Ciencias de la Información Vol. 46, No. 1, enero-abril, pp. 35 - 40, 2015

Método heurístico para la anotación automática de Imágenes en documentos HTML

Jorge Luis Betancourt González
Adisleydis Rodríguez Alvarez

En el presente artículo se expone una técnica heurística para la anotación automática de imágenes embebidas en documentos HTML, con el objetivo de expandir la búsqueda de imágenes utilizando consultas textuales en un motor de búsqueda Web. El método propuesto aprovecha la estructura de árbol presente en los documentos HTML, tratando de identificar los nodos que pueden aportar información relacionada con la imagen. Para la evaluación de la implementación realizada se utilizó el índice de concordancia para medir el desacuerdo de los jueces voluntarios respecto a la clasificación de un conjunto común de textos asociados a las imágenes; obteniéndose un índice de concordancia superior al 85%.

Palabras clave: anotación de imágenes, HTML, recuperación de información, web

RESUMEN

ABSTRACT

An automatic heuristic method for embedded image annotation in HTML documents is exposed. This method exploits the tree structure present in HTML documents trying to identify nodes that contain relevant information about the embedded image, and then using the text in these nearest nodes to expand the information collected about the image, increasing the recall of a Web Search Engine. The proposed heuristic was evaluated using the Agreement Index: the text contained in the identified nodes and the corresponding image was assessed and assigned a category of how well the text was related (i.e. described) with the image. In our test cases the calculated Agreement Index was over 85%, validating the proposed method.

Keywords: image annotation, HTML, information retrieval, search engine, web

Introducción

En Internet es posible encontrar un cúmulo enorme de información; el verdadero desafío está en descubrir el contenido adecuado para cada usuario: que satisfaga su necesidad de información o aquello que le gustaría leer, ver o escuchar. Los motores de búsqueda ayudan a resolver el problema anteriormente formulado, en especial si la necesidad de información puede ser formulado como una consulta de varias palabras clave(Abhinandan Das, 2007).

Debido a su gran popularidad, la Web está creciendo a ritmo acelerado; más importante aún, computadoras y conexiones de red más rápidas permiten a los creadores de contenidos en la Web mayor libertad para adicionar imágenes, gráficos y videos. Al mismo tiempo el interés de las personas en utilizar imágenes provenientes de la Web ha ido aumentando (Jaimes, et al., 2003).

Los motores de búsqueda permiten incluir algunas palabras clave para buscar, teniendo en cuenta que de forma general el mayor contenido de una página Web es textual, es lógico suponer que éste sea uno de los métodos más extendidos; sin embargo la búsqueda de imágenes es un caso diferente, una imagen aislada no tiene ningún contenido textual que la caracterice; es real que es posible asociarle ciertos metadatos[1], que a su vez pueden proveer información adicional sobre dicho recurso. No obstante, en la mayoría de los casos

[1] *Los metadatos son conjuntos de datos estructurados que describen información, calidad, condición y otras características de los contenidos.*

estos metadatos son insuficientes para determinar los elementos (objetos, temas, etc.) recogidos en la imagen; además, el propio carácter heterogéneo de la Web imposibilita asumir que todas las imágenes poseerán los metadatos necesarios.

En un estudio realizado por (R. Baeza-Yates, 2003) se concluyó que en el año 2001 la palabra clave «fotos» se registró como el segundo término más buscado en TodoCL[2] (motor de búsqueda chileno). Esto pone de manifiesto que desde hace ya más de 10 años los usuarios de los motores de búsqueda han centrado una buena parte de su atención a la búsqueda de imágenes.

Un tema desafiante desde el punto de vista tecnológico lo constituyó caracterizar los contenidos multimedia presentes en la Web. Primeramente se debe lidiar con enormes cantidades de datos distribuidos, además es necesario utilizar herramientas de análisis específicos a cada tipo de medio (imágenes, audio, videos) para determinar los elementos presentes. Específicamente con imágenes y videos, significa desarrollar herramientas que permitan determinar las características visuales: color, textura, forma, etc. Incluso implica el uso de algoritmos para detectar automáticamente los objetos de interés (rostros, símbolos, etc.) presentes (Jaimes, et al., 2003).

