ARTÍCULOS

TECNOLOGÍAS *BIG DATA* PARA ANÁLISIS Y RECUPERACIÓN DE IMÁGENES WEB

Sergio Rodríguez-Vaamonde, Ana-Isabel Torre-Bastida y Estibaliz Garrote

Sergio Rodríguez-Vaamonde es ingeniero de telecomunicaciones por la *Universidad del País Vasco* y diploma de estudios avanzados en ingeniería telemática. Doctorando en anotación automática y recuperación de imágenes en grandes colecciones de datos, cuenta con publicaciones en este ámbito. Trabaja como investigador en el grupo de procesamiento de imagen en el centro tecnológico *Tecnalia*, desarrollando productos de aplicación de las tecnologías de visión artificial a las TICs o el sector industrial.
http://orcid.org/0000-0003-1982-5128

sergio.rodriguez@tecnalia.com

Ana-Isabel Torre-Bastida es ingeniera informática por la *Universidad de Deusto* y máster en sistemas informáticos avanzados por la *Universidad del País Vasco*. Es doctoranda, centrando sus investigaciones en web semántica y *linked open data*, e investigadora colaboradora en el centro tecnológico *Tecnalia*, donde lleva a cabo trabajos de aplicación de las tecnologías semánticas y de *big data* a varios sectores.
http://orcid.org/0000-0003-3005-1100

isabel.torre@tecnalia.com

Estibaliz Garrote es doctora y licenciada en ciencias físicas e ingeniera en electrónica por la *Universidad del País Vasco*. Como especialista en aplicaciones de visión artificial ha participado en más de 50 proyectos de investigación nacional e internacional sobre control de calidad, procesos de fabricación, biometría, reciclado y accesibilidad. En los últimos años ha trabajado en el desarrollo de modelos neuroinspirados, colaborando en los grupos de Tomaso Poggio (*MIT*), John Mollon (*University of Cambridge*) y Thomas Serre (*Brown University*).
http://orcid.org/0000-0002-0268-0391

estibaliz.garrote@tecnalia.com

Tecnalia. División Industria y Transporte
Parque Tecnológico de Bizkaia
Ibaizabal Bidea, Edif. 202. 48170 Zamudio (Bizkaia), España

Resumen

Se aborda el análisis web desde el punto de vista de las imágenes, empleando tecnologías *big data*. Las imágenes cada vez tienen más peso en la web por lo que cualquier análisis que se realice deberá considerar este tipo de información. Los grandes volúmenes de imágenes existentes hacen necesaria la utilización de grandes infraestructuras de computación para realizar este tipo de trabajos, así como tecnologías de visión artificial específicas. Se muestran tecnologías *big data* que pueden ser utilizadas dentro del campo del análisis de imágenes a gran escala. Además, se propone una arquitectura que permite recuperar imágenes de una biblioteca de imágenes de forma eficiente y con un bajo coste computacional. Esta arquitectura puede servir como base para los análisis web e investigaciones que requieran un estudio detallado de las imágenes similares, sin la necesidad de disponer de hardware específico para ello.

Palabras clave

Big data, Imágenes, Procesamiento de imágenes, Visión artificial, Búsqueda, Búsqueda de imágenes, *Map-Reduce*, *LSH*, *Locality-sensitive hashing*.

Title: *Big data* technologies for image retrieval and analysis in web environments

Artículo recibido el 09-07-2014
Aceptación definitiva: 14-10-2014

Abstract

This paper addresses web analytics from the point of view of images, using big data technologies. Images are increasingly present on the web, and therefore any web analysis must consider them. The huge volume of available images requires the use of large computation infrastructures in order to generate this kind of analysis, as well as specific computer vision algorithms. Big data technologies are discussed that can be used for large-scale image processing. In addition, a novel distributed algorithm is proposed that can efficiently retrieve images from an online collection at low computational cost. This algorithm can be used in web analysis or future research requiring detailed image study, without requiring any special hardware.

Keywords

Big data, Images, Image processing, Computer vision, Search, Image search, *Map-Reduce*, LSH, Locality-sensitive hashing.

Rodríguez-Vaamonde, Sergio; **Torre-Bastida, Ana-Isabel**; **Garrote, Estibaliz** (2014). "Tecnologías *big data* para análisis y recuperación de imágenes web". *El profesional de la información*, v. 23, n. 6, noviembre-diciembre, pp. 567-574.

http://dx.doi.org/10.3145/epi.2014.nov.02

1. Introducción

Vivimos en un mundo totalmente digitalizado donde el número de usuarios, sensores y dispositivos electrónicos aumenta a cada segundo, en el que se estima que hay más teléfonos móviles conectados a internet que ordenadores (**Martin** *et al.*, 2013). Cada proceso digital o intercambio de información que éstos producen, genera a su vez una cantidad ingente de datos que se va sumando a la nube ya existente. Este fenómeno se ha denominado *big data* y es un concepto que se usa para referirse a los datos, a los retos y características especiales que éstos engloban y a las nuevas tecnologías desarrolladas para poder tratarlos. La magnitud del fenómeno es tal que los datos generados durante dos días en 2011, por ejemplo, fueron más que los acumulados desde el origen de la civilización hasta principios de 2003 (**Lyman**; **Varian**, 2003). Estas magnitudes asustan y por ello la comunidad científica lleva mucho tiempo buscando soluciones al problema de cómo tratarlos.

