Análisis de los modelos recientes de descripción del contenido de imágenes: desde la Anotación Automática de Imágenes hasta la Recomendación de etiquetas

Juan-Antonio Martínez-Comeche

Resumen: Dado el volumen cada vez mayor de imágenes que manejamos, y su creciente importancia en Internet, resulta imprescindible describir adecuadamente el contenido de estos materiales para garantizar su correcta difusión y recuperación. Dado el volumen de documentos que alcanzan estos repositorios fotográficos, la descripción manual por documentalistas resulta inoperante. En este artículo se describen los enfoques actuales de la investigación relativos a la descripción del contenido de imágenes, desde la Anotación Automática de Imágenes (Automatic Image Annotation) a la Recomendación de Etiquetas (Tag Recommendation). Se analizan los principales problemas que plantean estas técnicas, y las posibles vías de solución que afronta actualmente la investigación.

1. Introducción

En los últimos años se ha producido un incremento espectacular en el número de documentos visuales presentes en las fuentes de información más habituales. En Internet, por ejemplo, han surgido multitud de repositorios de documentos multimedia, muchos de los cuales superan con facilidad los millones de imágenes almacenadas. Sin ánimo de ser exhaustivos, podemos señalar algunos de los ejemplos más conocidos: Google Images (www.google.com) tiene indexadas más de 10 mil millones de imágenes, Flickr (www.flickr.com) supera los seis mil millones de imágenes, YouTube (www.youtube.com) alberga más de 120 millones de vídeos, Shutterstock (www.shutterstock) almacena más de 30 millones de fotos y alrededor de 1 millón de vídeos, fotolia (www.fotolia.com) cuenta con 25 millones de imágenes, Morguefile (www.morguefile.com) dispone de un fondo de 13 millones de imágenes, Dreamstime posee 9 millones de imágenes, y 500px (500px.com) tiene en su colección unos 6 millones de fotos.

Estos números hablan por sí mismos de la importancia creciente de los fondos de carácter visual en bibliotecas, archivos y repositorios digitales, y de la imperiosa necesidad de organizar adecuadamente estos materiales para facilitar su organización, gestión y recuperación. Los documentalistas especializados en este tipo de materiales, conscientes de la importancia de su tarea, han identificado desde hace tiempo las características que deben describirse con la finalidad de garantizar una adecuada representación de los materiales visuales (Valle Gastaminza, 2002): origen de la imagen (autor, momento y lugar de creación...), temática de la imagen (personas, lugares, acciones...), relaciones (con otros documentos visuales o textuales, con personas...) y atributos morfológicos de la imagen (tipo de soporte, formato, color, plano, punto de vista...).

Como no podía ser de otra forma, las ventajas de la normalización en la descripción alcanzan a los recursos audiovisuales en cuanto se percibe su valor documental. Dicha normalización tiene dos principales vías de plasmación: la aplicación de normas de descripción generales a este tipo de materiales (Casado de Otaola; Panizo Santos, 2004) tanto en el ámbito bibliotecario (el formato MARC) como en el ámbito archivístico (la norma ISAD(G)); y, por otra parte, la atención prestada al control terminológico en este ámbito. Algunas instituciones como el Instituto Getty o la Library of Congress han dedicado un notable esfuerzo para el desarrollo de conocidos vocabularios controlados pensados expresamente para la descripción de imágenes: el Art

and Architecture Thesaurus (AAT) y el Thesaurus for Graphic Materials (TGM). Todo ello junto con un notable aumento en los últimos años de manuales dedicados a la gestión de unidades de información que albergan este tipo concreto de documentación (Boadas; Casellas; Suquet, 2001) y, por tanto, interesados especialmente en la descripción y difusión de recursos visuales (Sánchez Vigil, 1999).

Cuando el fondo archivístico o la colección de imágenes es de volumen fijo o al menos de crecimiento moderado, bien porque sea de carácter histórico o bien porque el ritmo de ingreso de nueva documentación se conoce de antemano, y siempre que dicha labor de descripción pueda ser afrontada por la institución en términos de personal, tiempo y costes, cobra pleno sentido la tarea de análisis manual de la documentación conforme a las normas existentes con la ayuda de tesauros y, en general, de vocabularios controlados (Arano, 2005).

En cambio, repositorios digitales de imágenes como Flickr, YouTube o Shutterstock, con un volumen ingente de documentación y un movimiento constante de ingreso, no pueden plantearse la posibilidad de realizar una descripción manual de sus colecciones. Es, pues, la necesidad de describir estos millones de documentos gráficos en permanente aumento de manera rápida y adecuada, a fin de posibilitar la recuperación eficaz de estos recursos, lo que fuerza la consideración de otros modelos distintos a los proporcionados por la tradición documental.