La recuperación de imágenes ha sido un área de investigación muy activa desde la década del 70 del pasado siglo, con la ayuda indispensable de dos comunidades de investigación : administración de bases de datos y visión por computadoras. Estas comunidades estudiaron la recuperación de imágenes desde diversos ángulos, centradas fundamentalmente en atributos textuales y visuales respectivamente (Yong Rui, 1999). Incluso en nuestros tiempos, teniendo en cuenta el volumen creciente de información en la Web y la cantidad de imágenes disponibles, puede resultar computacionalmente costoso realizar un procesamiento de los atributos visuales de las imágenes. Sin embargo, es posible explotar el hecho de que la Web es un medio básicamente textual y las imágenes que se adicionan a un sitio Web suelen situarse cerca del texto donde se explica la imagen.

Investigaciones previas

El lenguaje de marcado HTML provee mecanismos semánticos que de ser usados decforma correcta pueden describir la información que aparece en una imagen. Las imágenes, son insertadas utilizando una etiqueta HTML img dicha etiqueta posee un atributo alt que permite precisar un texto alternativo, el cual es mostrado automáticamente cuando el navegador no puede visualizar la imagen. Este texto alternativo puede ser considerado como un buen descriptor, brindando información sobre el contenido de la imagen (…). Sin embargo, en un estudio realizado sólo una pequeña fracción de las imágenes en la colección poseían dichos descriptores; adicionalmente la calidad de los mismos no resulta la mejor debido a que en muchos casos contienen pocas palabras, referidas en su mayoría al nombre del fichero; en correspondencia con lo reportado en (Baeza-Yates, Ruiz-del-Solar, Verschae, Castillo, & Hurtado, 2002).

Los autores (C. Frankel, 1996) proponen un sistema que indexa automáticamente imágenes recolectadas de la WWW[3]. Las imágenes se recuperan de forma automática y clasificadas en categorías basadas en el texto que las rodea. Tambien los atributos visuales son captados para construir un motor de búsqueda que permite realizar consultas por el contenido visual. Por su parte los autores de (J.R. Smith, 1997) elaboraron un sistema similar que además incorpora la detección automática de rostros. El enfoque propuesto combina varias de las ideas encontradas en la literatura y su adaptación a nuestro caso particular, igualmente la investigación se centra en algunas técnicas para la detección del texto en torno a las imágenes embebidas en documentos HTML.

Materiales y métodos

En el presente artículo se usa la terminología expuesta en (R. Baeza-Yates, 2003). Una página se define como un documento indexado por el crawler. Un sitio es un servidor Web lógico identificado por un sub-dominio (por ejemplo, intranet.uci.cu que pertenece al dominio uci.cu). El rastreo de la Web (efectuado por el componente denominado crawler) es aquel proceso en el cual se recolecta de forma rápida y eficiente las páginas Web y la estructura de enlaces que las interconectan, con el objetivo de indexarlas y que sirvan de base de información al proceso de búsqueda (Manning, Raghavan, & Schütze, 2008).

El primer paso de todo motor de búsqueda es el proceso de rastreo, en el cual la Web es recorrida por un crawler que se encarga de extraer la información relevante de las páginas Web y posteriormente indexarlas de forma que esté disponible en la interfaz utilizada para realizar las consultas.

Primeramente se parte de un conjunto de dominios de interés, los cuales constituyen el punto de partida del recorrido. El crawler se encarga de obtener las páginas Web correspondientes a los dominios de partida y extraer todos los enlaces. La información extraída de las páginas Web es almacenada en diferentes cores[4] de Solr. Existe ademas otra instancia del crawler

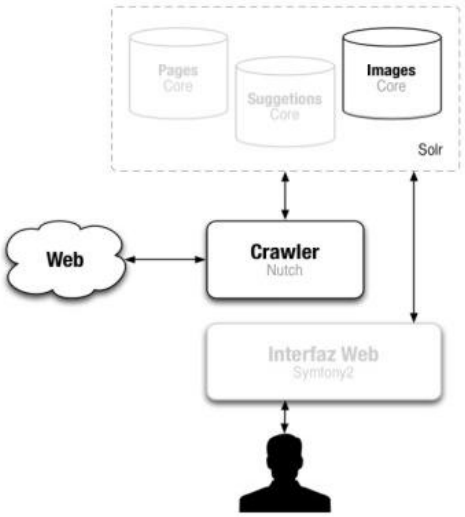

Figura 1: Arquitectura general de un sistema de recuperación de información Web. Se muestran además los componentes utilizados como base para la presente investigación.