> Asumir el problema del *big data* es enfrentarse a tres retos fundamentales: almacenamiento, procesamiento y acceso

El concepto de *big data* engloba grandes cantidades de datos de distintos dominios (que pueden ser complejos, crecientes y variables), junto con las técnicas necesarias para poderlos recolectar, almacenar, gestionar y analizar. Esta definición amplía la inicialmente establecida por *Gartner* en 2012 (**Beyer**; **Laney**, 2012). Para cualquier organización que se enfrenta al problema de tratamiento de sus datos, *big data* representa la frontera que hay que traspasar para hacer efectiva su gestión en tiempo y coste. Cualquiera que asuma este problema se enfrenta a tres retos fundamentales a resolver: almacenamiento, procesamiento y acceso.

Los datos recolectados pueden clasificarse según su naturaleza, formato y estructura en dos grandes grupos: estructurados, que siguen un modelo que proporciona una metainformación que ayuda en el procesamiento; y no estructurados o datos en crudo, que son más arduos de procesar y por lo tanto su análisis conlleva más tiempo y esfuerzo. En este segundo grupo se encuentran las imágenes y los vídeos, un tipo de información que cada vez está tomando más relevancia en la Web (**Rodríguez-Vaamonde**; **Ruiz-Ibáñez**; **González-Rodríguez**, 2012). Un ejemplo que muestra la tendencia de crecimiento del volumen de imágenes en internet son los 60 millones de fotografías intercambiadas en la red social *Instagram* cada día (*Instagram*, 2014).

2. *Big data* y la recuperación de imágenes web por su contenido

Desde una perspectiva *big data*, en el ámbito web muchas veces es necesario recuperar una imagen de una biblioteca distribuida. En el análisis de una determinada página web son muchos los escenarios que se pueden encontrar: es posible requerir la obtención de la fuente de las imágenes que aparecen en dicha página, también es posible analizar si sus imágenes han sido publicadas en terceras páginas, o reconocer a una persona en las fotografías de dicha web.

En todos estos casos el problema central es analizar el contenido de las imágenes de interés localizadas en una web para buscar y encontrar una imagen en una biblioteca de imágenes determinadas. Como se puede intuir, cuando se está hablando de grandes cantidades de datos la búsqueda en miles o millones de imágenes será una búsqueda computacionalmente cara y difícil de ejecutar en equipos informáticos básicos.

La pregunta clave en este contexto es si la tecnología evoluciona y crece a la misma velocidad que lo están haciendo estos datos. Para el campo en el que se centra este artículo, el análisis de imágenes, la respuesta es no. Trabajos como los de **Guo** y **Dyer** (2005) o **White** *et al.* (2010), explican como la infraestructura (recursos de almacenamiento y computación necesarios) es difícilmente adquirible para llevar a cabo aplicaciones de análisis de imagen a gran escala, por lo tanto existen pocos investigadores que se aventuren en esta área. Esto a su vez provoca que el número de trabajos dedicados al procesamiento de grandes volúmenes de imágenes o videos sea relativamente escaso y monopolizado por grandes grupos de investigación.

El tratamiento de la imagen, por su relevancia y complejidad, debe tener un espacio propio en el mundo *big data*. Si prosiguen las tendencias actuales y la sentencia de que "los datos son el nuevo petróleo del siglo XXI" acuñada por

Andreas Weigend es cierta, el potencial que se pueda extraer de estos datos dependerá irremediablemente de las tecnologías y algoritmos desarrollados para ello, y en el caso de las imágenes está claro que se necesita potenciar ambos.

Para ello presentamos en primer lugar un resumen de las principales tecnologías *big data* existentes y su aplicabilidad a los datos en formato imagen. Abarcar el 100% de las tecnologías queda fuera del alcance de este estudio, por lo que nos centraremos en el caso particular de la recuperación rápida de imágenes en base a su contenido y presentaremos una arquitectura que contiene los ingredientes necesarios para hacer búsquedas rápidas de imágenes que permitan un amplio abanico de análisis web sin necesidad de una gran inversión en infraestructura.

> En el caso de las imágenes, la tecnología no evoluciona y crece a la misma velocidad que lo están haciendo los datos

3. Tecnologías *big data* aplicables a imágenes

Las necesidades que deben cumplir las tecnologías del *big data* se basan en el procesamiento eficiente de grandes cantidades de datos con un tiempo reducido o tolerable. La complejidad que añade el hecho de que los datos se encuentren en formato imagen es otra variable a considerar.

En general las tecnologías de mayor relevancia para datos no estructurados como imágenes son las bases de datos *NoSQL* (**Leavitt**, 2010) y los modelos de programación *Map-Reduce* (**Bajcsy** *et al.*, 2013), ambas relacionadas con el procesamiento de datos en lotes. Por otro lado están los *CEP* (*complex event processing*), los *IMDG* (*in-memory data grids*), o los sistemas de computación distribuida, para el procesamiento de datos en tiempo real. En la tabla 1 se muestra una taxonomía de estas tecnologías en forma de cuadrante ordenado por volumen de datos y tiempos de rendimiento de las tecnologías *big data*.

En los siguientes apartados se detallan las tecnologías más relevantes en los ámbitos anteriores, centrándose en aquellas especialmente útiles para el procesamiento de imágenes a gran escala.