En este artículo se analizan las soluciones que actualmente se proponen en relación al problema del análisis del contenido de imágenes a gran escala, esto es, la descripción de las diversas características de las imágenes mediante palabras clave. Este proceso suele denominarse habitualmente anotación o etiquetado de imágenes.

2. Características del documento visual

Desde un enfoque estrictamente documental, la consideración de materiales audiovisuales presenta dificultades añadidas a las que plantean los recursos textuales, debido a que el código empleado en la transmisión de mensajes es radicalmente distinto. En efecto, la imagen comporta el empleo de atributos como el color, la forma, la textura o las relaciones espaciales (González; Woods, 2008), frente a las palabras como los atributos esencialmente considerados en el procesamiento documental de los documentos textuales.

Este diferente código tiene consecuencias relevantes de cara al quehacer documental. La lengua se caracteriza por partir de un código común a emisores y receptores de mensajes verbales, consecuentemente aceptado y compartido por todos los hablantes. Desde un enfoque semiótico, siguiendo la clasificación de Peirce (Peirce, 1960), ello se expresa diciendo que los signos empleados en las lenguas (las palabras, por ejemplo) son símbolos, y, en consecuencia, representan de manera arbitraria y convencional una realidad exterior. El que una palabra como 'mesa' no mantenga ninguna relación con los objetos reales que representa (como símbolo que es), tiene como contrapartida ventajosa que siempre que se utiliza 'mesa' en la comunicación de un mensaje, el receptor de inmediato puede asociar con precisión a qué objetos alude el mensaje.

La consecuencia de este hecho, relevante de cara a la representación y recuperación de documentos, es que la lengua permite al documentalista saber con gran precisión el contenido del mensaje, al menos a un nivel primigenio (las posibles interpretaciones de un texto quedan fuera de la labor estrictamente documental), pudiendo reflejar con objetividad y corrección dichos contenidos.

En cambio, las imágenes se caracterizan por constituir signos no convencionales, lo que puede implicar dificultades si se pretende representar los mensajes informativos que transmiten. Desde un enfoque semiótico, Peirce incluye las imágenes dentro de la categoría de iconos, pues presentan una relación de semejanza con la realidad exterior designada. Al analizar los iconos,

Peirce pone precisamente como ejemplo los retratos, en cuanto que la imagen (el signo) reproduce de manera semejante la realidad que desea comunicarse (Dubois, 1983, esp. s.v. Icono). Las implicaciones, desde el punto de vista del documentalista que debe representar el contenido del mensaje visual, son importantes y convendría tenerlas en consideración. Si una imagen es un icono, el único mensaje objetivo que transmite -comparable al de un mensaje verbal- es la propia imagen en sí misma (Martínez-Comeche, 1995), al menos considerando el mismo nivel primigenio que hemos destacado al analizar los mensajes verbales (esto es, eliminando las posibles interpretaciones posteriores de los textos). En resumen, la imagen -como icono- transmite, a este nivel primigenio, exclusivamente el mensaje de una cierta composición bidimensional de colores, con determinadas formas y texturas, que mantienen entre sí ciertas relaciones espaciales.

Esta diferente naturaleza semiótica entre un mensaje icónico (imágenes) y un mensaje simbólico (palabras) conlleva, como puede deducirse, desigualdades notables en los niveles de descripción documental que admiten unos y otros mensajes. La desigualdad más llamativa es que las palabras permiten al ser humano expresar conceptos intelectualmente muy complejos -la paz, la solidaridad...-, de manera que un documentalista podrá representar de manera sencilla y fiel dichos conceptos o sentimientos si está describiendo el contenido de un documento textual en el que aparecen, mientras que esos mismos conceptos o sentimientos complejos, en una imagen, no son consustanciales al mensaje original, sino que implican necesariamente una labor suplementaria -ajena a la imagen en sí- de interpretación por parte del receptor de la imagen. Dicho con un ejemplo, una fotografía de un recién nacido en brazos de su madre, que vela su sueño y le mira con cariño, puede fácilmente sugerir el concepto de maternidad e inspirar sentimientos de ternura y amor, e incluso puede que fuese la intención del autor al tomar dicha fotografía, pero en puridad el mensaje icónico transmitido es el de personas (por las formas y distribución de colores) en una posición espacial concreta.

