CAPÍTULO IV

DATIFICACIÓN EN LOS ARCHIVOS DIGITALES DE LOS MEDIOS DE COMUNICACIÓN: NUEVOS RETOS

Ángel Luis Garrido Marín
SID Research Group
Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza, Zaragoza (España)

Carlos Bobed Lisbona
SemLIS Research Group
University of Rennes / CNRS / IRISA, Rennes (Francia)

Resumen

En los últimos años los archivos digitales de los medios de comunicación han sufrido una gran transformación, pasando de usar métodos artesanales a su casi completa digitalización. De forma paralela, hemos asistido en la Web a un crecimiento exponencial de contenidos que son también de gran ayuda en la elaboración de noticias para los periodistas. El problema viene cuando un redactor está elaborando un nuevo contenido y necesita apoyo documental, y se ve desbordado por la gran cantidad de información que tiene a su disposición. A las habituales y numerosas informaciones que proporcionan las agencias se suman los contenidos propios del archivo digital y toda la vasta información (y desinformación) que puede proporcionar Internet. En este trabajo presentamos una visión general de un sistema de gestión y explotación del archivo digital destinado a las redacciones y a los equipos de documentación de los medios de comunicación. El sistema hace uso y combina técnicas de procesamiento del lenguaje natural, el aprendizaje automático y la Web Semántica, realizando tareas de datificación que permiten gestionar los grandes volúmenes de información actuales. Los distintos módulos del sistema han sido ya evaluados parcialmente, tanto con conjuntos de datos sintéticos como con reales, obteniendo resultados prometedores. Nuestra colaboración con medios de comunicación ha sido una constante durante estos últimos años, y las publicaciones en congresos y revistas internacionales avalan los resultados obtenidos.

Palabras claves: Informatización de archivos, Industria de la información, Internet, Automatización de bibliotecas, Web Semántica, Linked Data

1. Introducción

1.1 Grandes volúmenes de información: Big Data y datificación

En la actualidad, el concepto de *Big Data* se ha convertido en un término muy popular, ampliamente extendido en ámbitos académicos, tecnológicos, empresariales, etc. y con una frecuente aparición en medios de comunicación. A pesar de su popularización, su definición no está siempre tan clara, y además hay que tener en cuenta que el concepto ha ido evolucionando a lo largo del tiempo, pero siempre ha estado ligado a la gestión y el análisis de grandes volúmenes de información que no pueden ser tratados de forma convencional (Gandomi, 2015). La dificultad en el tratamiento estriba en los límites de las herramientas informáticas (hardware y software) usadas para su captura, gestión y posterior procesamiento.

Laney (2001) definió y caracterizó el concepto de *Big Data* en base a tres dimensiones: *Volumen, Variedad* y *Velocidad*. El *Volumen* es quizás la característica más asociada al concepto de *Big Data*. Entre las distintas causas que podemos encontrar para explicar el significativo aumento de datos disponibles hoy en día en cualquier ámbito se encuentran, como más destacadas, el abaratamiento del almacenamiento, junto con el aumento de datos generados a través del uso de la Web, las redes sociales, los dispositivos móviles, y todo ello unido al aumento de usos de sensores en todo tipo de procesos y actividades. Sin embargo el valor de esta información puede tener un ciclo de vida muy corto, quedando obsoleta en poco tiempo. Este tipo de apreciación enlaza con la dimensión de *Velocidad*, que es la que mide la frecuencia de creación de los datos, y determinando el margen disponible para dar una respuesta adecuada a su procesamiento y análisis. La *Variedad* en *Big Data* se basa en la diversidad de los tipos de datos (estructurados, semi-estructurados, o desestructurados) y de las distintas fuentes de donde son obtenidos (texto, imagen, video, audio, sensores, flujos de datos, etc.). Esta potencial riqueza de los datos que aporta la variedad, por otra parte aumenta el grado complejidad tanto en su almacenamiento como en su procesamiento y análisis. En los últimos años, se han ido añadiendo nuevas dimensiones, como el *Valor* y la *Veracidad* (Zikopoulos 2011). Cuando se habla de *Valor* se quiere hacer hincapié en que la recuperación del dato en sí mismo no aporta realmente valor si no es útil para la toma de decisiones de los usuarios de la base de datos, por lo que un aspecto importante del *Big Data* es que debe aportar *valor real* para ser considerado como tal, lo que es complejo a medida que aumenta el volumen y la complejidad. Por otra parte, la *Veracidad* determinará la calidad de los resultados y la confianza en los mismos.

Adicionalmente, se han definido otras V's como la *Variabilidad* (Zikopoulos 2013), referido a los rápidos cambios de cantidad y valor que pueden

experimentar los datos, causando picos en las cargas de datos y obligando a adaptar las herramientas para su captura. La *Visualización* (Keim, 2013) está relacionada con la tarea de como representar visualmente los datos recuperados de manera que sean legibles y accesibles, lo cual muchas veces tampoco es una tarea nada trivial. Por último, la *Validez (Khan, 2014)* indica el grado de corrección y precisión de los datos recuperados con respecto a uso para el que están destinados.

