factor determinante en el desarrollo y éxito del trabajo presentado.
En el marco de DIGMAP, se ha desarrollado un servicio de *gazetteer* [22] que integra datos de múltiples fuentes entre las que destacan GeoNames[3] y el directorio de periodos históricos ECAI [25]. El servicio sigue la interfaz Web XML y el modelo de datos propuesto por el *gazetteer* de la Biblioteca Digital de Alejandría – ADL[8]– [14], incorporando algunos cambios relativos a la manipulación de referencias temporales.
Como es evidente, el contenido del *gazetteer* propuesto impactará directamente en los resultados obtenidos de las pruebas realizadas puesto que se depende de él para la correcta interpretación de las referencias en el texto analizado. En particular, los métodos para desambiguar entidades dependen de las relaciones jerárquicas entre registros.

3.2 Extracción de información geográfica

Tal y como fue dicho previamente, la extracción de entidades (referencias) geográficas implica tres procesos: el reconocimiento de las referencias de lugares, desambiguar las referencias encontradas y la asignación de un entorno geográfico a las mismas.
Para el reconocimiento de referencias de lugares se utiliza el típico método NER basado en semillas complementado con búsquedas en *gazetteers*. Dependiendo de la importancia del lugar y la forma como este escrita (ej. iniciando con mayúscula)

[8] Alexandria Digital Library

es posible asignar un valor con el que se identifiquen fácilmente áreas geográficas relevantes. Para cada clase en la clasificación jerárquica del *gazetteer* se asigna un valor (puntuación) $s \in [0,1]$, de esta forma se asigna un peso a las entidades según su tipo: los continentes son más importantes que los países, los países son más importantes que las provincias, las provincias que las ciudades, etc. De este modo, una vez ejecutada la clasificación es fácil descartar aquellos elementos que no cumplan con determinado criterio, en este caso para un reconocimiento simple se descartan aquellos lugares con tipo de clase $s \leq 0.5$. Con este método se presenta ambigüedad máxima en áreas o nombres secundarios, pero es suficiente para el reconocimiento de áreas importantes y conocidas.

Para regiones más pequeñas, se ejecutan búsquedas en listas separadas en las que se encuentran registrados todos los nombres de lugares que componen el *gazetteer*. Sin embargo, no se utilizan procedimientos básicos de coincidencia sino que se pondera con los elementos (textos) circundantes que normalmente guardan un alto de correlación descriptiva si el texto intermedio es geográfico, expresiones como Municipalidad o Distrito son palabras encontradas junto a los nombres analizados, la ventana que se tiene en cuenta para este análisis es de tres palabras a izquierda y derecha; sólo se considera la referencia encontrada si está acompañada de un tipo de lugar encontrado en la ventana de búsqueda. Este método resulta de gran efectividad si se tiene en cuenta el carácter multilingüe del sistema propuesto.

Para desambiguar las referencias encontradas, se requiere del uso del *gazetteer* y un conjunto de heurísticas. Del paso anterior las referencias a pequeñas regiones cuentan ya con el tipo de lugar asociado. Para los demás nombres identificados, se sigue el mismo método y se buscan tipos de lugar en el texto circundante, aunque el reconocimiento no depende de que se encuentre alguna coincidencia. En los casos en que se tiene un tipo de lugar (jerarquía), la consulta que se hace al *gazetteer* combina el nombre y su tipo, y sólo aquellos elementos que coincidan tanto en tipo como en nombre son identificados. Para los nombres a los que no se haya asociado un tipo, simplemente se consulta el nombre exacto en el *gazetteer*.

Cuando la consulta retorna aquellos elementos coincidentes, se clasifican de acuerdo a una puntuación que refleja una heurística *Sentido predeterminado*. Para un elemento *f*, y en el caso de nombres a los que se ha asociado un tipo, la puntuación está dada por el total normalizado de los elementos hijo que para *f* son definidos en el *gazetteer;* la idea es que aquellos lugares con más subtipos son más factibles a ser referenciados. Para entidades sin tipo asociado, la calificación corresponderá al valor *s* previamente asociado ($s \in [0,1]$), de igual modo, es más probable que aquellas entidades con un valor *s* mayor sean referenciadas.