El reconocimiento de este hecho no quiere decir que el documentalista deba renunciar a incluir en las descripciones de los documentos visuales estos conceptos y sentimientos. De hecho, todas las normativas y modelos de descripción de recursos audiovisuales incluyen en mayor o menor medida estos elementos, y todos admiten que el documentalista encargado de la descripción incluya los conceptos abstractos que la imagen sugiera.

El problema no existe cuando se realiza una descripción manual de estos recursos, sino cuando abordamos la descripción del contenido de imágenes a gran escala mediante palabras clave de manera automática. Desde un punto de vista informático es factible inferir ciertos aspectos temáticos en un documento textual a partir de las distribuciones de las palabras en el documento y de las relaciones lingüísticas existentes entre las palabras de una lengua. En cambio, inferir estos mismos aspectos temáticos en un documento visual a partir simplemente de las distribuciones de color, de las formas y de las relaciones espaciales entre dichos atributos en la imagen es una labor muy complicada.

Este fenómeno es conocido, en el ámbito de la Recuperación de imagen, como vacío semántico. Si distinguimos entre las propiedades intrínsecas -o de bajo nivel- de las imágenes digitales (las consustanciales al código visual: color, textura, forma y relaciones espaciales) y las propiedades extrínsecas (todo atributo o cualidad no propiamente visual, pudiéndose discernir entre propiedades de nivel medio -objetos, personas, luz-oscuridad, día-noche, verano-invierno...- y propiedades de nivel alto -autor, título, fecha, formato o temas, entre otros-), el vacío semántico alude a la facilidad con que un sistema automático de recuperación de imagen maneja las características de bajo nivel (pues son los atributos propios del código visual con que trabajan) y la enorme dificultad que encuentran estos sistemas para detectar automáticamente propiedades de nivel medio y de nivel alto en función de los atributos de nivel bajo. Al contrario, los usuarios humanos expresan lingüísticamente con relativa facilidad las cualidades de nivel medio y alto, encontrando dificultades para expresar sus necesidades informativas mediante propiedades de bajo nivel.

El vacío semántico conlleva implicaciones importantes de cara a la búsqueda y recuperación de documentos visuales, pues el sistema automatizado de recuperación visual se verá en la obligación de efectuar algún tipo de equivalencia entre las palabras utilizadas por los usuarios en algunas consultas y las propiedades de bajo nivel (color, forma, textura...) con que ha representado las imágenes digitales de la colección.

En los siguientes epígrafes analizaremos algunas de las vías principales que ha abierto la investigación en el área de la Recuperación de Imagen para tratar de solventar el vacío semántico. Este área de conocimiento se denomina habitualmente Semantic-Based Visual Information Retrieval o Recuperación de Información Visual Basada en la Semántica (SBVIR).

3. La Anotación Automática de Imágenes

Desde el año 2000 aproximadamente (Duygulu; Barnard; Freitas; Forsyth, 2002) se viene investigando intensamente en la manera de establecer una correlación entre las propiedades de bajo nivel y las propiedades de nivel medio -objetos, personas...- o los aspectos semánticos implícitos en toda imagen.

La Automated Image Annotation o Anotación Automática de Imágenes (AIA) trata de equiparar las propiedades de nivel medio y las propiedades semánticas con su traducción correspondiente en propiedades de bajo nivel. El procedimiento seguido para ello consiste en asignar de manera automática palabras clave a las imágenes que representen dichas propiedades. Así, nombres de objetos (casa, coche, mesa, silla...), animales, acciones (correr, andar...), actos (fiesta, cumpleaños...), y denominaciones genéricas de lugares (exterior, interior), estaciones (invierno, verano) o momentos temporales (día, noche...) se añadirían a las imágenes, con la ventaja de poder aplicar directamente los algoritmos de recuperación textual en la recuperación de imágenes.

La asignación automática de palabras clave a las imágenes puede realizarse siguiendo dos procedimientos principales (Shah; Benton; Wu, Raghavan, 2008):

- Basada en Texto: Las imágenes son anotadas mediante el análisis de los textos que acompañan en ocasiones a dichas imágenes.
- Basada en Imagen: Las imágenes son anotadas infiriendo propiedades semánticas a raíz del análisis de las propias imágenes.

La Anotación Automática Basada en Texto ha sido utilizada principalmente en el ámbito de la World Wide Web, pues en la Web abundan las imágenes que se hallan inmersas en un contexto de carácter textual. Los motores de búsqueda como Google Image Search (http://images.google.com), por ejemplo, analizan cualquier texto relacionado con la imagen (el texto alrededor de la imagen en la misma página, el texto de los pies de foto, el texto del nombre del fichero que almacena la imagen, el texto del título de la imagen si lo tiene, el texto del título de la página web donde se incluye la imagen, e incluso el texto de los enlaces exteriores a dichas imágenes) para etiquetar las imágenes presentes en Internet, habiéndose desarrollado diversos modelos de ponderación de esta información textual con el objeto de mejorar los resultados en las búsquedas (Shen; Ooi; Tan, 2000).