Ligado al concepto de *Big Data* está el concepto de *datificación (Markus, 2017)* cuyo significado es *la acción de convertir en datos una información, un proceso o actividad*. Los principales objetivos que se persiguen al realizar una datificación se encuentran frecuentemente entre los enumerados a continuación:

1. *Recuperar datos existentes* en el objeto de la datificación, pero que por el volumen, la velocidad, o su variabilidad, resultan costosos de obtener. Por ejemplo, recuperar los datos de sensorización de una máquina en los periodos previos y posteriores a que se producen microcortes de tensión.
2. *Extraer algún tipo de información* concreta contenida o deducible de los datos existentes, por ejemplo, al querer obtener todos los planos en los que aparece una determinada marca comercial en una filmoteca digital.
3. *Monitorizar* los datos producidos por el objeto de la datificación. Por ejemplo, vigilar el comportamiento, segundo a segundo, de cada una las turbinas de un avión comercial durante los distintos trayectos que realiza.
4. *Optimizar* un comportamiento asociado al objeto de la datificación, como podría ser el encontrar el mejor momento para realizar acciones de compra-venta de acciones en un entorno bursátil.
5. *Buscar patrones* en un conjunto masivo y cambiante de datos, por ejemplo, al tratar de encontrar factores comunes en los comportamientos y las preferencias de los clientes que realizan compras en una web comercial, para mejorar los sistemas de recomendación asociados a dicha web.

Con la datificación, los datos se pueden cuantificar de manera que pueden ser tabulados y analizados. Su digitalización y almacenamiento en diferentes formatos va a ser el que va a permitir que se pueda realizar un análisis de ese *Big Data,* de cara a cumplir uno o varios de los objetivos enumerados anteriormente.

Las posibles fuentes de datos sobre las que realizar procesos de datificación son muy variadas, pero normalmente van a estar relacionadas con soportes informáticos: las acciones o los comportamientos de los usuarios de un software, las actividades de las propias aplicaciones o programas que utilizan,

los eventos de los sistemas operativos sobre los que se ejecutan, los contenidos de las bases de datos donde se guarda la información, o todo lo que registran los sistemas de sensorización, serían algunos ejemplos habituales.

1.2. Los archivos digitales en los medios de comunicación

Definimos un *archivo digital* (Lavoie, 2004) como una organización de contenidos almacenada mediante soportes informáticos con el objetivo de preservar el acceso a una determinada información por parte de una comunidad. Los archivos tienen una gran relevancia en ámbitos públicos y privado, albergando contenidos de todo tipo: administrativos, legales, artísticos, históricos, culturales, o industriales, por citar algunos de los más representativos. Estos archivos, habitualmente se encuentran implementados mediante una base de datos documental que contiene información de tipo texto, imágenes, audio, y/o vídeo, así como los atributos de los contenidos, y los índices que facilitan el acceso a los datos. Dicho archivo, en función de su volumen, la velocidad de la producción de nuevos elementos, y de la variabilidad de los mismos, es una fuente de datos susceptible de recibir la aplicación de un proceso de datificación para conseguir uno o varios de los objetivos comentados anteriormente.

Un ejemplo claro de este tipo de archivos en el que la gestión de la información puede convertirse en un problema de cierta complejidad sería el archivo digital de un medio de comunicación. Independientemente del tipo de soporte (radio, televisión, prensa, web, etc.) el archivo documental va a poder albergar todo tipo de contenidos multimedia en formatos digitales muy variables, y en muchas ocasiones, difíciles de gestionar. Dicha dificultad no solo va a venir dada por el tipo de fichero que hace las veces de soporte, sino también por el contenido del mismo: el lenguaje humano, las imágenes, y los sonidos. Las tres fuentes (Cambria 2014, Sonka 2014, Gold, 2011) representan tipos de información que, por su propia naturaleza, en nuestros días todavía constituyen de por sí un reto a la hora de realizar un procesamiento completo por parte de los sistemas informáticos.

A estos contenidos, que pueden ya tener una cierta magnitud en función de la antigüedad, de la frecuencia de aparición de nuevos contenidos, y de la riqueza del archivo, hay que añadir la presencia de otros atributos que describan a su vez a dichos contenidos, como por ejemplo la fecha, el autor, el tamaño, la fuente de procedencia, etc.

También es importante en este tipo de archivos tener en cuenta la presencia de *herramientas documentales* necesarias para la catalogación de estos fondos. Entre dichas herramientas, por citar las más usadas, encontramos las siguientes (Gilchrist, 2003):