[2] *http://www.todocl.cl*
[3] *Sistema de documentos de hipertexto y/o hipermedios enlazados y accesibles a través de Internet.*
[4] *Un «corede solr» no es más que un espacio lógico dentro de una misma instancia de Solr que posee una estructura única, en el caso en cuestión se utilizan diversos* cores*: el «corePages» es utilizado para indexar el contenido textual de las páginas Web y los metadatos asociados; por su parte el* coreImages *almacena los metadatos de las imágenes recolectadas y su versión en miniatura.*
[5] *Base64 es un grupo de esquemas de codificación binaria a texto que representa información binaria en una cadena ASCII.*

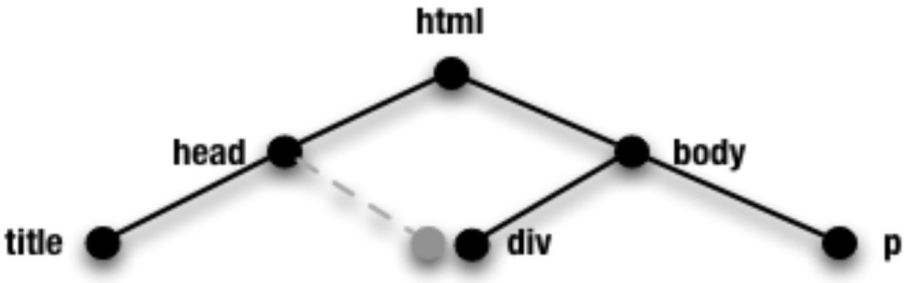

Figura 2: Estructura en árbol de un documento HTML

que se encarga de descargar las imágenes, transformarlas (generar la miniatura que se muestra en la interfaz web), codificarlas a Base64[5] y finalmente almacenarlas en un core destinado a las imágenes. Para cada imagen no solo se almacena su miniatura codificada en Base64, sino que además se extraen ciertos metadatos (dimensiones, URL, dominio, etc.). La Figura 1 muestra los componentes utilizados como base para la presente investigación.

Document Object Model

El DocumentObjectModel o DOM es un estándar independiente de la plataforma y lenguaje de programación para la representación e interacción con objetos en documentos HTML, XHTML y XML (W3C, 2013). De este modo los atributos y el texto se encuentran embebidos en los nodos (Chakrabarti, 2003).

Los nodos por documento se organizan en una estructura de árbol, denominada el árbol DOM (DOM Tree) con el nodo raíz denominado DocumentObject. Esta estructura de árbol (ver Figura 2) permite moverse en cualquier dirección, a través de padres e hijos (verticalmente) y utilizando los hermanos (horizontalmente). Dicha representación permite realizar cambios en el documento a través de ciertos métodos ejecutados sobre un objeto (W3C, 2013).

Nuestro enfoque

Dentro de este marco, en una estructura de árbol se ubican como nodos las etiquetas del lenguaje de marcado HTML, algunas de estas etiquetas son contenedores (por ejemplo las etiquetas div, span, p) o sea, son etiquetas que a su vez pueden contener otras etiquetas, convirtiéndose en la raíz de un subárbol. Las etiquetas img por su parte, siempre se ubicarán como hojas del DOM, pues son elementos que no actúan como contenedores. Teniendo en cuenta la estructura anteriormente expuesta se puede considerar que los nodos «hermanos» de un nodo imagen (Ver Figura 3), es decir aquellos nodos en el mismo nivel (dentro de un umbral de «cercanía»), con información textual, pueden describir o presentar información referente a dicha imagen. Por lo que puede constituir de gran importancia identificar y obtener dicho texto; aumentando, de esta forma, la posibilidad de una coincidencia entre la consulta introducida por un usuario y el texto asociado a la imagen.

Es conveniente destacar que, debido a la propia libertad que proporciona HTML a la hora de crear un documento, la idea anteriormente expuesta puede que no sea certera en todos los casos.

Un árbol puede definirse como un grafo $T = (V, E)$ tal que los elementos de V son los vértices y los elementos de E las aristas; T está mínimamente conectado: es decir, $(T - e)$ proporciona un grafo no conexo para cada arista $e \in T$, además, no posee ciclos y es máximamente acíclico: $(T + xy)$ provee un grafo con ciclos para cualquier par de vértices no adyacentes $x, y \in T$; entonces podemos concluir que es un árbol. En nuestro caso los vértices corresponden a las distintas etiquetas HTML y las aristas representan la relación de jerarquía que se establece entre las distintas etiquetas.