Tabla 1. Taxonomía de tecnologías de *big data* aplicables al procesamiento de imágenes

Volumen de datos	Pocos	Muchos
Tiempo real	**Analítica stream** -Proceso de eventos complejos -Base de datos en memoria	**Analítica en tiempo real** -Grid de datos en memoria -Plataforma especializada
Lotes	**Analítica de operación** -OLPT / OLAP -Base de datos relacional	**Analítica en lotes** -Plataforma *Map-Reduce* -Base de datos *NoSQL*

3.1. Bases de datos *NoSQL*

Se pueden definir como una nueva generación de almacenes de datos, que cumplen alguno de estos puntos: ser no relacionales, distribuidos, escalables o de código abierto. Tienen características comunes como:

- esquema libre, no siguen un modelo de datos rígido o uniforme;
- facilidad de replicación, lo que mejora la tolerancia a fallos y redundancia;
- API de acceso sencilla para el acceso y manipulación de datos;
- capacidad de tratamiento de una enorme cantidad de datos;
- estructura distribuida y escalabilidad horizontal.

Se clasifican en diferentes tipos. Una de las clasificaciones más completas es la de **Stephen Yen** (2009), que se muestra en la tabla 2 junto con diferentes alternativas para cada tipo.

Estos repositorios son la nueva opción de persistencia para aquellos casos en los que las necesidades de distribución y disponibilidad de los datos superen a la necesidad de un esquema complejo y rígido. Es el caso de las imágenes de las páginas web, ya que su generación es distribuida y el acceso a las mismas debe ser altamente disponible.

> Las tecnologías de mayor relevancia para imágenes son las bases de datos *NoSQL* y los modelos de programación *Map-Reduce* para el procesamiento de datos en lotes, y *CEP* e *IMDG* para el procesamiento en tiempo real

3.2. Procesamiento en lotes mediante *Map-Reduce*

En segundo lugar se encuentran los modelos de programación para la construcción de aplicaciones distribuidas. El objetivo de estos modelos es ejecutar un procesamiento determinado sobre un conjunto de datos (llamado lote de datos) de forma distribuida en diferentes localizaciones. Los modelos más extendidos de procesamiento en lotes se basan en dos operaciones básicas: *Map* y *Reduce* (**Dean**; **Ghe-**

Tabla 2. Clasificación de bases de datos *NoSQL* y productos existentes

Clases de bases de datos *NoSQL*	Ejemplos
Almacenamiento clave-valor en cache	*Memcached, Repcached, Coherence*
Almacenamiento clave-valor	*KeySpace, Flare, Schema Fre*
Almacenamiento clave-valor eventualmente consistente	*Dynamo, Voldemort, Dynomite*
Almacenamiento clave-valor ordenado	*LightCloud, MemCacheDB, Tokyo TIrant*
Servidores de estructuras de datos	*Redis*
Bases de datos de objetos	*SopeDB, DB40, Shoal*
Repositorio de documentos	*Couch DB, Mongo DB, Scalaris*
Repositorio de columnas	*Blgtable, Hbase, Cassandra DB, Hipertable*

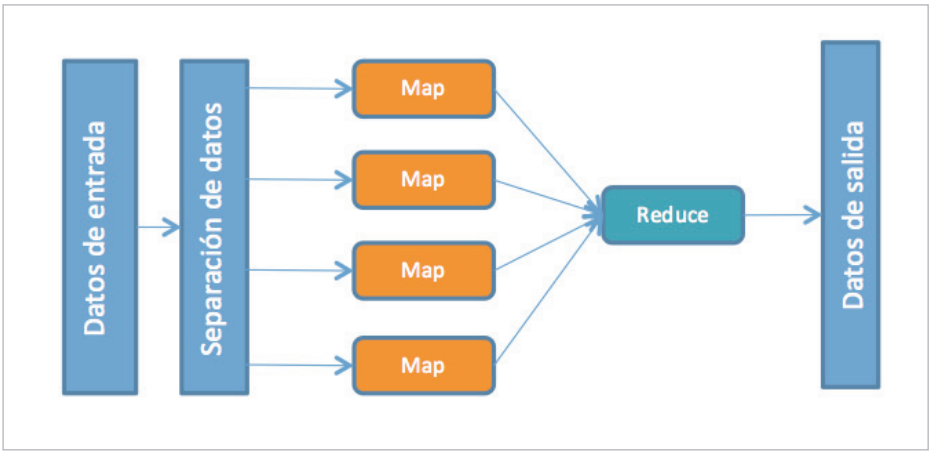

Figura 1. Diagrama de funcionamiento del modelo *Map-Reduce*

mawat, 2008). En la primera se distribuyen los datos por los nodos del clúster en forma de pares clave-valor y en la segunda se recogen aquellos pares clave-valor que cumplan los criterios del resultado deseado (figura 1).

Uno de los sistemas que ha popularizado esta metodología de trabajo ha sido *Hadoop*, un framework de código abierto para desarrollar y ejecutar aplicaciones distribuidas que procesen una gran cantidad de datos. Su simplicidad y fácil acceso le proporcionan una ventaja en el desarrollo y ejecución de grandes programas distribuidos. La forma de trabajo de *Hadoop* consiste en que los clientes envían sus trabajos a la nube, formada por el clúster de máquinas y es esta la que se encarga de su ejecución mediante el modelo *Map-Reduce* siendo transparente para el usuario la forma en que se haga. Es de un mecanismo muy utilizado en tareas que implican trabajos con grandes volúmenes de datos, como pueden ser el razonamiento de información (**Urbani** *et al.*, 2009) o la extracción de información de las imágenes (**Yan**; **Huang**, 2014).