Finalmente, para resolver aquellas entidades que tienen más de una coincidencia en el *gazetteer*, se hace uso de las heurísticas *un referente por discurso* y *referentes relacionados por discurso* para intentar ajustar las calificaciones asignadas.

Aquellos registros para los que se haya encontrado un registro padre o hijo en el conjunto de todas las referencias descubiertas en el documento tienen su calificación asignada aumentada por 0,2 hasta un máximo de 1.
El resultado final de este proceso es un conjunto de referencias con una lista de posibles candidatos a ser referenciados y ordenados por la calificación asignada. Esta calificación puede ser entendida como la probabilidad de que una referencia (lugar) determinada esté relacionada con el registro del *gazetteer*, conociéndose por tanto sus características geográficas.
Una vez se han reconocido y desambiguado las referencias, se combinan para encontrar el contexto geográfico general del documento. Esta tarea es realizada a través de una técnica similar a la propuesta en [2]. Para cada registro potencialmente referenciado en el documento, se utiliza la relación *part-of* definida en la estructura del *gazetteer* con el fin de extraer los registros padre hasta el nivel raíz. De este modo, se obtiene un conjunto *S* de posibles alcances geográficos. En este conjunto, todos los registros que son referenciados en el documento inician con la calificación asignada en el proceso de extracción e identificación. Estas calificaciones son propagadas y sumadas consecutivamente en los registros padre, para lo que se utiliza una función cuadrática para reducir la calificación propagada de acuerdo al nivel jerárquico en el que se esté heredando, una suma lineal sería asignar pesos sin control y los resultados serían erróneos. Por ejemplo, si se considera que un documento contiene referencias geográficas *A* & *B*, con calificaciones s_A & s_B, y que en el gazetteer existen las relaciones jerárquicas *part-of* del tipo *C/B/A* & *C/B*, la calificación asignada a la referencia *A* será s_A, para el *B* sería $s_B + (s_A * 0.75)$ y para el registro *C* $s_B + (s_A * 0.75^2)$.
El resultado final de este proceso es una lista de candidatos del entorno geográfico del documento ordenado por las calificaciones agregadas. Una vez más, esta calificación puede ser vista como la probabilidad de que un registro determinado corresponda al alcance geográfico del documento.

3.3 Extracción de información temporal

En forma análoga al caso expuesto para información geográfica, la extracción de información temporal es también dividida en tres procesos: reconocer, desambiguar y asignar alcance.
El proceso de reconocimiento implementa de nuevo las técnicas NER con semillas, búsquedas en el *gazetteer* y consultas en registros que contienen nombres de periodos históricos. Se inicia con identificaciones básicas como la contrastación de los nombres propios (aquellos que inician con mayúscula) identificados. Esta aproximación es complementada con reglas de expresión para el reconocimiento de fechas y otro tipo de expresiones que impliquen tiempo. Tal es el caso del algebra de Allen, con el que se identifican y asignan valores a expresiones como *En los*

inicios, *Antes*, *Durante* & *Después*. La *Tabla 1* muestra la descripción de las relaciones tenidas en cuenta para el análisis de este tipo de expresiones.

Tipo de relación	*Ejemplo*	*Relación de limites**
x before y **y after x**	xxxx yyyy	$x^+ < y^-$
x meets y **y met-by x**	xxxx yyyy	$x^+ = y^-$
x overlaps y **y overlap-by x**	xxxx yyyy	$x^- < y^- < x^+$; $x^+ < y^+$
x during y **y includes x**	xxxx yyyyyyyyyy	$x^- > y^-$; $x^+ < y^+$
x starts y **y started by x**	xxxx yyyyyyyyyy	$x^- = y^-$; $x^+ < y^+$
x finishes y **y finished by x**	xxxx yyyyyyyyyy	$x^+ = y^+$; $x^- > y^-$
x equals y	xxxx yyyy	$x^- = y^-$; $x^+ = y^+$

*. Las relaciones $x^- < x^+$ & $y^- < y^+$ son válidas para todos los tipos de relación.

Tabla 1. Componentes básico del Algebra de Allen. Basado en [10]

Las fechas reconocidas con expresiones regulares son convertidas en una representación de tiempo canónica a través de reglas. Para desambiguar los nombres de los periodos históricos identificados, se hacen consultas simultáneas al *gazetteer* junto con las heurísticas mencionadas. Se inicia con consultas simples que seleccionen los registros coincidentes con un nombre exacto. A cada uno de los registros seleccionados se asigna una calificación de 1/*n*, en donde *n* es el número total de registros resultantes de la consulta para la referencia temporal identificada. Para aquellas referencias temporales que tengan más de una coincidencia se usa la heurística *referentes relacionados por discurso* para ajustar la calificación asignada.

