IMAGENES SIMILARES DE FACEBOOK

Carolina Guevara[1], Jos Fajardo[2], Hemerson Morn[3], Felix Drouet[4]

[1] Universidad de Guayaquil
rosa.guevarap, jose.fajardod, hemerson.moranp, felix.drouetv @ug.edu.ec

***Abstract*—El uso de la coincidencia grfica aproximada para la minera subgrafo frecuente ha sido identificado en diferentes aplicaciones como una necesidad. Para satisfacer esta necesidad, varios algoritmos de minera se han desarrollado y se han utilizado en diferentes dominio de la ciencia. En este trabajo, se presenta un nuevo marco para la clasificacin de imgenes basado en el grfico, donde se utiliza un algoritmo para la minera subgrafo frecuente. En este momento, la correspondencia aproximada grfica se aproxima nos permiten desarrollar clasificadores robustos en presencia de distorsiones en este tipo de datos. La experimentacin se realiz sobre una coleccin de imgenes real, donde cada imagen se representa como un grfico. Los resultados de esta experimentacin muestran la utilidad de los algoritmos aproximados y el marco propuesto.**

Palabras clave: minera de grafos aproximados; cotejo aproximado de grafos; representacin de imgenes; clasificacin de imgenes; seleccin de caractersticas.

I. INTRODUCCION

En los ltimos aos se ha incrementado la necesidad de convertir grandes volmenes de datos en informacin til. Los objetos en muchas de estas bases de datos estn o pudieran estar representadas como grafos. Como respuesta a esta necesidad, varios autores han desarrollado tcnicas y mtodos para procesar estas bases de datos (JIMNEZ etal., 2010). Un ejemplo de estas tcnicas es el descubrimiento de patrones frecuentes (YUN y RYU, 2011).

La minera de patrones frecuentes, especialmente la deteccin de subgrafos frecuentes en colecciones de grafos es un problema en tareas de minera de grafos (GAGO-ALONSO et al., 2009; NIJSSEN and KOK, 2004; YAN and HUAN, 2002). En minera de subgrafos frecuentes, existen dos enfoques para evaluar la similitud de grafos, conocido como cotejo exacto de grafos y cotejo aproximado de grafos.

El cotejo exacto de grafos ha sido satisfactoriamente utilizado en diversas aplicaciones (EICHINGER and BHM, 2010; GAGO-ALONSO et al., 2010; JIANG et al., 2010a); sin embargo, existen problemas concretos donde un cotejo exacto no es aplicable con resultados satisfactorios (HOLDER et al., 1992). En ocasiones, los subgrafos muestran ligeras diferencias en los datos. Un ejemplo de estas diferencias se puede ver en el procesamiento de imgenes, donde estas diferencias pueden estar debido a ruido y distorsiones, o simplemente pueden presentar ligeras diferencias espaciales entre instancias de objetos iguales.

Esto significa que se debe tolerar cierto nivel de distorsiones geomtricas, variaciones semnticas o desajustes entre vrtices o aristas, mientras se realiza la bsqueda de los subgrafos frecuentes. Por este motivo, varios autores han expresado la necesidad del uso del cotejo aproximado entre grafos para la minera de subgrafos frecuentes en colecciones de grafos (BORGELT and BERTHOLD, 2002;HOSSAIN and ANGRYK, 2007; KETKAR et al., 2006; KOYUTRK et al., 2004). Estos autores defienden la idea de que se pudieran detectar subgrafos frecuentes con mayor inters para aplicaciones y usuarios.

En respuesta a esta necesidad, varios algoritmos han sido desarrollados para la minera de subgrafos frecuentes utilizando cotejo aproximado entre grafos para diferentes dominios de la ciencia como: anlisis de estructuras bioqumicas (CHEN et al., 2007; JIA et al., 2011; XIAO et al., 2008; ZHANG and YANG, 2008; ZOU et al., 2009), redes genticas regulatorias (SONG and CHEN, 2006); anlisis de circuitos, redes sociales, y anlisis de vnculos (HOLDER et al., 1992).