- *Listas controladas de términos*: Es la forma más sencilla de etiquetar con fines clasificatorios. Dado un determinado atributo (p. ej. "sección") se puede elegir una serie de valores fijos (siguiendo el ejemplo anterior: "nacional", "deportes", "opinión", etc.). Se procura disponer de ellos a priori (aunque la lista vaya siendo modificada) para mantener cierta consistencia en la posteriores búsquedas de noticias. Los elementos de una lista de términos no guardan ninguna relación entre ellos.
- *Taxonomía:* Es una lista de términos cuyos elementos guardan una relación jerárquica entre ellos (p. ej. "Lugar" podría tener como valores posibles "España" / "Andalucía" / "Jaén" / "Linares", siendo cada término superior jerárquicamente al siguiente). De esta manera y siguiendo el ejemplo, se podría deducir que un suceso que ocurre en Linares ha ocurrido también en Jaén, en Andalucía, y en España. Por otra parte, estamos asignando niveles de detalle a la clasificación, perteneciendo "España" al nivel más general, y "Linares" al más detallado.
- *Tesauro:* Según la norma ISO-2788, un tesauro es una taxonomía que cumple al menos estas tres condiciones: 1) relaciones de sinonimia o preferencia, 2) relaciones jerárquicas de tipo partitivo (todo-parte) o clase-subclase, y 3) relaciones asociativas: entre términos relacionados de forma pragmática, es decir, ni de forma jerárquica ni de sinonimia.
- *Ontologías:* Las ontologías se definen como especificación formal y explícita de una conceptualización (Gruber, 1993). Para que una herramienta documental pueda afirmarse que es una ontología debe estar en un formato que pueda ser procesado usando un software capaz de realizar inferencias. Es decir, una ontología debe estar basada en lógica formal y debe ser expresada como una colección de aserciones (p. ej. "[Mark Knopfler]" > "[nacer]" > ["Glasgow"]). La expresividad de las ontologías y su calidad a la hora de clasificar la información puede ser muy alta, pero también implica una mayor dificultad en su uso y en su implantación. Aunque el uso los tesauros está fuertemente arraigado en la comunidad de los profesionales de la documentación, poco a poco se están aplicando las ontologías para organizar y gestionar archivos documentales (Hodge, 2000), e incluso la norma ISO-25964 (Dextre Clarke, 2012) ya formaliza su uso.

En función de la envergadura del medio y los recursos empresariales de los que disponga, la solución dada a los archivos digitales de los medios de comunicación varía de unas empresas a otras. Si nos centramos en España

(Recio 2009), entre las grandes empresas encontramos: 1) soluciones comerciales completas dedicadas a medios de comunicación, 2) productos software dedicados a la gestión de archivos documentales de tipo más general, pero adaptados a este entorno, o 3) sistemas de archivo desarrollados a medida. En muchos casos se trata de sistemas legados que no se ha optado por actualizar, y que arrastran las limitaciones de su antigüedad. La presencia de departamentos de documentación y su capacidad de influencia es un factor clave en la correcta organización y adecuada evolución tecnológica del archivo digital del medio (Moral, 2014). En medios más pequeños, es muy común encontrar fuertes carencias tecnológicas en los archivos, por lo que éstos se limitan a ser una colección de ficheros y/o una base de datos más o menos desatendida.

1.3. Limitaciones y retos actuales en los archivos digitales de los medios

En los archivos digitales de los medios de comunicación pueden establecerse ciertos patrones comunes, de menor a mayor en lo que respecta a inversión en medios tecnológicos:

1. Existe uno o varios almacenes de ficheros con los contenidos binarios originales: páginas, imágenes, vídeos, audios, etc.
2. El acceso a la información se gestiona a través de herramientas que indexan los contenidos y permiten una búsqueda ágil mediante el uso de un índice inverso, por lo que se pueden buscar elementos a través de una típica consulta basada en palabras clave.
3. En los mejores casos, existe una base de datos que incluye la posibilidad de gestionar el uso de los atributos de los contenidos.
4. Cuando existe un cierto personal dedicado a tareas de documentación, generalmente aparecen herramientas de gestión documental como las citadas anteriormente: listas controladas de términos, taxonomías, tesauros, y/o ontologías.

Otra característica común a este tipo de archivos es que pueden caracterizarse como *silos de* datos (Seaman, 2003), ya que habitualmente son repositorios de datos con una finalidad única, dedicados a un departamento concreto, y aislados del resto de la organización, y de otros repositorios externos (como los existentes en la Web), lo que limita en gran medida sus posibilidades de enriquecimiento y explotación.

Actualmente, los nuevos modelos de redacciones y en general el periodismo digital están demandando unos servicios muy concretos por parte de los servicios documentales de los medios de comunicación, y en consecuencia, por parte de sus archivos digitales: la hemeroteca digital, la contextualización documental y la elaboración de productos documentales (Guallar, 2011). La cada vez más escasa presencia de documentalistas en las planti-

llas, y las carencias técnicas en cuanto a software capaz de proporcionar dichos servicios está dificultando estas labores, y pone en peligro la calidad de los trabajos periodísticos.

En este trabajo presentaremos una nueva aproximación para realizar la datificación de un archivo de un medio de comunicación cuyos contenidos estén basados en texto en lenguaje natural. En particular, nuestra propuesta consiste en la creación de un sistema que combina análisis y síntesis para realizar y mejorar las distintas tareas documentales requeridas, creando y enriqueciendo nuevos contenidos en base a los ya existentes.

El resto del artículo está estructurado de la siguiente manera: los propósitos perseguidos por este estudio se detallan en la Sección 2, "Objetivos Generales". La Sección 3, "Método", describe la aproximación propuesta. Los experimentos y los resultados hasta el momento se presentan en la Sección 4, "Resultados". Finalmente, la Sección 5, "Discusión y conclusiones", analiza y resume lo presentado en los anteriores apartados, enumerando asimismo las futuras líneas de trabajo.