Debido a que los nombres de los periodos de tiempo pueden variar de acuerdo a la ubicación geográfica (ej. La expresión ‘periodo de la revolución’ puede tener varios significados en diferentes partes del mundo), aquellos registros que tengan asociada una referencia geográfica y que sea también referenciada en el documento son sensibles a incrementar su calificación en el intervalo [0.2, 1]. Esta misma regla se aplica para los registros cuyos periodos de tiempo se superponen.

Para reconocer el alcance temporal del texto analizado se descartan todos los registros en los que la calificación asignada sea menor a 0.5. De este modo, el alcance esta dado por el periodo de tiempo comprendido entre una fecha inicial correspondiente a la referencia temporal más antigua en el documento, y como es obvio, a una fecha final que coincide con la referencia temporal más reciente.
En los casos en que la extracción de información temporal no es suficiente para suministrar algún resultado o sencillamente falla el proceso, el alcance temporal se trata de identificar con elementos externos como *RSS feeds,* o algún otro medio disponible.
Contando con fechas de inicio y final, es fácil consultar en el sistema aquellos elementos que se quieran filtrar mediante las expresiones del Algebra de Allen implementando la relación de límites de la *Tabla 1*. A través de constantes es posible determinar periodos relativos a expresiones como *En los inicios*.

4 Analizador semántico

Se ha desarrollado un servicio Web[9] para el análisis semántico de expresiones geotemporales implementando los conceptos descritos hasta este punto como parte del proyecto DIGMAP. El objetivo de este servicio es extraer información de carácter geotemporal de un documento o documentos cualquiera. La interfaz de acceso para el sistema implementado se basa en un formato XML resultante de complementar la propuesta realizada por el OGC en [18] y con el objetivo de tener acceso a i) diferentes tipos de referencias geotemporales en el texto analizado, utilizando GML para codificarlas, ii) las puntuaciones establecidas para desambiguar las entidades que así lo requiriesen asociadas a su respectiva referencia, iii) el alcance geotemporal asignado al documento y iv) múltiples salidas usando filtros XSLT.
Para evaluar el tanto el desempeño como el rendimiento del sistema propuesto, se han realizado dos tareas. Primero, se cuantifico el desempeño del servicio a través de un análisis de cargas de trabajo con Apache JMeter, posteriormente, se calificó la calidad de los resultados comparando las referencias temporales y geográficas identificadas automáticamente con un proceso realizado por humanos. La colección de datos utilizado para las pruebas pertenece a un conjunto de metadatos del catalogo de DIGMAP, que contiene descripciones textuales en múltiples idiomas.

