GeoNews: Generación automática de contextos geográficos para programas de noticias a través de HbbTV

Moisés Vilar, Sebastián Villarroya, José R.R. Viqueira, and José M. Cotos

Computer Graphics and Data Engineering Group (COGRADE)
Centro de Investigación en Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela (USC)
Santiago de Compostela, Spain
{moises.vilar, sebastian.villarroya, jrr.viqueira, manel.cotos}@usc.es

Abstract. Varios estudios recientes han detectado nuevos hábitos en la audiencia de televisión relacionados con el acceso a través de otros dispositivos a información que complemente los contenidos de televisión. En este documento se describe una solución preliminar para la generación de contextos geográficos para programas de noticias en lengua castellana y para su visualización sincronizada en un televisor a través de tecnología HbbTV.

Keywords: HbbTV, Geoetiquetado, Contexto geográfico, GIS

1 Introducción

La sincronización de los contenidos de televisión con el comportamiento de distintos tipos de aplicaciones es un tema que ha ganado en importancia en los últimos años. Así, algunos estudios recientes realizados por operadoras de televisión demuestran que las acciones publicitarias que combinan contenidos de televisión con aplicaciones en dispositivos móviles (smartphones o tablets) multiplican su eficacia. Relacionado con esto, el estándar para televisión híbrida HbbTV (Hybrid Broadcast Broadband TV) combina la emisión de contenidos de televisión con aplicaciones que pueden llegar al dispositivo del usuario tanto por la emisión broadcast como por conexiones de datos de banda ancha [2]. A diferencia de las plataformas de televisión inteligente (Smart TV) ofertadas por los principales fabricantes de televisores, las aplicaciones HbbTV se incluyen en la misma señal de televisión permitiendo de este modo sincronizar la emisión con el contenido de las mismas. Esto permite que las aplicaciones muestren información contextual ampliada al contenido emitido. Un ejemplo típico es complementar un anuncio publicitario con información adicional del producto correspondiente.

El objetivo principal de este trabajo es la generación automática de contextos geográficos (mapas) para programas de noticias en lengua castellana a partir de los textos de los subtítulos, que ya son obligatorios por ley en España para el 90% de las emisiones de televisión públicas y para el 75% de las privadas.

Para conseguir este objetivo es necesario superar los siguientes retos. El primer reto consiste en detectar las entidades geográficas relevantes nombradas en los subtítulos. La principal dificultad de este reto está provocada por la ambigüedad inherente al lenguaje natural. El segundo reto consiste en la generación de un mapa que permita al espectador localizar las entidades detectadas. Este mapa debe de usar entidades de referencia bien conocidas por el espectador para identificar la posición relativa de las entidades detectadas, por lo que su resultado final depende del conocimiento geográfico del espectador. El último reto tiene que ver con la identificación de subconjuntos de subtítulos para los cuales se quiere generar un único mapa. La principal dificultad en este reto es que el resultado final esperado es muy dependiente de las preferencias de la audiencia. Así, habrá personas que prefieran un sistema muy dinámico en el que el mapa proporciona mucho detalle y cambia muy frecuentemente y habrá usuarios que prefieran un sistema más estático en el que en un único mapa de mucho menos detalle se muestre el contexto geográfico de muchos más subtítulos. La versión actual del sistema proporciona soluciones iniciales sólo para los dos primeros retos.

El resto de este documento se organiza como sigue. La Sección 2 resume el trabajo relacionado. La solución propuesta se describe brevemente en la Sección 3. Por último, la Sección 4 sintetiza líneas de trabajo futuro.

2 Trabajo relacionado

El geoetiquetado de contenido web es un tópico que ha sido ampliamente estudiado en la última década [4], siendo uno de los pilares fundamentales de la nueva área de Recuperación de Información Geográfica (GIR), en la que confluyen investigadores de Recuperación de Información (IR) y Sistemas de Información Geográfica (GIS). De forma genérica, las soluciones propuestas determinan el ámbito geográfico de cada documento mediante las dos siguientes tareas. Primero obtienen un conjunto de entidades geográficas relevantes para el documento referenciadas en algún repositorio externo, habitualmente un gazetteer. Segundo, agregan las geometrías de la entidades relevantes para calcular una geometría única que represente el ámbito geográfico del documento completo. La segunda tarea es bastante dependiente del ámbito de aplicación. Para la primera sin embargo existen muchas soluciones de investigación de propósito general e implementaciones de referencia tanto comerciales [3] como de código abierto [1]. Esta segunda tarea se subdivide a su vez en dos etapas. En una primera fase se identifican nombres de entidades geográficas en los textos y en la segunda se les asignan coordenadas geográficas a estos nombres. La ambigüedad del lenguaje es el principal reto a resolver en las dos etapas [5], ambigüedades de tipo geo/non-geo en la primera etapa (León ciudad y león animal) y ambigüedades de tipo geo-geo en la segunda etapa (Santiago Galicia y Santiago Chile). Varios trabajos de investigación recientes centran sus esfuerzos en la determinación del ámbito de geográfico de contenidos de documentos de noticias [8, 6].