> Para un sistema de visión artificial, lo realmente importante es la representación de la imagen en base a un descriptor visual que sea lo más compacto posible, solventando el problema del almacenamiento

3.3. Tiempo real

En este grupo se enmarcan aquellas soluciones capaces de cumplir con los requisitos de temporalidad del análisis de datos. El resultado del análisis debe de ser inmediato, en tiempo real, para que no pierda su valor y quede obsoleto. Aquí se encuentran dos opciones: los *CEP* (*complex event processing*) y los *IMDG* (*in-memory data grids*).

Los *CEP* son una tecnología de red emergente, que permite adquirir información a partir de sistemas distribuidos de mensajes, bases de datos o aplicaciones, todo ello en tiempo real. Permiten a las organizaciones definir, manejar y predecir eventos y condiciones en una red compleja y heterogénea de actores e informantes. Se centran en el procesamiento de eventos como un método para rastrear y analizar grandes volúmenes de información sobre diferentes fenómenos y acciones producidas en un contexto.

En los últimos tiempos ha surgido un nuevo concepto denominado *IMDG*, con el que se denomina a las tecnologías capaces de procesar grandes conjuntos de datos con baja latencia en un contexto de tiempo real. Para ello, estas tecnologías paralelizan el almacenamiento de los datos gracias a su alojamiento particionado en memoria principal (haciendo que los datos estén cerca de la aplicación que los consume).

En este nuevo contexto se combinan técnicas de cacheo distribuido con potentes herramientas de análisis y gestión, para ofrecer una solución completa que permita gestionar grandes cantidades de datos de naturaleza volátil mediante un clúster de ordenadores.

En la tabla 3 se muestran una serie de productos, así como su fabricante y el lenguaje de programación que aceptan.

Esto ofrece grandes posibilidades a los usuarios, ya que la toma de decisiones será en tiempo real, mejorando la productividad y experiencia de uso**.**

4. Propuesta de uso de tecnologías de *big data* aplicadas a imagen

Una vez vistas las tecnologías que se pueden emplear, en este apartado se propone una arquitectura software para construir sistemas más complejos y completos de análisis web en base a imágenes. El objetivo de esta propuesta es doble:

- aliviar la carga computacional y ser capaces de realizar búsquedas rápidas sobre grandes colecciones de imágenes;
- demostrar que el uso de tecnologías genéricas *big data* no requiere disponer de grandes infraestructuras de computación para el procesamiento y recuperación de grandes cantidades de imágenes, eliminando la barrera de entrada de numerosos investigadores a este tipo de estudios.

Tabla 3. Ejemplos de sistemas *in-memory data grids* (*IMDG*)

Producto	Lenguaje	Empresa
Oracle Coherence	Java	*Oracle*
Ehcache BigMemory	Java	*Terracota*
GemFire	Java	*VMware*
JBoss Infinispan	Java	*RedHat*
Ncache	.net	*AlaChisoft*
WebSphere eXtreme	Java	*IBM*

A continuación se definirá el flujo de trabajo mínimo para el análisis del contenido de una imagen, se describirá la arquitectura propuesta para la búsqueda de imágenes a escala web y se mostrará cómo la propuesta puede ser considerada como base para siguientes trabajos en este ámbito ya que mejora en gran medida los sistemas básicos actuales.

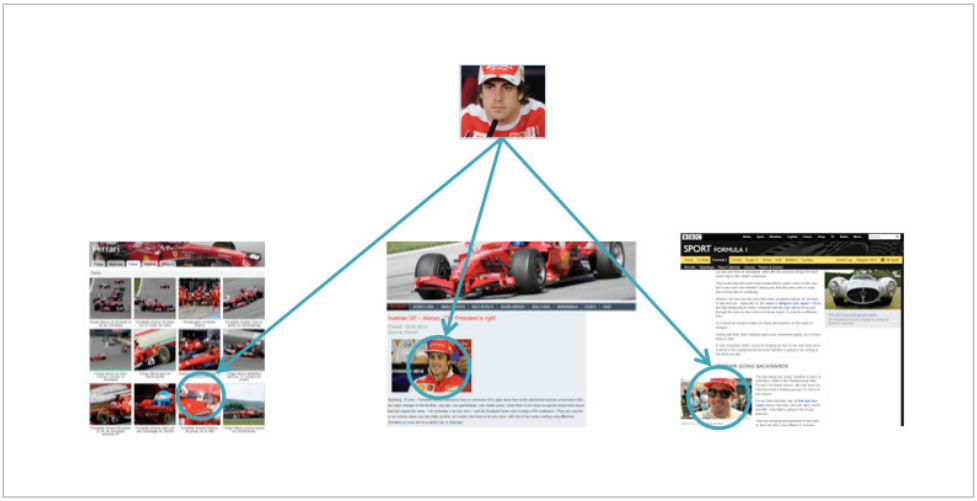

Figura 2. Ejemplo de uso en el que se quiere buscar las páginas web con una fotografía de Fernando Alonso para analizar su aparición en internet

4.1. Análisis de imagen, almacenamiento y recuperación

Una imagen se describe como un conjunto de píxeles en formato matricial que contiene información sobre la escena que representa. Este conjunto de píxeles en crudo no permite a los sistemas de computación conocer el contenido, sino que es necesario generar descriptores visuales de la imagen.