2. Objetivos Generales

El objeto de este artículo es proponer el diseño de un sistema que asista a los redactores y a los documentalistas en las tareas que realizan de forma habitual sobre los archivos digitales de los medios de comunicación. De esta manera, el personal del departamento de documentación puede aumentar su productividad y enfrentarse con un personal limitado a grandes volúmenes de datos. Para ello el sistema propuesto debe ser capaz de:

1. *Automatizar el archivado:* No solo hablamos del correcto guardado en el archivo digital de los ficheros físicos que representan el soporte del contenido (imágenes, texto, videos, etc.), sino de la obtención e inserción los datos de cada uno de estos contenidos.
2. *Clasificar y etiquetar los contenidos:* Aparte de datos descriptivos básicos, en aquellos entornos en los que existan herramientas de clasificación y etiquetado (listas controladas, tesauros, ontologías, etc.) el sistema ha de ser capaz de etiquetarlos y/o clasificarlos de acuerdo a los criterios establecidos por el medio de comunicación.
3. *Crear resúmenes:* El sistema será capaz de resumir uno o varios elementos basados en texto en lenguaje natural, condensando en un nuevo elemento de texto las ideas más importantes en base a unos parámetros predefinidos de extensión y simplificación.
4. *Realizar búsquedas avanzadas:* En ocasiones es necesario recoger mucha información para poder elaborar una consulta concreta y aparentemente sencilla sobre el archivo documental, como por ejemplo saber cuántas veces ha sido campeón un equipo en una determinada competición deportiva entre dos fechas. Probablemente

ese dato no figura de forma explícita en ningún documento del archivo, pero, al recoger toda la información referente a ese equipo existente en el archivo, un humano podría fácilmente *deducirla*. Si algo es deducible, se entiende que a través de herramientas informáticas es también *calculable* si le damos al sistema recursos para manejar ese tipo de información y poder realizar las mismas deducciones lógicas y así extraer esa información.

5. *Personalizar la información:* Un aspecto muy importante en este tipo de archivos es tener en cuenta *para qué y para quién* son los datos que se recuperan del mismo. La finalidad informativa y el perfil propio de los usuarios, en este caso los periodistas, han de tener un protagonismo relevante a la hora de determinar la información a recuperar y cómo presentarla para su consulta, llegando al extremo de atender de forma completamente personalizada las necesidades documentales.
6. *Elaborar informes, fichas, y dossiers:* Por último, y esta es la tarea más compleja, en ocasiones lo que se le reclama a los departamentos de documentación es la realización de documentos que contengan información elaborada a partir de los contenidos del archivo. Estos documentos pueden ser pequeños informes, fichas temáticas y/o dossiers de cierta profundidad. Esta síntesis de información es uno de los grandes retos hoy en día para un sistema de información, ya que para ello requiere cumplir gran parte los anteriores objetivos, sobre todo en lo referente a la extracción de información concreta y en la capacidad de resumir.

En entornos sin personal dedicado a la documentación, la presencia de un sistema capaz de realizar estas tareas puede suplir hasta cierto punto la falta de documentalistas, pero sin presencia humana se pierde el factor de supervisar y corregir al sistema. Por el momento queda fuera de alcance de esta propuesta que el sistema sea capaz de realizar sus tareas de forma completamente desatendida y aprendiendo de sus errores.

3. Método

Para la consecución de estos objetivos se propone un marco de trabajo en el que agregamos por una parte herramientas de análisis que contribuyan a la datificación del archivo, y por otra parte herramientas de síntesis que contribuyan a su enriquecimiento. En este apartado en primer lugar ese estudiará dichas herramientas por separado y luego presentaremos cómo nuestro sistema las utiliza de forma coordinada para las distintas tareas antes mencionadas.

3.1. Herramientas de análisis

Las herramientas de análisis de los contenidos textuales del archivo digital de las que nos serviremos son las siguientes:

1. *Lematización*: Consiste en la obtención del *lema* de cada palabra, también denominada *forma canónica de una palabra*. Por ejemplo, las palabras profesor, profesores y profesora tendrían como lema "profesor". En el caso de los verbos, la forma canónica es el infinitivo, sea cual sea la forma conjugada. Este proceso ayuda a reducir el espacio de datos y facilita las búsquedas de elementos relacionados independientemente de las palabras clave introducidas por el usuario.
2. *Detección de Named Entities:* Las *Named Entities* hacen referencia a nombres propios (personas, organizaciones, lugares, etc.), que aparecen capitalizados. Un ejemplo claro sería "Los Ángeles", referido a la ciudad norteamericana. Tiene más lógica tratar esas dos palabras de forma conjunta que procesar el texto considerando de forma separada "Los" y "Ángeles".
3. *N*-gramas: Un n-grama, simplificando, es una asociación de varias palabras que aparecen seguidas. Los textos también pueden estudiarse por medio de esta herramienta, ya que en muchas ocasiones determinadas secuencias de palabras que aparecen siempre en el mismo orden pueden ser relevantes para clasificar los textos en los que aparecen. Un ejemplo claro sería el n-grama "la crisis de Cataluña", formada por cuatro palabras que aparecen juntas en un gran número de ocasiones y aportan mucha información en lo que respecta al contenido de la noticia.
4. *Análisis de frecuencias:* La frecuencia de aparición de los términos anteriormente descritos es otra herramienta de gran utilidad a la hora de estudiar la relevancia de unas noticias u otras con respecto a una determinada temática. Estas frecuencias pueden estudiarse de forma local (dentro de un mismo contenido) o de forma global (con respecto a todo el archivo). Un texto con alta frecuencia de palabras tales como "médico", "enfermero", "paciente", "hospital" o "consulta", muy probablemente esté relacionado con el tema "Sanidad".
5. *Sintaxis y morfología:* El estudio de las características del lenguaje usado en los contenidos también aporta información valiosa de cara a realizar diferentes análisis del archivo digital, y en algunos casos nos pueden permitir discernir el rol de un término. Por ejemplo, el significado de la palabra "poder" como sustantivo o como verbo es completamente diferente.