3 Solución propuesta

La arquitectura global de sistema desarrollado se muestra en la Fig.1. La entrada al sistema consta de un archivo de audio y vídeo de un telediario y sus correspondientes subtítulos. Un componente de radiodifusión se encarga de emitir la señal de audio y vídeo a través del canal broadcast proporcionado por el distribuidor, acompañado de la señalización necesaria y de la URL de la aplicación web que se encarga de mostrar el contexto geográfico en el dispositivo HbbTV. En cada instante de tiempo, el dispositivo HbbTV solicita a través de la URL recibida la aplicación web de contexto geográfico, recibiendo el código HTML y Javascript necesario para mostrar en cada momento un mapa obtenido de un servidor de mapas WMS adaptado al contexto de la noticia que se está emitiendo.

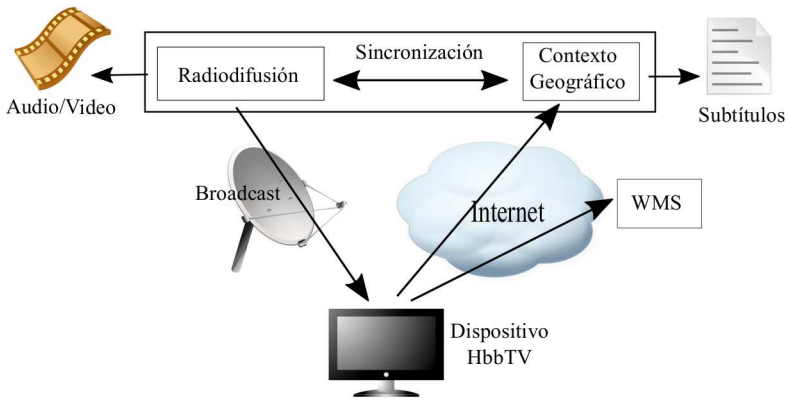

Fig. 1. Arquitectura del sistema.

La generación del ámbito geográfico para cada subtítulo se realiza mediante los tres siguientes pasos.

Parseado de los subtítulos. Se están evaluando dos soluciones para esta tarea, un reconocedor de entidades nombradas (NER) [7] y una solución del ámbito de la anotación semántica [9]. Estas soluciones proporcionan listados de nombres de entidades relevantes para los subtítulos que se corresponden con localizaciones geográficas.

Resolución de entidades geográficas. Los nombres de entidades del paso anterior se utilizan para obtener sus coordenadas de un servicio de gazetteer. Se está evaluando dos servicios, Geonames [1] y Nominatim [2].

[1] http://www.geonames.org/
[2] http://nominatim.openstreetmap.org/

Cálculo del ámbito geográfico. Para cada subtítulo, se calcula el rectángulo mínimo (MBR) que contiene a todos las entidades resueltas en el paso anterior. El ámbito geográfico a mostrar será un rectángulo que debe contener al rectángulo que se acaba de calcular y al menos a una entidad geográfica bien conocida por el usuario que le sirva como referencia visual. El sistema dispone por lo tanto de una colección de entidades bien conocidas adaptada en este caso a lo que se estima es el conocimiento geográfico medio de un espectador de telediarios en España. Para la obtención de la entidad bien conocida de referencia se están evaluando varios métodos basados en el cálculo de distancias, perímetros, áreas y aspectos. Para la generación del mapa final con el rectángulo obtenido se está utilizando un servicio de mapas WMS con cartografía de OpenStreetMaps [3].

4 Trabajo futuro

El trabajo futuro se centra en la determinación de los subconjuntos de subtítulos que deberían mostrarse conjuntamente en el mismo contexto geográfico, la evaluación exhaustiva y mejora de la eficacia del procedimiento de generación del propio contexto y la habilitación de la personalización del sistema para cada espectador mediante retroalimentación.

References

1. Clavin: Cartographic location and vicinity indexer. https://clavin.bericotechnologies.com/, consultada en Abril de 2015.
2. Hbbtv specification. https://www.hbbtv.org/, consultada en Abril de 2015.
3. Yahoo pacespotter. https://developer.yahoo.com/boss/geo/docs/key-concepts.html, consultada en Abril de 2015.
4. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: Geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 273–280 (2004)
5. Buscaldi, D.: Approaches to disambiguating toponyms. SIGSPATIAL Special 3(2), 16–19 (2011)
6. D'Ignazio, C., Bhargava, R., Zuckerman, E., Beck, L.: Cliff-clavin: Determining geographic focus for news articles. In: NewsKDD workshop, 201h ACM SIGKDD conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA (2014)
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370 (2005)
8. Samet, H., Sankaranarayanan, J., Lieberman, M.D., Adelfio, M.D., Fruin, B.C., Lotkowski, J.M., Panozzo, D., Sperling, J., Teitler, B.E.: Reading news with maps by exploiting spatial synonyms. Commun. ACM 57(10), 64–77 (2014)
9. Vidal, J.C., Lama, M., Otero-García, E., Bugarín, A.: Graph-based semantic annotation for enriching educational content with linked data. Know.-Based Syst. 55, 29–42 (2014)

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