Big data y la recuperacin de imgenes web por su contenido. Desde una perspectiva big data, en el mbito web muchas veces es necesario recuperar una imagen de una biblioteca distribuida. En el anlisis de una determinada pgina web son muchos los escenarios que se pueden encontrar: es posible requerir la obtencin de la fuente de las imgenes que aparecen en dicha pgina, tambin es posible analizar si sus imgenes han sido publicadas en terceras pginas, o reconocer a una persona en las fotografas de dicha web. En todos estos casos el problema central es analizar el contenido de las imgenes de inters localizadas en una web para buscar y encontrar una imagen en una biblioteca de imgenes determinadas. Como se puede intuir, cuando se est hablando de grandes cantidades de datos la bsqueda en miles o millones de imgenes ser una bsqueda computacionalmente cara y difcil de ejecutar en equipos informticos bsicos.

La pregunta clave en este contexto es si la tecnologa evoluciona y crece a la misma velocidad que lo estn haciendo estos datos. Para el campo en el que se centra este artculo, el anlisis de imgenes, 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 computacin necesarios) es dificilmente adquirible para llevar a cabo aplicaciones de anlisis de imagen a gran escala, por lo tanto existen pocos investigadores que se aventuren en esta rea. Esto a su vez provoca que el numero de trabajos dedicados

al procesamiento de grandes volumenes de imgenes o videos sea relativamente escaso y monopolizado por grandes grupos de investigacin. 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 petrleo del siglo XXI acuada por

Por otro lado, la minera de subgrafos frecuentes se ha utilizado satisfactoriamente en clasificacin de imgenes (BA-HADIR and SELIM, 2010; JIANG and COENEN, 2008; JIANG et al., 2010b). No obstante, casi todos los enfoques han estado basados en el uso del cotejo exacto. En este trabajo, se propone un esquema basado en grafos para la clasificacin de imgenes. Este esquema utiliza los subgrafos frecuentes como caractersticas obtenidas mediante algoritmos de minera de subgrafos frecuentes en una coleccin de imgenes reales. Utilizando este esquema se evalan los algoritmos mediante la clasificacin de imgenes.

II. TRABAJOS RELACIONADOS

1 Clasificacin de imgenes utilizando minera de subgrafos frecuentes aproximados Tipo de artculo: Artculo original Temtica: Inteligencia artificial, Procesamiento de imgenes, Reconocimiento de patrones

Resumen: Se aborda el anlisis web desde el punto de vista de las imgenes, empleando tecnologas big data. Las imgenes cada vez tienen ms peso en la web por lo que cualquier anlisis que se realice deber considerar este tipo de informacin. Los grandes volmenes de imgenes existentes hacen necesaria la utilizacin de grandes infraestructuras de computacin para realizar este tipo de trabajos, as como tecnologas de visin artificial especficas. Se muestran tecnologas big data que pueden ser utilizadas dentro del campo del anlisis de imgenes a gran escala. Adems, se propone una arquitectura que permite recuperar imgenes de una biblioteca de imgenes de forma eficiente y con un bajo coste computacional. Esta arquitectura puede servir como base para los anlisis web e investigaciones que requieran un estudio detallado de las imgenes similares, sin la necesidad de disponer de hardware especfico para ello.

2 Tecnologas big data para anlisis y recuperacin de imgenes web Rodrguez-Vaamonde, Sergio; Torre-Bastida, Ana-Isabel; Garrote, Estibaliz (2014). Tecnologas big data para anlisis y recuperacin de imgenes web. El profesional de la informacin, v. 23, n. 6, noviembre-diciembre, pp. 567-574.

Resumen: El uso del cotejo aproximado de grafos para la minera de subgrafos frecuentes se ha identificado en diferentes aplicaciones como una necesidad. Con el fin de enfrentar este reto, varios algoritmos de minera han sido desarrollados y han sido utilizados en varios dominios de la ciencia. En este trabajo, se presenta un nuevo esquema para la clasificacin de imgenes basado en grafos donde se utiliza un algoritmo para la minera de subgrafos frecuentes. Esta vez el enfoque del cotejo aproximado de grafos nos permite desarrollar clasificadores robustos ante distorsiones presentes en este tipo de datos. La experimentacin se realiza sobre una coleccin real de imgenes representadas en forma de grafos. Los resultados de esta experimentacin muestran la utilidad del uso de los algoritmos aproximados y del esquema propuesto.