Un descriptor visual se puede definir como la representación matemática del contenido de la imagen. Así, el descriptor visual más básico es un histograma de color, donde en un conjunto limitado de valores representa estadísticamente el número de píxeles contabilizados de cada color en una imagen. Además de este, existen múltiples descriptores que tratan de representar una información de mayor nivel semántico como los bordes presentes (**Lowe**, 2004) o las formas básicas que aparecen en la imagen (**Ferrari**; **Jurie**; **Schmid**, 2010).

Para un sistema de visión artificial, lo realmente importante es la representación de la imagen en base a un descriptor visual que sea lo más compacto y representativo posible. En este punto entra en juego el concepto de almacenamiento. Si bien una imagen puede ocupar unos cientos o miles de kilobytes, un descriptor apenas ocupa unas decenas. El problema relacionado con *big data* es que en grandes bibliotecas de imágenes (como puede ser la propia web), existirán millones de estos descriptores que representan las imágenes y por ello el espacio de almacenamiento requerido será inmenso. En este punto es donde las tecnologías *big data* son de gran utilidad.

Una vez se tiene toda la información almacenada, la clave es el acceso a la misma. En el caso que se está analizando el objetivo es recuperar las imágenes por su contenido y, específicamente, se desea recuperar las que sean lo más similares visualmente a una imagen dada (figura 2).

Para poder realizar esta búsqueda, el procedimiento directo es comparar los descriptores visuales con una métrica que mida el grado de similitud por su contenido visual. Estas métricas no son más que distancias computadas para cada uno de los dos vectores *x* e *y* a comparar y algunas de las más utilizadas se presentan en la tabla 3.

El problema de esta aproximación es claro: si se tienen millones de imágenes, comparar y calcular una distancia para cada vector de forma directa es una tarea que llevará una cantidad excesiva de tiempo.

Debido a estos dos principales problemas derivados de los datos y comunes a cualquier sistema *big data*, es necesario utilizar tecnologías adaptadas al uso de grandes cantidades de datos.

4.2. Arquitectura software de recuperación rápida de imágenes similares

En el apartado anterior se ha mostrado que la aproximación directa de la recuperación de imágenes similares no es algo factible cuando se trata de bibliotecas de un gran volumen como pueden ser las relacionadas con el análisis web. Para poder hacer este tipo de análisis y recuperación, se propone la utilización de tecnologías *big data* que permitan un acceso eficiente a toda la información disponible en imágenes. La propuesta se puede resumir en la figura 3.

El primer problema es el almacenamiento eficiente de las imágenes de entrada. Estas imágenes pueden entrar al sistema de diferentes formas, en función del análisis al que se esté dedicando esta arquitectura. En el caso de análisis de imágenes web, la entrada sería por un robot automático que obtuviese las imágenes de las páginas web. En el caso

Tabla 4. Distancias utilizadas para medida de similitud entre descriptores visuales de imágenes

Distancia	Fórmula
Euclídea	$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
Coseno	$d(x,y)=1-\cos(\theta)=1-\frac{x\,y'}{\sqrt{(x\,x')(y\,y')}}$
Chi Cuadrado	$d(x,y)=\frac{1}{2}\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert^2}{x_i+y_i}$

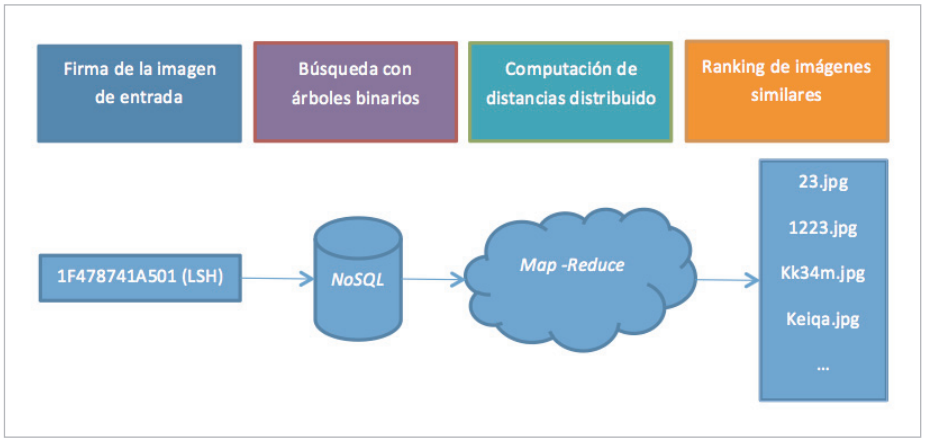

Figura 3. Fases de la arquitectura propuesta para la recuperación de imágenes similares a gran escala

de búsqueda de una imagen concreta, se dispondría de una base de datos de imágenes y habría que introducir manualmente la nueva imagen de consulta.

En cualquier caso el problema principal consiste en almacenar las imágenes y, además de la propia imagen, es crucial almacenar sus descriptores visuales. Ya que esta no es una información estática y es dependiente del análisis o búsqueda a realizar, es necesario disponer de un almacén de datos lo suficientemente flexible. Además se debe poder guardar una gran colección de datos, como son el origen de las imágenes (web, fecha de acceso, etc.), anotaciones manuales o comentarios. Para ello, cualquier base de datos *NoSQL* de las presentadas en la tabla 1 es una opción válida ya que en general permiten disponer de un esquema flexible y son capaces de tener réplicas o nodos distribuidos. Esta última característica la hace ideal para el análisis de imágenes web en cualquier punto geográfico, ya que permite replicaciones distribuidas en todo el mundo.