La Anotación Automática Basada en Texto también se ha mostrado útil en entornos académicos, donde las imágenes comparten espacio con los textos de artículos científicos especializados. En estos casos, es frecuente el empleo de ontologías o vocabularios controlados específicos del área de conocimiento para evitar la dispersión terminológica habitual en un texto especializado -como el empleo de diversas denominaciones y distintos enunciados del lenguaje natural para un mismo fenómeno- (Hu; Dasmahapatra; Lewis; Shadbolt, 2003).

La Anotación Automática Basada en Imagen, por el contrario, no cuenta con ningún contexto que ayude en el etiquetado de imágenes, sino que las propias imágenes son la base para establecer las palabras clave que describen su contenido.

Los distintos métodos que pueden englobarse bajo esta denominación comparten un procedimiento común, consistente en la utilización de técnicas de aprendizaje máquina, esto es, técnicas que permiten a las máquinas aprender o inferir un cierto conocimiento a raíz de una serie de ejemplos (Duda; Hart; Stork, 2001).

A su vez, las técnicas de aprendizaje máquina empleadas en la Anotación Automática Basada en Imagen pueden subdividirse en (Athanasakos; Stathopoulos; Jose, 2010):

- Técnicas supervisadas de aprendizaje máquina: se caracterizan por partir de un conjunto previo de imágenes etiquetadas a modo de ejemplo.
- Técnicas no supervisadas de aprendizaje máquina: no cuentan con un conjunto previo de imágenes etiquetadas. Estas técnicas se denominan habitualmente de agrupamiento o clustering.

En las técnicas supervisadas de aprendizaje máquina, se debe contar previamente con un conjunto de imágenes anotadas manualmente, denominado conjunto de prueba. A raíz de este conjunto de prueba, y mediante un cierto algoritmo o procedimiento, la máquina infiere una cierta equivalencia entre las propiedades de bajo nivel (color, forma, textura...) de las imágenes de prueba y las palabras clave asignadas previamente. Una vez entrenado el sistema, se etiquetan las imágenes restantes de la colección digital conforme a la equiparación establecida por el sistema entre las propiedades de bajo nivel y las propiedades de alto nivel.

A su vez, dentro de esta aproximación, pueden distinguirse dos enfoques: en el primero de ellos, se segmentan las imágenes en zonas, de manera que en el grupo de imágenes de prueba se dispone de un conjunto de espacios de color y de un listado de palabras clave. La anotación consiste en desarrollar la correlación entre ambos conjuntos.

Uno de los primeros intentos de Anotación Automática Basada en Imagen (Duygulu; Barnard; Freitas, Forsyth, 2002) siguió precisamente este enfoque. El método ideado consistió esencialmente en desarrollar una tabla de equivalencias que cuantificaba la probabilidad de que un cierto espacio hubiese sido etiquetado con una cierta palabra clave.

El segundo enfoque que se ha probado dentro de las técnicas supervisadas de aprendizaje considera las imágenes en su integridad, sin segmentarlas, asumiendo que la anotación automática mejorará si se equiparan imágenes íntegras y palabras clave. Entre los estudios realizados con este enfoque podemos citar los de Feng (Feng; Shi; Chua, 2004). En ellos esencialmente se calcula la distribución de un número predeterminado de colores (en este caso, 12) en cada imagen, originando los denominados histogramas de color, que permiten comprobar los colores predominantes en cada imagen. De igual forma, se analiza la textura de la imagen mediante el análisis de las frecuencias de la imagen. A raíz de estos datos correspondientes a propiedades de bajo nivel, Feng utiliza un algoritmo SVM (Support Vector Machine, en español Máquinas de Vectores de Soporte) para generar el modelo de correlación entre las propiedades de color y textura de las imágenes y las palabras clave asignadas al conjunto de imágenes de prueba.

La principal ventaja del empleo de técnicas no supervisadas es que no requieren de un conjunto de imágenes de prueba anotadas previamente. El algoritmo de clustering elegido, por ejemplo, el de los vecinos más cercanos (Boiman; Shechtman; Irani, 2008), identifica las imágenes más semejantes entre sí a raíz de los datos correspondientes a las propiedades de bajo nivel. Repitiendo reiteradas veces este procedimiento de agrupamiento, se generan tantos grupos (clusters) como se desee a partir de la colección digital de imágenes, grupos que pueden posteriormente ser etiquetados con las mismas palabras clave.