A. Big data y la recuperacin de imagenes web por su contenido

Desde una perspectiva big data, en el mbito web muchas veces es necesario recuperar una imagen de una biblioteca distribuida. En el anlisis de una determinada pgina web son muchos los escenarios que se pueden encontrar: es posible requerir la obtencin de la fuente de las imgenes que aparecen en dicha pgina, tambin es posible analizar si sus imgenes han sido publicadas en terceras pginas, o reconocer a una persona en las fotografas de dicha web.
En todos estos casos el problema central es analizar el contenido de las imgenes de inters localizadas en una web para buscar y encontrar una imagen en una biblioteca de imgenes determinadas. Como se puede intuir, cuando se est hablando de grandes cantidades de datos la bsqueda en miles o millones de imgenes ser una bsqueda computacionalmente cara y difcil de ejecutar en equipos informticos bsicos.
La pregunta clave en este contexto es si la tecnologa evoluciona y crece a la misma velocidad que lo estn haciendo estos datos. Para el campo en el que se centra este artculo, el anlisis de imgenes, 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 computacin necesarios) es difcilmente adquirible para llevar a cabo aplicaciones de anlisis de imagen a gran escala, por lo tanto existen pocos investigadores que se aventuren en esta rea.

Esto a su vez provoca que el nmero de trabajos dedicados al procesamiento de grandes volmenes de imgenes o videos sea relativamente escaso y monopolizado por grandes grupos de investigacin. 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 petrleo del siglo XXI acuada por Andreas Weigend es cierta, el potencial que se pueda extraer de estos datos depender irremediablemente de las tecnologas y algoritmos desarrollados para ello, y en el caso de las imgenes est claro que se necesita potenciar ambos.
Para ello presentamos en primer lugar un resumen de las principales tecnologas big data existentes y su aplicabilidad a los datos en formato imagen. Abarcar el cien por ciento de las tecnologas queda fuera del alcance de este estudio, por lo que nos centraremos en el caso particular de la recuperacin rpida de imgenes en base a su contenido y presentaremos una arquitectura que contiene los ingredientes necesarios para hacer bsquedas rpidas de imgenes que permitan un amplio abanico de anlisis web sin necesidad de una gran inversin en infraestructura.

B. Tecnologias big data aplicables a imagenes

Las necesidades que deben cumplir las tecnologas del big data se basan en el procesamiento eficiente de grandes

cantidades de datos con un tiempo reducido o tolerable.
La complejidad que aade el hecho de que los datos se encuentren en formato imagen es otra variable a considerar. En general las tecnologas de mayor relevancia para datos no estructurados como imgenes son las bases de datos NoSQL (Leavitt, 2010) y los modelos de programacin Map-Reduce (Bajcsy et al., 2013), ambas relacionadas con el procesamiento de datos en lotes. Por otro lado estn los CEP (complex event processing), los IMDG (in-memory data grids), o los sistemas de computacin distribuida, para el procesamiento de datos en tiempo real. En la tabla siguiente se muestra una taxonoma de estas tecnologas en forma de cuadrante ordenado por volumen de datos y tiempos de rendimiento de las tecnologas big data. En los siguientes apartados se detallan las tecnologas ms relevantes en los mbitos anteriores, centrndose en aquellas especialmente tiles para el procesamiento de imgenes a gran escala.

TABLE I
TAXONOMIA DE TECNOLOGIAS DE BIG DATA APLICABLES AL PROCESAMIENTO DE IMAGENES

VOLUMEN DE DATOS	POCOS	% MUCHOS
Tiempo real	Analitica stream	% Analitica tiempo real
	eventos complejos	% Grid datos en memoria
	BD en memoria	% Plataforma especializada
Lotes	Analtica operacin	% Analtica lotes
	OLPT / OLAP	%Map-Reduce
	BD relacional	% BD NoSQL

III. DATOS

TABLE II
DATOS ESTADISTICOS DEL PROYECTO.

MEDIA	MEDIANA	VARIANZA	% DESV.ESTANDAR
25.5	25.5	222.83	14.93

A. Arquitectura software de recuperacin rpida de imgenes similares

Para poder hacer este tipo de anlisis y recuperacin, se propone la utilizacin de tecnologas big data que permitan un acceso eficiente a toda la informacin disponible en imgenes.
La propuesta se puede resumir en la figura siguiente. El primer problema es el almacenamiento eficiente de las imgenes de entrada. Estas imgenes pueden entrar al sistema de diferentes formas, en funcin del anlisis al que se est dedicando esta arquitectura. En el caso de anlisis de imgenes web, la entrada sera por un robot automtico que obtuviese las imgenes de las pginas web.