Por consiguiente se define que a un nodo I de tipo imagen, se asocia el conjunto de nodos H que en un mismo nivel se encuentren a una distancia N_h . De igual se forma se asocian el subconjunto V de nodos pertenecientes niveles superiores que se encuentren a una distancia N_v, o sea: $I \leftarrow H \cup V$.

Selección de Nodos vecinos

Teniendo en cuenta lo planteado, surge una interrogante importante: ¿Cuántos elementos (n) se pueden asociar a una imagen? Tal como se muestra en (Chakrabarti, 2003) donde se aplica un enfoque similar, la absorción de atributos de componentes vecinos de forma indiscriminada no mejora la calidad de la clasificación realizada, o sea no aporta datos relevantes al atributo analizado. En nuestro enfoque existe una particularidad: los nodos vecinos son seleccionados en dos direcciones (vertical y horizontal) lo cual permite introducir variables y alterarlas de forma diferenciada. En el caso de la selección vertical es importante destacar que un nivel superior incluye la información de los niveles inferiores del subárbol, lo cual puede generalizar significativamente el texto asociado a una imagen.

Si tomamos como ejemplo el subárbol que se muestra en la Figura 4 podemos observar que incluso la elección de un tope superior para la dirección vertical (N_v) puede tener un gran efecto: Se puede advertir que el subárbol señalado tiene una baja probabilidad de estar relacionado con el nodo imagen. Si establecemos $N_v = 2$ entonces se incluye la información textual de todo el subárbol mostrado; si bien es

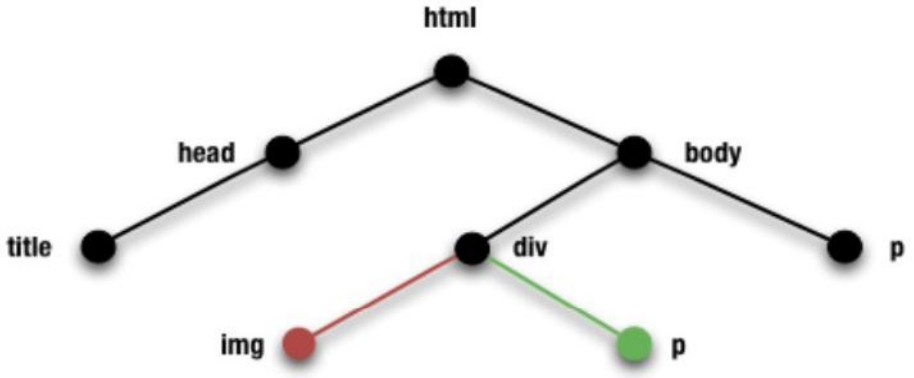

Figura 3: Nodo de tipo párrafo (marcado en verde) cuyo contenido está posiblemente relacionado con la imagen (resaltada en color rojo)

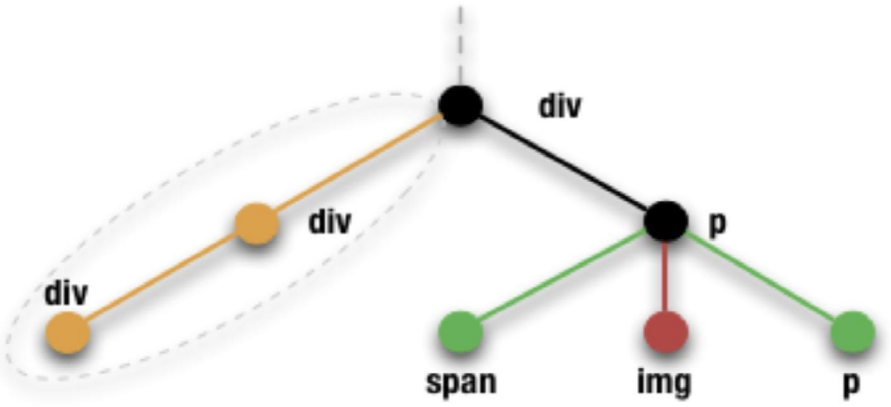

Figura 4: Ejemplo de generalización de la selección vertical de nodos asociados a una imagen

cierto que se incluiría el texto deseado, esto puede causar que una misma imagen coincida con criterios muy diversos y no relacionados entre sí, disminuyendo la precisión del sistema y aumentando la exhaustividad, lo cual puede ser perjudicial para un sistema de recuperación de información.