6. *Patrones:* El uso de patrones que relacionen los anteriores elementos entre sí de forma abstracta será también una potente herramienta, sobre todo aplicada en el escenario de extracción de información. Por ejemplo, el patrón "[persona], [nacido en/originario de/ oriundo de] + [Lugar]" es una herramienta muy interesante para encontrar datos relacionados con el lugar de nacimiento de personas que aparecen en el archivo.
7. *Correferencias:* La detección de correferencias es otro elemento sumamente interesante en el estudio analítico de textos, sobre todo en textos periodísticos donde el sujeto de una misma noticia puede aparecer de muchas formas diferentes, pero todas ellas hacen referencia a esa misma entidad. Por ejemplo, en un mismo texto podemos encontrar para hacer referencia a Rafael Nadal las siguientes acepciones: "Rafael Nadal", "Nadal", "el mallorquín", "el tenista español", "el jugador", etc.
8. *Aprendizaje automático:* Es una herramienta muy valiosa, consistente en dar una gran cantidad de problemas resueltos al ordenador de forma que, por medio de modelos matemáticos y estadísticos, la máquina sea capaz de encontrar patrones comunes que le permitan resolver nuevos problemas similares. Por ejemplo, si le damos una gran cantidad de artículos periodísticos etiquetados como "deportes" u "otros temas", cualquier sistema de aprendizaje automático podrá de manera sencilla generar un clasificador capaz de acertar en un alto porcentaje si un nuevo texto es una noticia de deportes o no.

Todas estas herramientas han sido objeto a lo largo de los últimos años de múltiples estudios científicos, demostrando su utilidad a la hora de realizar tareas de análisis de información, por lo que podemos afirmar que su uso está suficientemente asentado y extendido.

3.2. Herramientas de síntesis

En lo referente a herramientas de síntesis, el sistema propuesto incluye una serie de subsistemas prácticamente independientes, pero que funcionan de forma coordinada. Todos ellos se sirven de las anteriores herramientas de análisis para realizar parte de su trabajo. En la Figura 1 se muestra un gráfico donde se puede ver como se combinan y comunican entre sí con el fin de conseguir los objetivos propuestos en la Sección 2.

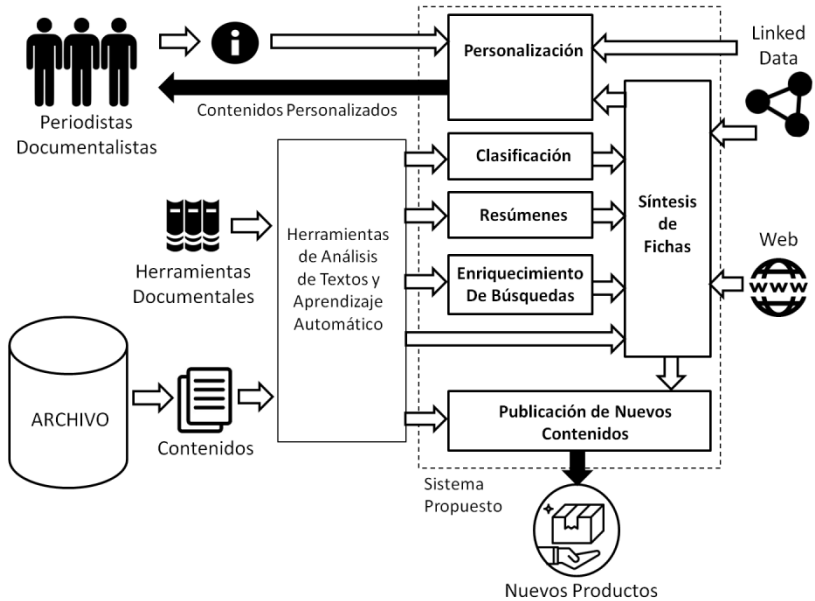

Figura 1. Arquitectura del sistema propuesto, en el que se muestran los subsistemas de síntesis y su relación con el resto de elementos externos al sistema.