IV. METDOLOGIA

Dado un conjunto de imgenes representados en forma de grafos pre-etiquetados propuesto por Riesen y Bunke (RIESEN and BUNKE, 2008) se utilizan los algoritmos para la MSFA con el objetivo de obtener todos los subgrafos frecuentes de esta coleccin. Luego, estos subgrafos son utilizados para construir los vectores de caractersticas de las imgenes

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	$d(x,y) = \frac{1}{2}\sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert^2}{x_i + y_i}$

originales. Finalmente, se emplea un clasificador usando estos vectores como datos para realizar la clasificacion de dichas imagenes.

A. Clasificacin de imgenes

Teniendo en cuenta los subgrafos frecuentes obtenidos, se construyen los vectores de caractersticas, entonces una imagen es representada como un vector de caractersticas, donde es el nmero total de subgrafos identificados. Por tanto, se construye una matriz donde el nmero de las filas corresponde al nmero de grafos (imgenes) en la coleccin y el nmero de las columnas corresponde a la cantidad de subgrafos frecuentes (caracteristicas). Cada valor de caracterstica puede ser asignado utilizando una configuracin binaria o una configuracion de similitud. En la configuracin binaria, una celda de la matriz es si la caracterstica ocurre en la imagen de la coleccin y en otro caso . Una celda de la matriz en la configuracin de similitud es el mayor valor de similitud de una ocurrencia de la caracterstica en la imagen. El valor de la similitud de cada caracterstica se obtiene mediante utilizando la funcin de similitud que le corresponda a cada algoritmo en particular.

Para la clasificacin mediante el clasificador SVM (de sus siglas en ingles, Support Vector Machine) se utiliz el paquete libSVM1. En el caso de APGM, se usa la matriz indizada por las etiquetas de los vrtices, las cuales representan el tipo de vertice.

B. Clasificacion de imagenes

Teniendo en cuenta los subgrafos frecuentes obtenidos, se construyen los vectores de caractersticas, entonces una imagen es representada como un vector de caractersticas, donde es el numero total de subgrafos identificados. Por tanto, se construye una matriz donde el nmero de las filas corresponde al nmero de grafos (imgenes) en la coleccin y el nmero de las columnas corresponde a la cantidad de subgrafos frecuentes (caracteristicas).

Cada valor de caracterstica puede ser asignado utilizando una configuracin binaria o una configuracin de similitud. En la configuracin binaria, una celda de la matriz es si la caracterstica ocurre en la imagen de la coleccin y en otro caso . Una celda de la matriz en la configuracin de similitud es el mayor valor de similitud de una ocurrencia de la caracteristica en la imagen. El valor de la similitud de cada caracterstica se obtiene mediante utilizando la funcin de similitud que le corresponda a cada algoritmo en particular.

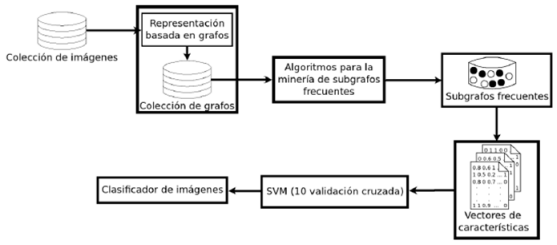

Figura 1. Esquema de clasificación basada en grafos.

C. Esquema para la clasificacion de imagenes

En esta seccion se presenta el esquema propuesto para mostrar la utilidad de los metodos aproximados en tareas de clasificacion de imagenes. Los detalles de la clasificacion de imagenes son introducidos con el esquema propuesto.