> Para un sistema de búsqueda por similitud no es factible computar para cada imagen de la colección la distancia a la imagen de consulta; es fundamental utilizar algún sistema de búsqueda aproximada de las más similares

Sobre este almacén de datos *NoSQL* es necesario construir el sistema de recuperación de figuras. No es factible computar para cada imagen de la colección la distancia a la imagen de consulta. Por ello es fundamental utilizar algún sistema de búsqueda aproximada de las más cercanas.

El algoritmo más utilizado en el campo del análisis de imágenes (**Kulis**; **Grauman**, 2009) es *locality-sensitive hashing* (LSH) (**Slaney**; **Casey**, 2008), por lo que éste será el algoritmo base para la recuperación de figuras. Este algoritmo permite generar una firma numérica (o *hash*) para cada descriptor o conjunto de descriptores de imagen, de tal forma que aquellos vectores que tengan una distancia euclídea muy baja, y por tanto sean vectores muy similares, posean la misma firma numérica. Este tipo de algoritmos es muy útil para encontrar entradas similares dentro de grandes colecciones de datos, por ejemplo buscando páginas web similares (**Slaney**; **Casey**, 2008), por lo que es lógica su aplicación a los descriptores visuales de imágenes.

Una vez se tienen las firmas para todas las imágenes de la colección, la búsqueda de las similares es sencilla: dada una imagen de consulta, se generará su *hash*. Con cada firma se buscará en toda la base de datos las que posean la misma firma y todas ellas serán las imágenes más similares.

Para hacer esta comparación, se puede pensar en que existe el mismo problema de búsqueda que antes, pero nada más lejos de la realidad. Las bases de datos *NoSQL* actuales para *big data*, disponen de técnicas de indexación y búsqueda rápida de un número único, como puede ser el algoritmo de búsqueda en árbol binario de la base de datos *NoSQL MongoDB* o el uso de cualquier tecnología *IMDG* de las propuestas. Por ello la búsqueda ya no se circunscribe a calcular una distancia entre vectores sino a usar una arquitectura de índices para encontrar un número concreto.

Tras este paso ya se dispone de un conjunto de imágenes de la biblioteca similares a la de la entrada. En función de la analítica web que se esté ejecutando, quizá conocer este número es suficiente. En muchos casos la recuperación de figuras tiene como objetivo aquella que más se parece a la de entrada. En este caso, es obligatorio calcular la distancia concreta, pero ya que se ha obtenido un conjunto de imágenes parecidas, se puede calcular la distancia sobre ese conjunto de unas decenas de imágenes similares en unos pocos segundos, en vez de sobre el total de la biblioteca.

Para este último paso, también se va a aprovechar el almacén de datos *NoSQL* distribuido propuesto en el inicio. Ya que las figuras pueden estar almacenadas en localizaciones diferentes y que cada cálculo de la distancia entre la imagen de consulta y la similar es independiente, es posible usar el paradigma *Map-Reduce* expuesto con anterioridad. Este modelo permitirá en la función *Map* el cálculo de la distancia entre cada imagen similar y la de consulta, ejecutándose en cada nodo de la red distribuida de almacenamiento. Por otro lado, el método *Reduce* se encargará de ordenar todas las distancias y podrá generar el ranking final de imágenes similares útiles para la analítica.

5. Resultados y conclusiones

En este artículo se ha mostrado cómo es necesario aplicar tecnologías de *big data* cuando se trata del almacenamiento y acceso de grandes volúmenes de imágenes, como es el caso del análisis de imágenes en la web. Se han expuesto un conjunto de tecnologías que potencialmente pueden ser aplicables a este campo y también se ha realizado una propuesta de arquitectura software que puede ser utiliza-

da en cualquier análisis web que requiera conocer imágenes presentes en las páginas web.

Para comprobar si esta arquitectura es la más adecuada, se ha realizado un experimento comparativo de diferentes métodos de recuperación de imágenes similares del estado del arte. Para ello, se ha utilizado la base de datos *ImageNet* (**Deng** *et al.*, 2009), que posee más de 14 millones de imágenes. De ellas se han generado diferentes subconjuntos aleatorios, en los que se han agrupado las imágenes de cada una de las 7 pruebas realizadas. En cada prueba el tamaño del conjunto de imágenes que conformaba la base de datos ha sido incremental: empezando con 100, cada prueba ha añadido 200.000 adicionales hasta 1,2 millones. Con estas bases de datos se han ejecutado 10 búsquedas de imágenes similares y se ha medido el tiempo (en segundos) que se ha tardado en computar la distancia a todos los documentos, almacenando el valor del tiempo medio en esas 10 imágenes. Todas las pruebas se han realizado en un único PC, con 16 núcleos de computación a 3GHz, con 32GB de memoria RAM.