A continuación se detallan las funcionalidades de cada uno de los subsistemas de síntesis:

a) *Clasificación*: Uno de las tareas más estudiadas dentro del área del procesamiento del lenguaje natural es la clasificación y la categorización de contenidos (Korde, 2012). Aún así, todavía es un problema abierto que requiere en muchos casos de soluciones especializadas, como la que proponemos en nuestro sistema basada en combinar reglas, aprendizaje automático, y ontologías (Garrido, 2011).

b) *Resúmenes automáticos:* La posibilidad de realizar resúmenes de forma desasistida se remonta a la década de los 50 (Luhn, 1958). Las técnicas han ido evolucionando desde entonces y la calidad de los resúmenes ha ido mejorando con los años (Nenkova, 2011). A la hora de hacer resúmenes existen diferentes técnicas, pero en nuestro sistema usaremos una aproximación extractiva (Gupta, 2010), mejorada mediante el uso de herramientas de aprendizaje automático (Garrido-Bobed, 2017).

c) *Enriquecimiento de búsquedas:* Mediante técnicas de procesamiento del lenguaje natural, nuestro sistema realiza de forma combinada expansiones de las consultas y simplificación de términos al identificar *Named Entities* (Buey 2014).

d) *Personalización de la información*: El sistema también incluye un subsistema de recomendación híbrido basado tanto en el conocimiento que disponemos sobre el usuario como en su actividad, así como en el contenido de las noticias. Este subsistema se ha optimizado en lo referente a desambiguación de términos mediante el uso de información externa semántica localizada en la Web conocida como datos enlazados o *Linked Data* (Bizer 2009).
e) *Síntesis de fichas:* Por último, el sistema permite diseñar y rellenar de forma automática fichas relativas a *entidades* (personas, lugares, temas, organizaciones, obras, etc.) que aparecen en los contenidos del archivo (Garrido-Sangiao, 2017). Para ello, se crea en primer lugar de forma desasistida un catalogo de entidades, etiquetando después todo el archivo en base a dichas entidades. Las entidades se desambiguan para conseguir una mayor precisión del sistema de nuevo sirviéndonos de técnicas relacionadas con la Web Semántica (Berners-Lee, 2001). Después, para cada entidad se crea una ficha descriptiva combinando tanto información procedente del archivo (utilizando para ello las herramientas de síntesis anteriores), como información procedente de la Web.
f) *Publicación de nuevos contenidos*: En base a las fichas obtenidas, la información se puede volver a publicar con nuevos formatos y múltiples fines. Su consumo se puede realizar o bien desde el propio archivo, creando por ejemplo páginas web automáticas que entren a formar parte del archivo digital, o bien de forma externa, mediante la creación de servicios API-REST, diseñando mapas conceptuales, o publicando los contenidos en formato Linked Data consultable a través de consultas SPARQL.

4. Resultados

En esta sección se presenta un resumen de los resultados de investigación relacionados con el sistema propuesto. Debido a su extensión, se encuentran repartidas en varias publicaciones, cada una de ellas relacionadas con uno de los subsistemas de síntesis presentados en la Sección 3.2:

- *Clasificación automática:* las publicaciones más relevantes relacionadas con este tema son (Garrido, 2011) y (Garrido, 2012), donde se presenta el sistema NASS (News Annotation SemanticSystem), que combina el uso de reglas, aprendizaje automático y ontologías, consiguiendo más un 95% de precisión y exactitud en las pruebas experimentales. Este clasificador se especializó para descriptores geográficos (Garrido, 2013), y finalmente se realizó un amplio estudio experimental en distintos medios de comunicación donde se probó la validez del sistema (Garrido, 2014)

- *Optimización de búsquedas:* En este tema destacan los trabajos relacionados con el sistema SQX-Lib, que en los experimentos fue capaz de mejorar los resultados de más del 95% de las consultas realizadas por los periodistas en un medio de comunicación real (Buey, 2014).
- *Resúmenes automáticos:* Las aportaciones introducidas por el sistema se muestran en (Garrido-Bobed, 2017), donde se propone la personalización de la elaboración de resúmenes de noticias en base a su tipología. Esta tipología se obtiene a través de procesos de aprendizaje automático, consiguiéndose una mejora sustancial con respecto a otras herramientas de creación automática de resúmenes.
- *Extracción de información*: Un aspecto importante a evaluar es la capacidad de conseguir información precisa dentro de grandes volúmenes de información basada en texto en lenguaje natural. A este respecto destacan los trabajos de (Rincon, 2014), (Buey 2016) y (Garrido-Buey, 2016), que aunque no están directamente relacionados con los archivos digitales de los medios de comunicación, permitieron avanzar en el desarrollo de las técnicas que más adelante se usarían en el sistema descrito en este trabajo.
- *Detección de entidades y elaboración de fichas:* La dificultad de creación de este subsistema ha motivado el desarrollo de varias publicaciones a lo largo de los últimos años. Las primeras propuestas (Garrido, 2014) y (Garrido, 2015) sentaron las bases del diseño del subsistema. En (Garrido, 2016) se presenta NEREA como herramienta de creación automática del catálogo de entidades y POIROT como herramienta de desambiguación. El sistema es posteriormente ampliado (Garrido-Sangiao, 2017) con la herramienta SOPHIE, que asiste al usuario en la creación de las fichas de forma automática en base al catálogo de entidades creado. La combinación de estas herramientas consigue muy buenos resultados frente a otros sistemas de detección de entidades y los experimentos realizados en medios de comunicación reales arrojan un resultado de más del 70% de precisión en la tarea de rellenado de las fichas.

Actualmente se está trabajando por un lado en mejorar los resultados de extracción de información y creación de fichas, así como en la creación de nuevos contenidos, y por otro lado en el subsistema de personalización de información. Los trabajos están muy avanzados y se confía en publicar a corto plazo los resultados obtenidos.