[9] http://geoparser.digmap.eu/

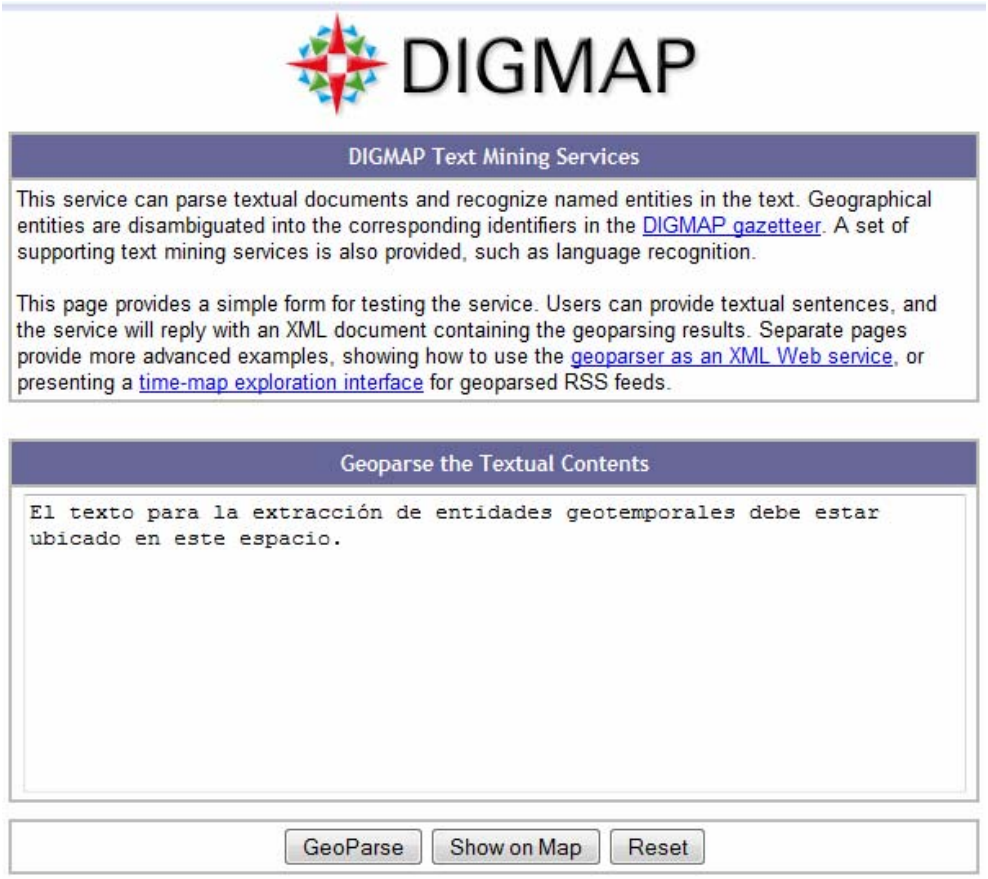

Figura 2. Interfaz de usuario del analizador semántico geotemporal.

Con Apache JMeter se simularon múltiples llamadas al servicio con el fin de medir los tiempos de respuesta en diferentes condiciones. Utilizando un ordenador portátil convencional como servidor y realizando las pruebas en la misma máquina (eliminando de este modo retardos debidos a red) se realizó un test correspondiente a un máximo de 5 de llamadas de usuario por hilo en un periodo de dos segundos. Las simulaciones de carácter aleatorio seleccionaron un conjunto de 100 ejemplos extraídos del corpus Reuters-21758.

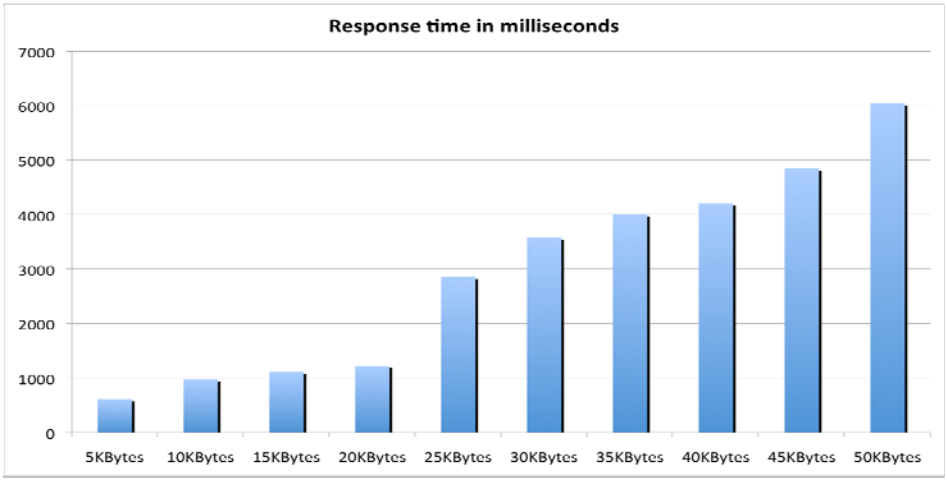

Figura 3. Resultados de las pruebas con JMeter.

La *Figura 3* presenta los resultados obtenidos demostrando que el sistema es capaz de soportar de forma apropiada y escalable los llamados. De aquí también se concluye que procesar textos largos y completos no es un problema que reduzca la fiabilidad y eficiencia del sistema.

Para la prueba también se tuvo en cuenta el tamaño de los documentos, los textos seleccionados se clasificaron en 10 conjuntos con incremento de 5Kb [5 -50Kb]. La *Figura 3* muestra un comportamiento casi lineal en el incremento del tiempo necesario para procesar los textos, se observa como para un fichero de 50Kb (superior al contenido de texto de una página Web convencional) y cinco llamadas simultáneas al sistema el tiempo de respuesta no excede los seis segundos.