Entre los inconvenientes señalados en relación a las técnicas de aprendizaje supervisado destacan la necesidad de decidir previamente el número de clases que van a distinguirse, el número

de imágenes que es preciso considerar en el conjunto de prueba para que los resultados sean fiables, la obligatoriedad de introducir una variedad suficiente de imágenes en el conjunto de prueba, y la posibilidad de que ciertas imágenes se incluyan en una cierta clase y corrompan la correlación entre palabras clave y propiedades de nivel bajo en una clase.

Entre los inconvenientes señalados en relación a las técnicas de aprendizaje no supervisado destacan la necesidad de decidir previamente el número de clases en algunos algoritmos (como el de K-Medias), que la mayoría de las ocasiones son muy costosos en tiempo de computación y en recursos de memoria, y que los grupos o clusters generados pueden en ocasiones ser difíciles de anotar mediante palabras clave debido a la heterogeneidad de las imágenes que los componen.

Por último, en relación a la Anotación Automática de Imágenes en su globalidad, se han señalado recientemente (Athanasakos; Stathopoulos; Jose, 2010) dos aspectos negativos relativos a la metodología de evaluación seguida en los experimentos relacionados con estas técnicas: por una parte, la imposibilidad de comparar los resultados obtenidos entre distintas aproximaciones por el empleo de colecciones de prueba diferentes; y, por otra parte, la constatación de que en muchas ocasiones los aparentes buenos resultados son debidos a las características peculiares de las colecciones de prueba empleadas y no a la técnica o método particular de afrontar la equiparación entre propiedades de bajo y alto nivel.

Puede concluirse que los resultados obtenidos con estas técnicas de aprendizaje máquina todavía deben mejorarse para ser consideradas plenamente operativas. En definitiva, la Anotación Automática de Imágenes ha constatado que la representación de propiedades de alto nivel mediante atributos perceptuales de bajo nivel es una tarea extremadamente compleja que requiere aún de un gran esfuerzo investigador.

4. Recomendación de Etiquetas

Si la Anotación Automática de Imágenes no han logrado todavía eliminar el vacío semántico, la segunda gran posibilidad de anotación de imágenes a gran escala que permite evitar el vacío semántico, incorporando las propiedades semánticas a las imágenes, es el denominado etiquetado social.

Una de las características esenciales de la Web 2.0 consiste en la participación y colaboración del usuario en la creación de contenidos. El etiquetado social es una de sus expresiones más representativas, pues consiste en permitir que los usuarios añadan palabras clave (etiquetas) a los recursos de Internet (páginas web, imágenes, vídeos) sin basarse en ningún vocabulario controlado. Debido a esta ausencia de estructura terminológica previa, el etiquetado social se basa en comportamientos y estructuras sociales emergentes, así como en las estructuras lingüísticas y conceptuales de la comunidad de usuarios (Marlow; Naaman; Boyd; Davis, 2006). De ahí su estrecha relación con el término folcsonomía, pues el etiquetado social permite vislumbrar, dentro del conjunto de palabras utilizadas sin jerarquía ni relación predeterminada, una taxonomía de los conceptos y palabras más relevantes y emergentes dentro de la comunidad de usuarios (Vander Wal, 2005).

El etiquetado social conlleva, sin duda, beneficios. Entre otros, interesa destacar aquí que permite distribuir la carga de la anotación de imágenes entre todos los usuarios y creadores de dichos recursos, proporcionando una solución al problema analizado en este estudio en cuanto que los usuarios pueden asignar propiedades semánticas con las palabras de su elección.

Sin embargo, el etiquetado social también origina problemas relacionados con las características del vocabulario empleado. Entre ellos, se pueden destacar los siguientes:

- El denominado “Problema del Vocabulario” (Vocabulary Problem), por el que diferentes usuarios utilizan distintas etiquetas/palabras clave para describir las mismas cosas. Este

fenómeno fue analizado por primera vez por (Furnas; Landauer; Gomez; Dumais, 1987) en relación a cualquier sistema de computación. En su artículo, estos autor concluyeron que la probabilidad de que dos usuarios elijan el mismo término para describir el mismo objeto o la misma función es menor del 20%.