5. Discusión y conclusiones

La principal aportación de este trabajo es la presentación de la arquitectura completa de un sistema dedicado al estudiar el problema concreto de la gestión y explotación automática de los archivos digitales en el contexto concreto de un medio de comunicación. Para ello se ha propuesto un novedoso sistema que, sirviéndose de procesos de datificación, permite un mejor análisis de los grandes volúmenes de información manejados en este tipo de archivos y facilitan la síntesis de nuevos elementos documentales relacionados entre sí.

Hasta donde hemos podido investigar, no existen trabajos académicos que estudien este problema desde la óptica presentada en este trabajo, pero sí podemos citar algunos trabajos estrechamente relacionados con el nuestro, y que en algunos casos nos han resultado inspiradores. Entre los más antiguos podemos encontrar algunos de finales de los 90 como (Palmer, 1999) en los que se empieza a analizar la relevancia de los archivos digitales en los medios de comunicación, o un interesante estudio del archivo digital de fotografías de los periódicos (Markkula, 2000) que ya entonces observa las limitaciones de "simplemente indexar" el contenido como única acción documental. Más adelante, encontramos otro destacado trabajo que explica como el uso de la Web como un canal más dentro del medio de comunicación afectaba a las estructuras internas de las redacciones de los periódicos (Boczkowski, 2005), creando nuevas necesidades y al mismo tiempo ofreciendo nuevos recursos. Nos parece obligado reseñar también el trabajo realizado por (Deacon, 2007) que profundiza en los problemas y limitaciones de los archivos digitales basados en texto desde el punto de vista documental. Las nuevas posibilidades que ofrecen los archivos digitales se exploran por ejemplo en trabajos como (Nicholson, 2013), que abre la puerta incluso a nuevas formas de hacer periodismo a partir de datos extraídos de los archivos digitales de los medios.

En lo referente a sistemas comparables al que proponemos en este trabajo, es obligado citar NEPTUNO (Castells, 2014), un sistema de explotación del archivo digital basado en las técnicas semánticas. La arquitectura de dicho sistema, presentada a través de un caso de uso en un periódico real, está basada en la creación manual de una ontología que es poblada de datos a través de un proceso de migración del archivo actual, añadiéndosele después herramientas de búsqueda y explotación. La gran diferencia con nuestra aproximación es en el alcance y en los objetivos: nuestra propuesta está enfocada como se ha comentado en el presente trabajo a aumentar la productividad de los departamentos documentales, por lo que abarca no solo los procesos de recuperación de información si no también, los de clasificación, resumen, y síntesis en nuevos formatos pensados para una re-publicación posterior. Por otra parte, en nuestro enfoque la creación y el uso de

la ontología que describe el archivo no es la base del funcionamiento de nuestro sistema, sino un posible recurso más a utilizar.

Es también destacable el trabajo de (Elragal, 2017), en el que se presenta una ambiciosa hoja de ruta futura que afronta muchos de los problemas aquí comentados relacionados con la datificación de los archivos digitales en los medios de comunicación, pero que claramente está todavía en una fase muy preliminar.

Por concluir, las principales aportaciones de nuestra propuesta son la novedad que representa nuestra investigación, hasta ahora no abordado desde esta perspectiva en otros trabajos, así como las capacidades que presenta el sistema para abordar los grandes volúmenes de información basados en texto usados en los medios de comunicación para realizar tareas de documentación: archivos digitales, noticias de agencias, y la propia Web, imposibles de abarcar de manera efectiva por equipos reducidos dentro del limitado tiempo del que se dispone en las organizaciones dedicadas a la información. Por otra parte, es destacable también la aportación realizada por el sistema en lo referente a cubrir la demanda existente entre el personal de este tipo de compañías de herramientas capaces de producir nuevos contenidos elaborados a partir de los ya existentes.

Las futuras líneas de nuestro trabajo pasan por seguir profundizando por un lado en los estudios de los subsistemas de creación automática de fichas, resúmenes temáticos, e informes, por otro lado se espera mejorar la personalización de la capa de acceso a los archivos digitales, y por último todavía queda todo por decir en lo referente a la producción de nuevos contenidos. Los aspectos relacionados en cómo mejorar la coordinación de estos subsistemas entre sí de cara a optimizar resultados es otra de nuestras futuras líneas de investigación.

Referencias bibliográficas

Gandomi, A. , Murtaza H. (2015).Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137-144.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70.

Taylor, C. (1994). La ética de la autenticidad. Barcelona: Paidós.

Zikopoulos, P. &Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Corrigan, D.&Giles, J. (2013). Harness the power of big data: The IBM big data platform. McGraw-Hill.

Khan, M. A., Uddin, M. F., Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. IEEE Conference of the American Society for Engineering Education, 1-5.

Keim, D., Qu, H., Ma, K. L. (2013). Big-data visualization. IEEE Computer Graphics and Applications, 33(4), 20-21.

Markus, M. L. (2017). Datification, Organizational Strategy, and IS Research: What's the Score?. The Journal of Strategic Information Systems, 26(3), 233-241.

Lavoie, B. F. (2004). The open archival information system reference model: Introductory guide. Microform & imaging review, 33(2), 68-81.

Cambria, E., White, B. (2014).Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), 48-57.

Sonka, M., Hlavac, V., Boyle, R. (2014). Image processing, analysis, and machine vision. Cengage Learning.

Gold, B., Morgan, N., Ellis, D. (2011). Speech and audio signal processing: processing and perception of speech and music. John Wiley & Sons.