Por su parte la cantidad de nodos del mismo nivel (N_h) que serán asociados a la imagen (dirección horizontal) se encuentra sujeto a un problema similar; aunque es necesario resaltar que la diversidad introducida es menor en comparación con un similar desplazamiento en dirección vertical, pues como se ha demostrado al aumentar en k, la cantidad de nodos verticales a considerar se adiciona la información correspondiente a S subárboles de modo que:

$$S = \sum_{i=0}^{k} d(n_k),$$

donde $d(n_k)$ denota el grado o valencia del nodo n. A los propósitos del presente artículo se toman los valores $N_h = 2$ y $N_v = 2$; en la Figura 5 aparecen señalados los nodos que serán asociados a la imagen.

Resultados y discusión

La evaluación del sistema implementado posee de por sí una serie de retos. En primer lugar no hay una colección controlada que pueda ser utilizada para probar la precisión y exhaustividad del sistema, haciéndose necesario utilizar algunas métricas alternativas que permitan evaluar cómo se comportaba la heurística utilizada. Para ello se decidió utilizar una métrica de discrepancia, aplicada sobre un sistema de evaluación que permitirá evaluar los puntos de vistas de un conjunto de jueces que evaluaron que tan bien se relacionaba el texto recolectado con la imagen en cuestión. De modo que el proceso de evaluación se centró entorno a tratar de responder la siguiente pregunta: ¿existe un consenso general acerca de cuan bien describe el texto recolectado a la imagen asociada? Parece razonable asumir que la respuesta a esta interrogante en algún sentido debería guiar el proceso de validación de la solución propuesta; aún así, la solución a este problema no es obvia.

Estadígrafo de Kappa

Una primera opción para cuantificar el nivel de concordancia entre jueces es utilizar el estadígrafo Kappa (Cohen, 1960), una medida estadística de fiabilidad entre los evaluadores:

$$k = \frac{P - Pe}{1 - Pe} \qquad (1)$$

que es definida como la diferencia entre cuánto consenso está realmente presente $(P - P_e)$ comparado con el valor cohesión esperado de manera aleatoria $(1 - P_e)$. P es el valor de consenso entre los jueces y P_e es la probabilidad de coincidencia aleatoria. En particular se utiliza el estadígrafo Kappa de Fleiss (Green, 1997; Fleiss, 1971), una variante a la propuesta de Cohen que es aplicable a un número constante de evaluadores que clasifiquen una cantidad fija de elementos. Para la interpretación del valor que se obtuvo se utilizó la escala que se muestra en la Tabla 1, similar a la presentada por (Landis & Koch, 1977).

Tabla 1. Interpretación del valor de kappa

Kappa	Consenso entre jueces
< 0	Ninguna posibilidad de acuerdo
0.01-0.20	Ligero
0.21-0.40	Considerable
0.41-0.60	Moderado
0.61-0.80	Sustancial
0.81-0.99	Casi perfecto

Considerando las tres clasificaciones para evaluar (bueno, regular, malo) se observó un valor de kappa de 0.64 que puede ser interpretado como un consenso sustancial en la decisión de los evaluadores. En los comentarios recibidos de los voluntarios se observó la dificultad para asignar la categoría «regular». Es por ello que si restringimos la evaluación de kappa a sólo al subconjunto de imágenes y texto que fueron calificadas de bueno o malo se obtiene un consenso mucho mayor entre los jueces, con un valor 0.82, casi perfecto teniendo en cuenta la escala que se utiliza. Los resultados de la métrica utilizada reflejan que existe un buen consenso en los jueces respecto a la pertenencia de cada par (imagen y texto asociado) a la categoría definida, resultando que cerca del 90% de las imágenes que se recolectaron (y el texto asociado) recibieron una evaluación de bien o regular.

Índice de concordancia: Otra forma de evaluar el desacuerdo entre los voluntarios, es considerando una «matriz de costo».