Para la evaluación de la calidad de los registros obtenidos se llevaron a cabo comparaciones manuales directas, garantizando de primera mano la fiabilidad del análisis. Seleccionando los mejores registros del catalogo de DIGMAP, registrados previamente por bibliotecarios y cartotecarios, en los que evidentemente se encuentra almacenada información geotemporal de diferentes tipo y en múltiples idiomas, con expresiones de tiempo convencionales (dd/mm/yyyy) o periodos de tiempo. El análisis realizado demuestra que el método propuesto para la asignación del alcance geográfico supera las simples referencias utilizadas en nuestras pruebas de desarrollo. De los 511 recursos utilizados en el experimento, fue posible reconocer las referencias de lugar en un porcentaje altamente significativo. En términos de precisión, los resultados parecen ser de la calidad suficiente para su uso en aplicaciones del mundo real que impliquen el uso del contexto geográfico del documento. Esto es especialmente cierto si pensamos en aplicaciones generales

y de amplio alcance como DIGMAP, en donde la mayoría de los datos son explorados a nivel de grandes regiones geográficas.
En lo referente al reconocimiento de referencias temporales, gracias al registro de hechos históricos sus fechas se obtuvieron resultados aceptables en los periodos. La identificación de fechas en formato convencional no representa ningún problema siempre y cuando tengan un contexto real dentro del texto analizado, no obstante, para desambiguar los elementos relativos a periodos vinculados a la geografía no se apreció una evidente diferenciación, por lo que se podría decir que en este aspecto aún sigue todo por hacer. Encontrar métodos eficientes para desambiguar expresiones temporales relativas a la ubicación de hechos históricos es uno de los trabajos que se continuará desarrollando en esta investigación.
Exceptuando este problema con el análisis de expresiones temporales, los resultados obtenidos se pueden dar por buenos y con la calidad suficiente para que los conceptos y prototipos desarrollados sean implementados en otras aplicaciones.

5 Conclusión

Los registros de metadatos en bibliotecas y cartotecas digitales describen con frecuencia recursos que relacionan algún lugar en un momento específico. Debido a ello, las colecciones pueden ser organizadas de acuerdo a criterios que involucren su distribución espacial y/o temporal. Aunque la idea parece simple, los recursos están caracterizados a menudo por descripciones textuales y tanto los nombres de lugares como los períodos de tiempo presentan una alta ambigüedad; por ejemplo, ¿cuál es el significado detrás de la Guerra del golfo o Periodo de la revolución? La ambigüedad en este tipo de descripciones y en las entidades que representan los lugares y periodos de tiempo debe ser resuelta a fin de obtener una plena comprensión del contexto geotemporal involucrado.
Este artículo muestra como sencillas técnicas de extracción (o relativamente sencillas) son capaces de proporcionar resultados con calidad suficiente para ser utilizadas en procesos complejos y que requieren de una alta precisión en la identificación de entidades geotemporales. Queda demostrado como una correcta interpretación del contexto temporal de un documento puede ser una tarea difícil de automatizar resultando aún más compleja que la interpretación del contexto geográfico.
Se evidencia también como se hace necesario involucrar múltiples elementos externos para una correcta interpretación de las entidades identificadas y poder desambiguarlas con un grado de precisión aceptable. Por supuesto, el contexto temporal necesita de más control y soporte que el geográfico. Relaciones básicas como las contenidas en el algebra de Allen resultan muy útiles en este contexto.

En futuros trabajos, los esfuerzos se centrarán en tareas para mejorar los resultados obtenidos en el dominio temporal. Se puede contar con un conjunto mayor de heurísticas que podrían resultar en mejoras significativas y/o generales. Además de los tipos lugar, la demografía o el idioma del documento podrían ser utilizados como elementos de la heurística *sentido predeterminado*. Los límites involucrados en los métodos propuestos también podrían ser objeto de nuevos estudios, con el fin de ajustarlos a valores óptimos. Por último, el uso de métodos de procesamiento de lenguaje natural más avanzados, como el *part-of-speech tagging*, también podría ser tenidos en cuenta y así poder comparar la sensibilidad en expresiones geotemporales entre métodos básicos y avanzados.