- Relacionado con el anterior, se ha comprobado que los usuarios con mayor experiencia en el etiquetado social tienden a emplear etiquetas/palabras clave más específicas y más detalladas que los usuarios noveles. (animal – gato – gato persa – Felis silvestris catus longhair Persian)
- La Polisemia: una misma etiqueta/palabra clave posee significados distintos.
- La Sinonimia: diferentes etiquetas/palabras clave poseen idéntico significado. Estos dos aspectos son conocidos y han sido ampliamente discutidos en el campo de la Documentación (Morato; Sánchez-Cuadrado; Fraga; Moreno-Pelayo, 2008).
- La utilización de diferentes lenguajes por parte de los usuarios.
- El empleo de etiquetas/palabras clave dependientes del contexto. Alude al hecho de que ciertas etiquetas (“yo”, “mis amigos”, “Juan” -y cualquier otro nombre propio-) no pueden ser relacionadas con esas mismas imágenes por otros usuarios porque desconocen el contexto necesario para su desambiguación (¿quién es 'yo'?...)

Estos factores tienen como consecuencia común un empeoramiento de la precisión y de la exhaustividad en la recuperación (Golder; Huberman, 2005). En consecuencia, se han tratado de solventar estos problemas terminológicos. Las soluciones propuestas pueden dividirse en dos diferentes enfoques (Kolbitsch, 2007):

- Búsqueda de similaridades estructurales.
- Búsqueda de similaridades semántico-lingüísticas.

Conforme al primero de los enfoques enunciados, la solución a los problemas terminológicos consiste esencialmente en establecer vínculos o relaciones entre aquellas etiquetas/palabras clave que han sido o pueden ser utilizadas por distintos usuarios para describir los mismos recursos. Con ello no solamente se paliaría el “Problema del Vocabulario”, sino que se reducirían los problemas de sinonimia o del empleo de diferentes lenguas. Estos grupos o clusters de etiquetas/palabras clave se generarían aplicando las mismas técnicas de aprendizaje máquina no supervisado (algoritmos de agrupamiento o clustering) que hemos comentado al analizar la Anotación Automática de Imágenes.

La similaridad estructural no solo permite crear grupos de etiquetas/palabras clave similares. Pensando específicamente en el problema del empleo de un vocabulario más o menos especializado según los usuarios de ciertos recursos, se han aplicado variantes de algoritmos de agrupamiento o clustering jerárquicos (Heymann; Garcia-Molina, 2006) que permiten obtener taxonomías (en concreto, árboles jerárquicos de etiquetas/palabras clave) a partir de las folcsonomías que conforman la totalidad de las etiquetas.

El segundo enfoque busca la información que permita superar los problemas terminológicos en el exterior a la propia colección de etiquetas/palabras clave. Así, el procedimiento seguido consiste esencialmente en aplicar técnicas de modificación de la consulta, en concreto de expansión de la consulta mediante utilización de vocabularios controlados, técnicas muy conocidas en el ámbito de la Recuperación de Información. De esta manera, la consulta introducida por el usuario se amplía con sinónimos, hipónimos (términos más específicos dentro de un campo semántico, como 'descapotable' en relación al término 'coche'), hiperónimos (términos más genéricos dentro de un campo semántico, como 'fruta' en relación al término 'manzana'), holónimos (términos cuyo significado tiene componentes designados por otros términos, como 'casa' en relación al término 'cocina') o merónimos (términos cuyo significado es una parte de un todo designado por otro

término, como 'dedo' en relación al término 'mano'). Todos estos términos añadidos suelen tomarse de vocabularios de carácter general, siendo muy habitual el empleo de WordNet (Kolbitsch, 2007).

Si el etiquetado social permite superar el vacío semántico gracias a la colaboración de los usuarios de los propios repositorios, otra vía complementaria actual de investigación, complementaria de la Anotación Automática de Imágenes, trata de desarrollar sistemas automáticos que ayuden y asistan a los usuarios en la tarea de etiquetar y describir los recursos. Son las denominadas estrategias de Recomendación de Etiquetas.

La investigación en Recomendación de Etiquetas se ha sustentado en dos factores esenciales:

- El contexto en el que la imagen se produce o se incorpora al repositorio digital: se trata de aprovechar la situación espacio-temporal en que se produce o se ingresa la imagen en la colección, así como cualquier otro aspecto informativo complementario que permita sugerir al usuario nuevas etiquetas.
- Las etiquetas previamente asignadas por el usuario a la imagen: se trata de aprovechar la anotación parcial realizada por el usuario para sugerirle nuevas etiquetas que completen la descripción de la imagen.