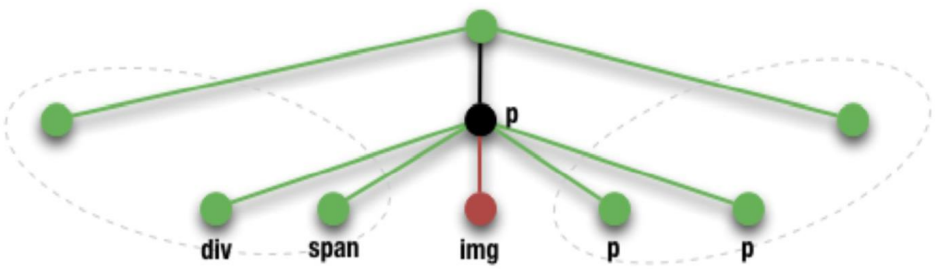

Figura 4: Ejemplo de generalización de la selección vertical de nodos asociados a una imagen

matriz es simétrica y cuadrada, sus índices para cada fila y columna son las posibles etiquetas (bueno, malo y regular), cada coordenada (a, b) corresponde al costo de sustituir la clasificación a por la clasificación b. Obviamente los elementos de la diagonal principal de esta matriz es cero. Para el experimento realizado se consideró la siguiente matriz de costo:

Tabla 1. Matriz de costo para el índice de concordancia

	Bien	Regular	Mal
Bien	0	0.5	1
Regular	0.5	0	0.5
Mal	1	0.5	0

El significado de esta matriz de costo indica que, dado dos jueces, si uno de ellos considera que el texto asociado no describe el contenido de la imagen (o sea una clasificación de mal) y el otro juez piensa lo contrario, entonces existe un mayor nivel de discrepancia, que si por ejemplo considerara que es regular el nivel de descripción del texto para con la imagen.

Dentro de este marco, dados dos revisores (o jueces) **i** y **l**, se define su nivel de concordancia respecto a las imágenes $S_i \cap S_l$ de la siguiente forma: para cada $j \in S_i \cap S_j$, la concordancia **Aj (i, l)** de y en es:

$$A_j(i,l) = 1 - cost(a,b) \quad (2)$$

$$AI(i,l) = \sum_{j \in S_i \cap S_j} A_j(i,l) / |S_i \cap S_j| = \sum_{j \in S_i \cap S_j} A_j(i,l) / O(i,l) \quad (3)$$

Constituyendo el parámetro $O(i, l)$, el número de imágenes clasificadas común a cada par de jueces. Se consideraron los valores correspondientes a la media, los valores máximo y mínimo del índice de concordancia para valores crecientes del número de imágenes en común entre los revisores **i** y **l**. En particular para cada valor x de solapamiento se restringió a todos aquellos pares (i, l) de revisores de modo que $O(i, l) \geq x$, tomando el máximo, mínimo y la media respectivamente. En la Figura 6. Índice de concordancia en función del número de imágenes solapadas, para valores $x \leq 70$ (se seleccionaron 5 pares de jueces con dicha cantidad de imágenes solapadas) se puede observar una gráfica de la evolución de estas métricas.

Conclusiones

El análisis de las métricas recolectadas durante la realización del experimento demostraron que la mayoría de los pares de revisores tenían poco solapamiento, o sea que no evaluaron las mismas imágenes; esto sugiere que puede realizarse este experimento a una mayor escala con el objetivo de cubrir una mayor cantidad de las imágenes recolectadas, de modo que el solapamiento entre cada par de jueces sea superior. El análisis posterior también demostró que a pesar de lo anteriormente enunciado, muchos pares de revisores solaparon sus clasificaciones de forma significativa, al menos lo suficiente como para arribar a algunas conclusiones de acuerdo a su índice de concordancia, probando que coincidían en la evaluación asignada.

La media del índice de concordancia nunca es superior al 89% y nunca menor al 73%. Además, el índice de concordancia no parece aumentar con el número de solapamientos. De hecho, para los valores mayores de 50 imágenes solapadas el índice comienza a decrecer, cuando el número de pares de revisores sobre los cuales la media es calculada son aún relativamente bajos (entre 7 y 10). Es probable que esta resultado deba ser corroborado en posteriores estudios de mayor magnitud, pero parece indicar que un índice no despreciable de desacuerdo entre los revisores puede no ser el resultado de ruido estadístico, sino a la naturaleza ambigua respecto a las categorías usadas; o sea ambigüedad en la decisión de si el texto seleccionado describe o no de forma satisfactoria la imagen asociada, pudiendo solucionarse con la selección de categorías más disjuntas.