de bsqueda de una imagen concreta, se dispondra de una base de datos de imgenes y habra que introducir manualmente la nueva imagen de consulta. En cualquier caso el problema principal consiste en almacenar las imgenes y, adems de la propia imagen, es crucial almacenar sus descriptores visuales. Ya que esta no es una informacin esttica y es dependiente del anlisis o bsqueda a realizar, es necesario disponer de un almacn de datos lo suficientemente flexible. Adems se debe poder guardar una gran coleccin de datos, como son el origen de las imgenes (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 opcin vlida ya que en general permiten disponer de un esquema flexible y son capaces de tener rplicas o nodos distribuidos. Esta ltima caracterstica la hace ideal para el anlisis de imgenes web en cualquier punto geogrfico, ya que permite replicaciones distribuidas en todo el mundo. Sobre este almacn de datos NoSQL es necesario construir el sistema de recuperacin de figuras. No es factible computar para cada imagen de la coleccin la distancia a la imagen de consulta. Por ello es fundamental utilizar algn sistema de bsqueda aproximada de las ms cercanas. El algoritmo ms utilizado en el campo del anlisis de imgenes (Kulis; Grauman, 2009) es locality-sensitive hashing (LSH) (Slaney; Casey, 2008), por lo que ste ser el algoritmo base para la recuperacin de figuras. Este algoritmo permite generar una firma numrica (o hash) para cada descriptor o conjunto de descriptores de imagen, de tal forma que aquellos vectores que tengan una distancia eucldea muy baja, y por tanto sean vectores muy similares, posean la misma firma numrica. Este tipo de algoritmos es muy til para encontrar entradas similares dentro de grandes colecciones de datos, por ejemplobuscando pginas web similares (Slaney; Casey, 2008), por lo que es lgica su aplicacin a los descriptores visuales de imgenes. Una vez se tienen las firmas para todas las imgenes de la coleccin, la bsqueda de las similares es sencilla: dada una imagen de consulta, se generar su hash. Con cada firma se buscar en toda la base de datos las que posean la misma firma y todas ellas sern las imgenes ms similares. Para hacer esta comparacin, se puede pensar en que existe el mismo problema de bsqueda que antes, pero nada ms lejos de la realidad. Las bases de datos NoSQL actuales para big data, disponen de tcnicas de indexacin y bsqueda
rpida de un nmero nico, como puede ser el algoritmo de bsqueda en rbol binario de la base de datos NoSQL MongoDB o el uso de cualquier tecnologa IMDG de las propuestas. Por ello la bsqueda ya no se circunscribe a calcular una distancia entre vectores sino a usar una arquitectura de ndices para encontrar un nmero concreto. Tras este paso ya se dispone de un conjunto de imgenes de la biblioteca similares a la de la entrada. En funcin de la analtica web que se est ejecutando, quiz conocer este nmero es suficiente. En muchos casos la recuperacin de figuras tiene como objetivo aquella que ms se parece a la de entrada. En este caso, es obligatorio calcular la distancia concreta, pero ya que se ha obtenido un conjunto de imgenes parecidas, se puede calcular la distancia sobre ese conjunto de unas decenas de imgenes similares en unos pocos segundos, en vez de sobre el total de la biblioteca. Para este ltimo paso, tambin se va a aprovechar el almacn de datos NoSQL distribuido propuesto en el inicio. Ya que las figuras pueden estar almacenadas en localizaciones diferentes y que cada clculo 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 permitir en la funcin Map el clculo de la distancia entre cada imagen similar y la de consulta, ejecutndose en cada nodo de la red distribuida de almacenamiento. Por otro lado, el mtodo Reduce se encargar de ordenar todas las distancias y podr generar el ranking final de imgenes similares tiles para la analtica.

V. RESULTADOS

Mediante los resultados experimentales, determinaremos la utilidad del uso de la minera de subgrafos frecuentes aproximados en tareas de clasificacion de imagenes. Esperamos realizar un codigo util y eficiente, es decir que trabaje de la manera especificada, reflejando sobre todo los conocimientos adquiridos durante el semestre.

VI. BIBLIOGRAFIA

Ajovalasit, A., Petrucci, G. and Scafidi. M., RGB photoelasticity applied to the analysis of membrane residual stress in glass. Measurement science and technology, vol 23, pp. 1-4, 2012

Azzam, R.M.A. The intertwined history of polarimetry and ellipsometry. Thin Solid Films, Volume 519, pp. 2584- 2588, 2011.

Kasimayan T. and Ramesh, K., Digital reflection photoelasticity using conventional reflection polariscope. Optics and Lasers in Engineering, vol 34, pp. 45-51, 2010

Kasimayan, T., Ramesh, K., Digital reflection photoelasticity using conventional reflection polariscope. ScienceDirect, vol 34, pp. 45-51, 2010.

Wang, Z., Bovik, A. C. H. R. and Sheikh, E., Simoncelli, P., Image quality assessment: from error visibility to structural

similarity. IEEE Transactions on Image Processing, vol 13, pp. pp. 600 612, 2004.

R. Dosselmann, X. and Yang, A., comprehensive assessment of the structural similarity index. Signal, Image and Video Processing, vol 5, pp. 8191, 2011.

Maldonado, M., Sanchez, G. and Branch, J., Registration of range images using a histogram based metric. Dyna, vol 79, (176), pp. 27-34, 2012.