En relación a la Recomendación de Etiquetas basada en el contexto, su objetivo es sacar partido, por ejemplo, del lugar y momento en que la imagen fue producida, o sacar rendimiento a las coordenadas espacio temporales en que la imagen se incorpora al repositorio y se procede a su anotación. En concreto, en el caso de las fotografías, actualmente no es tan infrecuente que las fotos digitales incorporen los metadatos que la propia cámara ha añadido a la imagen, en especial las coordenadas espaciales (gracias al sistema GPS incorporado) y el momento en que se captó la instantánea. Gracias a estos datos geográficos, el Sistema de Recomendación de Etiquetas puede sugerir etiquetas correspondientes a los países, regiones o accidentes geográficos (montes, ríos, lagos, mares...) donde se tomó la fotografía. De igual forma, basándose en la fecha y hora en que se realizó la foto, pueden fácilmente sugerirse etiquetas relativas a dicho momento (McParlane; Jose, 2013), como la estación del año (verano, invierno...) e incluso etiquetas relativas a los eventos (Rattenbury; Good; Naaman, 2007) a que corresponden las fotos (si el sistema ha grabado lugar y fecha de eventos tales como conciertos, maratones, fiestas, celebraciones...). Asimismo, también se ha investigado la recomendación de etiquetas al usuario basándose en el momento de la incorporación de la imagen al repositorio, dada la relación comprobada entre el instante en que los usuarios ingresan las fotos y el empleo de ciertas etiquetas (McParlane; Moshfeghi; Jose, 2013). Por ejemplo, estos autores muestran que la etiqueta “sunrise” o “sunset” se emplea por los usuarios con mucha mayor frecuencia cuando añaden fotografías en las horas del día correspondientes a la salida y puesta del sol. De igual forma, dentro de este apartado, se ha investigado el empleo del historial de etiquetas empleadas por cada usuario con anterioridad (las tendencias del usuario, en definitiva) para proporcionar sugerencias de nuevas etiquetas a dichos usuarios (Garg; Weber, 2008).

En relación a la Recomendación de Etiquetas basada en las previamente asignadas por el usuario a la imagen en el momento del etiquetado, los procedimientos investigados parten de los datos relativos al empleo simultáneo de todas las etiquetas del repositorio, junto con técnicas de aprendizaje máquina no supervisado (algoritmos de agrupamiento o clustering). El procedimiento consiste, esencialmente, en cuantificar la probabilidad de co-ocurrencia de las etiquetas del repositorio y las empleadas ya por el usuario (Sigurbjörnsson; Zwol, 2008).

5. Conclusiones

El estado actual de la investigación en la Anotación Automática de Imágenes no ha logrado

superar todavía el vacío semántico, esto es, la distancia entre las consultas basadas en palabras clave y las propiedades de bajo nivel (color, forma o textura) con que se representan las imágenes. El etiquetado social, esto es, la posibilidad de disponer de etiquetas asignadas manualmente por los usuarios, es, por tanto, actualmente imprescindible para permitir la recuperación de imágenes en los sistemas de recuperación visual a gran escala existentes. En consecuencia, la Recomendación de Etiquetas se ha convertido en un área de investigación de gran relevancia y de aplicación inmediata en los repositorios comerciales de recursos audiovisuales que cuentan con millones de documentos entre sus colecciones.

BIBLIOGRAFÍA

ARANO, S. (2005). Los tesauros y las ontologías en la Biblioteconomía y la Documentación. Hipertext.net, 2005, nº 3. Disponible en http://www.upf.edu/hipertextnet/numero-3/tesauros.html

ATHANASAKOS, K.; STATHOPOULOS, V.; JOSE, J. (2010). A Framework for evaluating automatic image annotation algorithms. En: Gurrin, C.; He, Y.; Kazai, G. et al. (eds). Advances in Information Retrieval, Proceedings. Berlin: Springer-Verlag, pp. 217-228.

BOADAS, J.; CASELLAS, Ll.E.; SUQUET, M.A. (2001). Manual para la gestión de fondos y colecciones fotográficas. Girona: Centre de Recerca i Difusió de la Imatge.

BOIMAN, O.; SHECHTMAN, E.; IRANI, M. (2008). In defense of nearest-neighbor based image classification. En: IEEE Conference on Computer Vision and Pattern Recognition. Ankorage: IEEE, pp. 1-8.

CASADO DE OTAOLA, L.; PANIZO SANTOS, I. (2004). Descripción de los materiales fotográficos en el fondo de Secretaría del Archivo Histórico Nacional (Madrid). En: Amador Carretero, M.P.; Robledano Arillo, J.; Ruiz Franco, M.R. (coords.). Imagen, cultura y tecnología: Segundas Jornadas. Madrid, pp. 39-54.

DUBOIS et al. (1983). Diccionario de Lingüística. Madrid: Alianza.

DUDA, R.O.; HART, P.E.; STORK, D.G. (2001). Pattern Classification. New York: John Wiley & Sons.