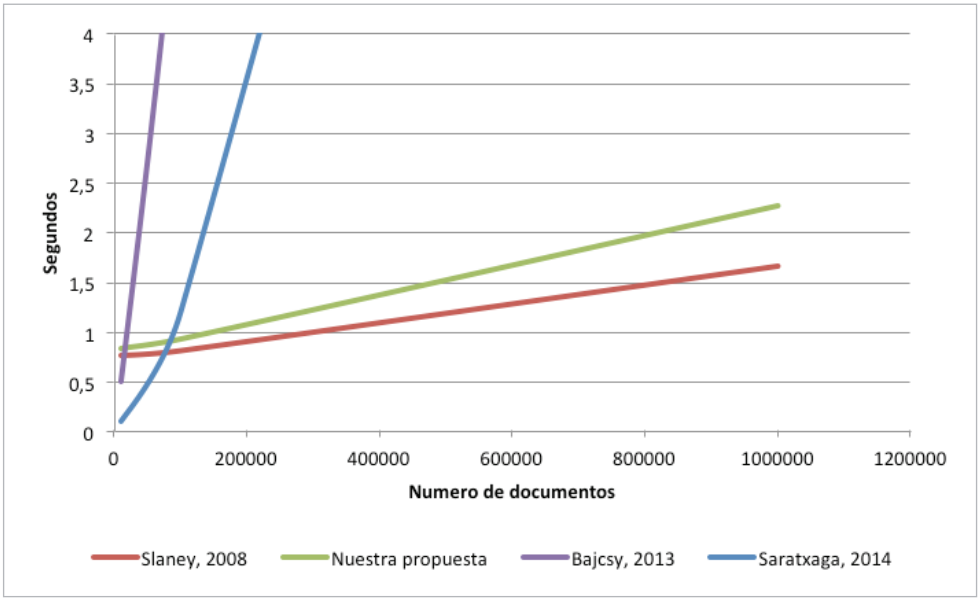

Figura 4. Comparativa de tiempos medios de búsqueda de una imagen usando cuatro aproximaciones diferentes

Esta operación de búsqueda se ha realizado para cuatro algoritmos, tres propuestos en el estado del arte, y el cuarto es la propuesta realizada:

- el primer algoritmo es el utilizado por la mayor parte de los trabajos, y trata de cargar en memoria RAM todas las imágenes y computar la distancia de forma exhaustiva (**Saratxaga** *et al.*, 2014);
- el segundo trabajo es computar la distancia exhaustiva utilizando el algoritmo *Map-Reduce* (**Bajcsy** *et al.*, 2013) donde cada nodo es un núcleo de procesamiento del PC;
- el tercero es utilizar la aproximación LSH de forma directa con los *hashes* cargados en memoria RAM (**Slaney**; **Casey**, 2008);
- la propuesta que realizamos es utilizar LSH para representar las imágenes, un *Binary Tree* para hacer una búsqueda aproximada y posteriormente *Map-Reduce* para obtener de forma precisa un subconjunto de distancias.

En todos ellos se ha comprobado cómo el resultado de las 10 imágenes más similares de la base de datos es el mismo, por lo que sólo varía el tiempo de ejecución, no la precisión de la búsqueda.

En la figura 4 se muestra cómo el uso de la aproximación propuesta es muy útil cuando aumenta el número de imágenes.

En el eje horizontal, se ven diferentes números de imágenes que se han introducido en la bases de datos de validación. En el eje vertical se ve el tiempo en segundos que se ha tardado de media en buscar una única imagen similar dentro de la biblioteca de validación.

El resultado muestra que el tiempo de búsqueda en las técnicas de búsqueda exacta aumenta de forma exponencial con el número de figuras almacenado, mientras que en las búsquedas aproximadas usando LSH el aumento es lineal. Como es de esperar, el acceso a memoria es mucho más rápido que el uso de árboles de indexación distribuidos, por ello la aproximación de LSH en memoria (**Slaney**; **Casey**, 2008) es la que menos tiempo tarda. Aun así, esto no es eficiente para los análisis web distribuidos, ya que la multitud de nodos hace que no sea viable mantener todo en memoria. Por ello, la aproximación distribuida propuesta es la más útil para este tipo de analíticas.

La conclusión de este artículo está claramente alineada con el impulso de la utilización de tecnologías *big data* en el campo del análisis de las imágenes para recuperación a escala web. La motivación principal es que cada vez existen más datos en forma de imágenes y videos, y su investigación y análisis se monopoliza en los grandes grupos de investigación con grandes infraestructuras dedicadas al manejo de estos datos. En este artículo se muestra cómo las tecnologías *big data* más populares se pueden aplicar a la búsqueda de documentos en formato de imagen. Por tanto el uso de herramientas estándar, principalmente de código abierto, y optimizaciones de las mismas generan que sea posible utilizar sistemas de computación clásicos, y no necesariamente grandes clústeres de computación, para la analítica de estos datos.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el *Gobierno Vasco* dentro del proyecto *Smartur* bajo el programa de investigación fundamental *Etortek*. También ha sido parcialmente financiado por la *Comisión Europea* bajo el proyecto *Biopool*. Los autores quieren agradecer el trabajo a todos los socios de este proyecto.
http://www.biopoolproject.eu

6. Bibliografía

Bajcsy, Peter; **Vandecreme, Antoine**; **Amelot, Julien**; **Nguyen, Phuong**; **Chalfoun, Joe**; **Brady, Mary** (2013). "Terabyte-sized image computations on Hadoop cluster platforms". En: *IEEE Intl conf on big data*, pp. 729-737.
http://dx.doi.org/10.1109/BigData.2013.6691645

Beyer, Mark A.; **Laney, Douglas** (2012). *The importance of big data: A definition*. Gartner.
http://www.gartner.com/id=2057415