En sentido general el elevado porciento del índice de concordancia demuestra que la heurística seleccionada eleva la precisión del sistema de búsqueda de imágenes sin introducir demasiada ambigüedad. El método expuesto en el presente artículo, al centrarse en los atributos textuales para la anotación de las imágenes puede ser extensible a otros formatos presentes en la Web. Además, no limita o interfiere con la aplicación de métodos de procesamiento de imágenes para la identificación de objetos y otras técnicas, constituyendo una alternativa más barata en términos de procesamiento a estos métodos.

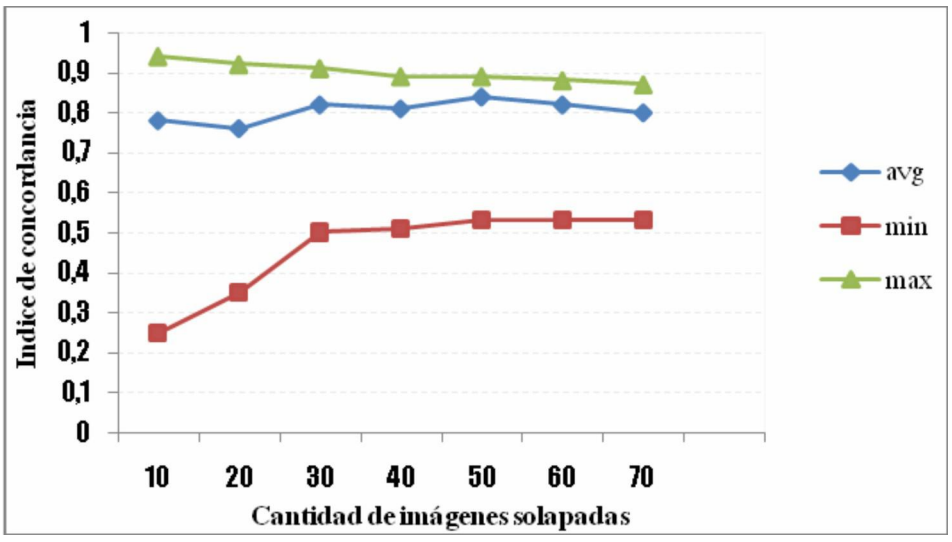

Figura 6: Índice de concordancia en función del número de imágenes solapadas

Bibliografía

R. Baeza-Yates, B. J.-J. (2003). Evolucioìn de la Web Chilena 2001-2002 (Evolution of the Chilean Web 2001 - 2002).Center for Web Research, Department. of Computer Science, Universidad de Chile.

W3C. (17 de 08 de 2013). Document Object Model (DOM). Recuperado el 17 de 08 de 2013, de W3C: http://www.w3.org

Yong Rui, T. S.-F. (1999). Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of Visual Communication and Image Representation , 39-62.

Abhinandan Das, M. D. (2007). Google News Personalization: Scalable Online Collaborative Filtering. WWW 2007 / Track: Industrial Practice and Experience .

Baeza-Yates, R., Ruiz-del-Solar, J., Verschae, R., Castillo, C., & Hurtado, C. (2002). Content-based Image Retrieval and Characterization on Specific Web Collections. Center for Web Research, Department of Computer Science, Universidad de Chile.

C. Frankel, M. S. (1996). «WebSeer: An Image Search Engine for the World Wide Web. University of Chicago Technical Report TR-96-14.

Chakrabarti, S. (2003). Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Psychological Bulletin, 20, 37-46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. Proceedings of the Twenty-Second Annual Conference of SAS Users Group, San Diego, USA .

J.R. Smith, S.-F. C. (1997). An Image and Video Search Engine for the World-Wide Web. Proc. of SPIE Storage & Retrieval for Image and Video Databases V, 3022, 84-95.

Jaimes, A., Ruiz-del-Solar, J., Verschae, R., Yaksic, D., Baeza-Yates, R., Davis, E., y otros. (2003). On the Image Content of the Chilean Web. Proceedings of the First Latin American Web Congress (LA-WEB 2003) .

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). An Introduction to Information Retrieval. Cambridge University Press.

Recibido: 11 de agosto de 2014.
Aprobado en su forma definitiva:
9 de febrero de 2015

Jorge Luis Betancourt González
Centro de Ideoinformática, Universidad de las Ciencias Informáticas (UCI), Boyeros, La Habana, Cuba
Correo-e.: jlbetancourt@uci.cu.

Adisleydis Rodríguez Alvarez
CALISOFT,
La Habana, Cuba.
Correo-e.: yisselec@yahoo.com