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies–an etymological note. Journal of Documentation, 59(1), 7-18.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220.

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Digital Library Federation, Council on Library and Information Resources.

Dextre Clarke, S. G., Zeng, M. L. (2012). From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly, 24(1).

Recio, M., Carlos, J., Sánchez Vigil, J. M., Serrada Gutiérrez, M. (2009). Nuevos paradigmas periodísticos y documentales en los periódicos digitales: estudio de casos en España. Investigación Bibliotecológica, 23(49), 43-65.

Moral, M. V. N. (2014). Sistemas de acceso y consulta en los diarios digitales españoles. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 28(62), 81-99.

Seaman, D. (2003). From isolation to integration: re-shaping the serials data silos. Serials, 16(2).

Guallar, J. (2011). La documentación en la prensa digital. Nuevas tendencias y perspectivas. Congreso Internacional de Ciberperiodismo y Web 2.0.

Garrido, A., Gómez, O., Ilarri, S., Mena, E. (2012). An experience developing a semantic annotation system in a media group. International Conference on Applications of Natural Language to Information Systems (NLDB), 333-338.

Korde, V., Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

Garrido, A.L, Bobed, C., Cardiel, O., Aleyxendri, A., Quilez, R.(2017) Optimization in Extractive Summarization Processes through Automatic Classification. International Conference of Computational Linguistics and Intelligent Text Processing (CICLING), pendiente de publicación.

Nenkova, A., McKeown, K. (2011). Automatic Summarization. Foundations and Trends in Information Retrieval, 5(2–3), 103-233

Gupta, V., Lehal, G. S. (2010). A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence, 2(3), 258-268.

Buey, M. G., Garrido, Á. L., Escudero, S., Trillo, R., Ilarri, S., Mena, E. (2014). SQX-Lib: Developing a Semantic Query Expansion System in a Media Group. European Conference of Information Retrieval (ECIR), 780-783.

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked data - The story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227.

Garrido, A. L., Sangiao, S., Cardiel, O. (2017). Improving the Generation of Infoboxes from Data Silos through Machine Learning and the use of Semantic Repositories. International Journal on Artificial Intelligence Tools, 26(05), 1760022.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. Scientific american, 284(5), 28-37.

Garrido, A. L., Gomez, O., Ilarri, S., Mena, E. (2011). NASS: News Annotation Semantic System. International Conference on Tools with Artificial Intelligence (ICTAI), 904-905.

Garrido, A. L., Buey, M. G., Ilarri, S., Mena, E. (2013). GEO-NASS: A semantic tagging experience from geographical data on the media. East European Conference on Advances in Databases and Information Systems (ADBIS), 56-69.

Garrido, A. L., Buey, M. G., Escudero, S., Peiro, A., Ilarri, S., Mena, E. (2014). The GENIE Project-A Semantic Pipeline for Automatic Document Categorisation. International Conference on Web Information System and Technologies (WEBIST), 161-171.

Rincón, J., Bobed, C., Garrido, A. L., Mena, E. (2014). SIWAM: Using Social Data to Semantically Assess the Difficulties in Mountain Activities. International Conference onWeb Information System and Technologies (WEBIST), 41-48.

Buey, M. G., Garrido, A. L., Bobed, C., Ilarri, S. (2016). The AIS Project: Boosting Information Extraction from Legal Documents by usingOntologies. International Conference on Agents and Artificial Intelligence (ICAART), 438-445.

Garrido, A. L., Buey, M. G., Muñoz, G., Casado-Rubio, J. L. (2016). Information Extraction on Weather Forecasts with Semantic Technologies. International Conference on Applications of Natural Language to Information Systems (NLDB), 140-151.

Garrido, A. L., Blázquez, P., Buey, M. G., Ilarri, S. (2015). Knowledge Obtention Combining Information Extraction Techniques with

Linked Data. International Conference on World Wide Web (WWW), 643-648.

Garrido, A. L., Ilarri, S., Sangiao, S., Gañán, A., Bean, A., Cardiel, O. (2016). NEREA: Named Entity Recognition and Disambiguation Exploiting Local Document Repositories. International Conference on Tools with Artificial Intelligence (ICTAI), 1035-1042.

Palmer, J. W., Eriksen, L. B. (1999). Digital newspapers explore marketing on the Internet. Communications of the ACM, 42(9), 32-40.

Markkula, M., Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. Information retrieval, 1(4), 259-285.

Boczkowski, P. J., Ferris, J. A. (2005). Multiple media, convergent processes, and divergent products: Organizational innovation in digital media production at a European firm. The Annals of the American Academy of Political and Social Science, 597(1), 32-47.

Deacon, D. (2007). Yesterday's papers and today's technology: Digital newspaper archives and 'push button' content analysis. European Journal of Communication, 22(1), 5-25.

Nicholson, B. (2013). The digital turn: Exploring the methodological possibilities of digital newspaper archives. Media History, 19(1), 59-73.

Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., Lorés, J. (2014). Neptuno: Semantic web technologies for a digital newspaper archive. European Semantic Web Symposium, 445-458.

Elragal, A., Päivärinta, T. (2017). Opening Digital Archives and Collections with Emerging Data Analytics Technology: A Research Agenda. Tidsskriftet Arkiv, 8(1).