Dean, Jeffrey; **Ghemawat, Sanjay** (2008). "MapReduce: simplified data processing on large clusters". *Communications of the ACM*, v. 51, n. 1, pp. 107-113.
http://static.googleusercontent.com/media/research.google.com/es//archive/mapreduce-osdi04.pdf

Deng, Jia; **Dong, Wei**; **Socher, Richard**; **Li, Li-Jia**; **Li, Kai**; **Fei-Fei, Li** (2009). "ImageNet: A large-scale hierarchical image database". *IEEE Computer vision and pattern recognition* (*CVPR*).
http://dx.doi.org/10.1109/CVPR.2009.5206848

Ferrari, Vittorio; **Jurie, Frederic**; **Schmid, Cordelia** (2010). "From images to shape models for object detection". *Intl journal of computer vision*, v. 87, n. 3, pp. 284-303.
https://lear.inrialpes.fr/pubs/2010/FJS10/vitto_final_ijcv.pdf
http://dx.doi.org/10.1007/s11263-009-0270-9

Guo, Guodong; **Dyer, Charles R.** (2005). "Learning from examples in the small sample case: face expression recognition". *IEEE Transactions on systems, man, and cybernetics, Part B: Cybernetics*, v. 35, n. 3, pp. 477-488.
ftp://ftp.cs.wisc.edu/computer-vision/repository/PDF/guo.2005.smc.pdf
http://dx.doi.org/10.1109/TSMCB.2005.846658

Instagram (2014). *Instagram press stats*.
http://instagram.com/press

Kulis, Brian; **Grauman, Kristen** (2009). "Kernelized locality-sensitive hashing for scalable image search". En: *IEEE 12th Intl conf on computer vision*, pp. 2130-2137.
http://www.cs.utexas.edu/~grauman/papers/iccv2009_klsh.pdf
http://dx.doi.org/10.1109/ICCV.2009.5459466

Leavitt, Neal (2010). "Will NoSQL databases live up to their promise?". *IEEE Computer*, v. 43, n. 2, pp. 12-14.
http://leavcom.com/pdf/NoSQL.pdf

Lowe, David G. (2004). "Distinctive image features from scale-invariant keypoints". *Intl journal of computer vision*, v. 60, n. 2, pp. 91-110.
http://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf
http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

Lyman, Peter; **Varian, Hal R.** (2003). *How much information?* Berkeley, CA: University of California at Berkeley, School of Information Management and Systems.
http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf

Martín, David; **López-de-Ipiña, Diego**; **Alzua-Sorzabal, Aurkene**; **Lamsfus, Carlos**; **Torres-Manzanera, Emilio** (2013). "A methodology and a web platform for the collaborative development of context-aware systems". *Sensors*, v. 13, n. 5, pp. 6032-6053.
http://www.mdpi.com/1424-8220/13/5/6032
http://dx.doi.org/10.3390/s130506032

Rodríguez-Vaamonde, Sergio; **Ruiz-Ibáñez, Pilar**; **González-Rodríguez, Marta** (2011). "Uso combinado de tecnologías semánticas y análisis visual para la anotación automática de imágenes y su recuperación". *El profesional de la información*, v. 21, n. 1, enero-febrero, pp. 27-33.
http://www.computervisionbytecnalia.com/wp-content/uploads/2012/12/2012_JointUseSemantics.pdf
http://dx.doi.org/10.3145/epi.2012.ene.04

Saratxaga, Cristina L.; **Picón, Artzai**; **Rodríguez-Vaamonde, Sergio**; **López-Carrera, Ángel**; **Echazarra, Jone**; **Bereciartua, Arantza**; **Garrote, Estibaliz** (2014). "Plataforma de búsqueda de imágenes histológicas por similitud visual". En: *XVII Congreso nacional de informática de la salud*.
http://www.computervisionbytecnalia.com/wp-content/uploads/2014/04/PlataformaBusquedadeImagenesHistologicas.pdf

Slaney, Malcolm; **Casey, Michael** (2008). "Locality-sensitive hashing for finding nearest neighbors". *IEEE Signal processing magazine*, v. 25, n. 2, pp. 128-131.
http://web.iitd.ac.in/~sumeet/Slaney2008-LSHTutorial.pdf

Urbani, Jacopo; **Kotoulas, Spyros**; **Oren, Eyal**; **Van-Harmelen, Fran** (2009). "Scalable distributed reasoning using MapReduce". En: *ISWC '09 Procs of the 8th Intl semantic web conf*, pp. 634-649.
http://www.few.vu.nl/~jui200/papers/ISWC09-Urbani.pdf
http://dx.doi.org/10.1007/978-3-642-04930-9_40

White, Brandyn; **Yeh, Tom**; **Lin, Jimmy**; **Davis, Larry** (2010). "Web-scale computer vision using mapreduce for multimedia data mining". En: *Procs of the 10th Intl workshop on multimedia data mining*, pp. 1-10.
http://www.umiacs.umd.edu/~lsd/papers/brandyn-kdd-cloud.pdf
http://dx.doi.org/10.1145/1814245.1814254

Yan, Yuzhong; **Huang, Lei** (2014). "Large-scale image processing research cloud". En: *Cloud computing 2014, The 5th Intl conf on cloud computing, grids, and virtualization*, pp. 88-93.
http://www.thinkmind.org/index.php?view=article&articleid=cloud_computing_2014_4_20_20069

Yen, Stephen (2009). "NoSQL is a horseless carriage". NoSQL Oakland.