Agradecimientos. Esta investigación ha sido parcialmente financiada por el programa *eContentPlus* de la *European Science Foundation* en el marco del proyecto ECP-2005-CULT-038042 (DIGMAP).

Referencias

[1] Allen, J.F. 1991. “Temporal reasoning and planning”. In *Reasoning about plans*. Edited by J.F. Allen, H.A. Kautz, R.N. Pelavin and J.D. Tenenberg. Morgan Kaufmann, San Francisco - USA. pp. 1-67.

[2] Amitay E., Har'El N., Sivan R. and Soffer A. 2004 Web-a-where: geotagging Web content. *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*

[3] Bates M. J. and Wilde D. N. 1993 *An analysis of search terminology used by humanities scholars: the Getty Online Searching* Project Report Number 1. Library Quarterly, 63(1)

[4] Boguraev B. and Ando R. K. 2005 TimeML-compliant text analysis for temporal reasoning. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*

[5] Buckland M. and Lancaster L. 2004 *Combining Place, Time, and Topic : The Electronic Cultural Atlas Initiative.* D-Lib Magazine, 10(5)

[6] Chen Y., Di Fabbrizio G., Gibbon D., Jana R., Jora S., Renger B. and Wei B. 2007 GeoTracker: Geospatial and temporal RSS navigation. *Proceedings of the 16th World Wide Web conference*

[7] Chinchor N. 1998 *Proceedings of the 7th Message Understanding Conference*

[8] Clough P. and Sanderson M. 2004 A proposal for comparative evaluation of automatic annotation for geo-referenced documents. *Proceedings of the 1st Workshop on Geographic Information Retrieval*

[9] Cohen W. and Sarawagi S. 2004 Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*

[10] Drakengren T. and Jonsson P. "*Eight maximal tractable subclasses of Allen's algebra with metric time*" Journal of Artificial Intelligence Research vol. 7, pp. 25-45, 1997.

[11] Gale W., Church K. and Yarowsky D. 1992 One sense per discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*

[12] Garbin E. and Mani I. 2005 Disambiguating toponyms in news. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*

[13] Harpring P. 1997 The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. *Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, Archives and Museum Informatics*

[14] Hill L. and Zheng Q. 1999. Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. *Proceedings of the American Society for Information Science Annual Meeting*

[15] Jones C., Abdelmoty A., Finch D., Fu G. and Vaid S. 2004 The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Proceedings of the 3rd International Conference on Geographic Information Science*

[16] Jones C. and Purves R. *2006 GIR'05*: *The 2005 ACM workshop on Geographical Information Retrieval*, ACM SIGIR Forum, 40(1)

[17] Kornai A. 2003 *Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References*

[18] Lansing J. 2001 Geoparser service draft candidate implementation specification. OGC Discussion Paper 01-035

[19] Leidner J. 2007 *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*, Ph.D. thesis, School of Informatics, University of Edinburgh, Scotland, UK

[20] Li H., Srihari K. R., Niu C. and Li W. 2002 Location normalization for information extraction. *Proceedings of the 19th Conference on Computational Linguistics*

[21] Malouf R. 2002 Markov models for language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*

[22] Manguinhas, H., Martins, B., Siabato, W. & Borbinha, J. 2008 "The DIGMAP Geo-Temporal Web Gazetteer Service" *Proceedings of the 3th International Workshop Digital Approaches to Cartographic Heritage*.

[23] Manov D., Kiryakov A., Popov B., Bontcheva K., Maynard D. and Cunningham H. 2003 Experiments with geographic knowledge for information extraction. *Proceedings of the HTL/NAACL-03 Workshop on Analysis of Geographic References*

[24] Mikheev A., Moens M. and Grover C. 1999 Named entity recognition without gazetteers. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*

[25] Petras V., Larson R. R. and Buckland M. 2006 Time period directories: a metadata infrastructure for placing events in temporal and geographic context. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*

[26] Pustejovsky J., Castano J., Ingria R., Saurı R., Gaizauskas R., Setzer A., Katz G., and Radev D. 2003 TimeML: Robust specification of event and temporal expressions in text. *Proceedings of the AAAI Spring Symposium on New Directions in Question-Answering*

[27] Sang E. T. K. and De Meulder F. 2003 Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. *Proceedings of the 7th Conference on Natural Language Learning*