DUYGULU, P.; BARNARD, K.; FREITAS, N.; FORSYTH, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. En: Proceeding of the 7th. European Conference on Computer Vision. London: Springer-Verlag, pp. 97-112.

FENG, H.; SHI, R.; CHUA, T. (2004). A bootstrapping framework for annotating and retrieving WWW images. En Proceedings of the ACM Conference on Multimedia. New York: ACM, pp. 960-967.

FURNAS, G.W.; LANDAUER, T.K.; GOMEZ, L.M.; DUMAIS, S.T. (1987). The vocabulary problem in human-system communication. Communication of the ACM, 30, 11, pp. 964-971.

GARG, N.; WEBER, I. (2008). Personalized, Interactive Tag Recommendation for Flickr. En

Proceedings of the 2008 ACM Conference on Recommender Systems, ACM, pp. 67-74.

GOLDER, S.; HUBERMAN, B.A. (2005). The Structure of Collaborative Tagging Systems. HP Labs technical report. Disponible en http://www.hpl.hp.com/research/idl/papers/tags/ [Consultado: 11/11/2013]

GONZÁLEZ, R.C.; WOODS, R.E. (2008). Digital Image Processing. Third de. New Jersey, Pearson Education.

HEYMANN, P.; GARCIA-MOLINA, H. (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Stanford InfoLab Technical Report 2006-10. Disponible en http://ilpubs.stanford.edu:8090/775/1/2006-10.pdf.

HU, B.; DASMAHAPATRA, S.; LEWIS, P.; SHADBOLT, N. (2003). Ontology-based medical image annotation with description logics. En: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence. Sacramento (California): IEEE, pp. 77-82.

KOLBITSCH, J. (2007). WordFlickr: A Solution to the Vocabulary Problem in Social Tagging Systems. Proceedings of I-MEDIA '07 and I-SEMANTICS '07. Graz (Austria), pp. 77-84.

MARLOW, C.; NAAMAN, M.; BOYD, D.; DAVIS, M. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. En: Proceedings of the 17th Conference on Hypertext and Hypermedia. New York: ACM, pp. 31-39.

MARTÍNEZ-COMECHE, J.A. (1995). Teoría de la información documental y de las instituciones documentales. Madrid: Síntesis.

McPARLANE, P.J.; JOSE, J. (2013). Exploiting Time in Automatic Image Tagging. En Proceedings of the 35th European Conference on Information Retrieval Research, ECIR 2013. Berlin: Springer-Verlag, pp. 520-531.

McPARLANE, P.J.; MOSHFEGHI, Y.; JOSE, J. (2013). On contextual Photo Tag Recommendation. En: Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13). ACM, pp. 965-968.

MORATO, J.; SÁNCHEZ-CUADRADO, S.; FRAGA, A.; MORENO-PELAYO, V. (2008). Hacia una web semántica social. El Profesional de la Información, 17, 1, pp. 78-85.

PEIRCE, Ch.S. (1960). Collected Papers. Cambridge (Mass.): Harvard University Press.

RATTENBURY, T.; GOOD, N.; NAAMAN, M. (2007). Towards automatic extraction of event and place semantics from Flickr tags. En: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07). ACM, pp. 103-110.

SÁNCHEZ VIGIL, J.M. (1999). El universo de la fotografía. Prensa, edición, documentación. Madrid: Espasa.

SHAH, B.; BENTON, R.; WU, Z.; RAGHAVAN, V. (2008). Automatic and Semi-Automatic Techniques for Image Annotation. En: ZHANG, Y. (Ed.). Semantic-Based Visual Information

Retrieval. London: IRM, 2008, pp. 112-134.

SHEN, H.; OOI, B.; TAN, K. (2000). Giving meanings to WWW images. En: Proceedings of the ACM Conference on Multimedia. Marina del Rey (California), pp. 39-47.

SIGURBJÖRNSSON, B.; ZWOL, R. van (2008). Flickr Tag Recommendation based on Collective Knowledge. En: Proceedings of the 17th International Conference on World Wide Web (WWW '08). ACM, pp. 327-336.

VALLE GASTAMINZA, F. del (2002). Perspectivas sobre el tratamiento documental de la fotografía. En: Amador Carretero, M.P; Robledano Arillo, J.; Ruiz Franco, M. R. (coords). Imagen, cultura y tecnología: Primeras Jornadas. Madrid, pp. 165-178.

VANDER WAL, T. (2005). Folksonomy definition and Wikipedia. Disponible en http://www.vanderwal.net/random/entrysel.php?blog=1750 [Consultado: 11/11/2013]