

Uso de inteligencia colectiva en el desarrollo de una aplicación Web 2.0 empleando técnicas de búsqueda basada en contenidos para la recuperación de mamografías digitalizadas

Yuliana Pérez-Gallardo, Giner Alor-Hernández, Ruben Posada-Gomez, Guillermo Cortes-Robles

División de estudios de Posgrado e Investigación,
Instituto Tecnológico de Orizaba

Av Oriente 9 No. 852, 94320, Orizaba, Ver.,

México. Teléfono/Fax: +52(272) 725 7056

{gallardo.yuli}@gmail.com, {galor, rposada, gcortes}@itorizaba.edu.mx

Paper received on 02/07/10, Accepted on 21/09/10.

Resumen. En la actualidad los motores de búsqueda tradicionales tales como Google han facilitado enormemente la recuperación de imágenes sobre la Web. Sin embargo, los resultados obtenidos por éstos no son a menudo útiles para los usuarios, por lo general debido a dos problemas: (1) no es posible definir una consulta acerca del contenido de una imagen, y (2) La semántica de las palabras clave no se considera, existiendo problemas de polisemia y sinonimia. Este trabajo presenta una arquitectura basada en capas con el fin de desarrollar una aplicación Web 2.0 que use técnicas de búsqueda basada en contenidos para la recuperación de imágenes a través de Inteligencia Colectiva. Como caso de estudio, los usuarios realizan búsquedas de imágenes médicas (mamografías digitalizadas) intentando eliminar los problemas semánticos y de contenido.

Palabras Clave: CBIR, Web 2.0, Inteligencia Colectiva, Mamografías.

1. Introducción

Actualmente los motores de búsqueda tradicionales facilitan de gran manera la recuperación de imágenes sobre la Web, ejemplos de ello son Google y Bing. Sin embargo, los resultados recuperados por éstos motores no son a menudo útiles para los usuarios, comúnmente debido a dos problemas: en primer lugar 1) la semántica de las palabras clave no se toma en consideración, cuando el usuario proporciona a

un motor de búsqueda tradicional, una palabra clave ambigua, la respuesta consiste en todos los elementos relacionados con los diversos significados de la palabra, mientras que el usuario está probablemente interesado en sólo uno de ellos. Esta ambigüedad se genera por la polisemia, que es la propiedad que poseen las palabras que, en distintos contextos tienen varios significados diferentes [69], y la sinonimia, que es la relación semántica de identidad o semejanza de significados entre determinadas palabras (llamadas sinónimos) u oraciones [2]. En segundo lugar, 2) no es posible definir una consulta acerca del contenido de una imagen, se necesita que los resultados de recuperación provean las imágenes que una persona seleccionaría, al buscarlas por sí mismo en un repositorio donde se almacenan. Esta situación presenta un problema, ya que es difícil de codificar en un algoritmo de manera automática, la idea que tiene una persona sobre el significado semántico de una imagen. Las personas tienden a describir las imágenes basándose en los objetos que están representados en ellas, este tipo de descripción produce resultados de correspondencia entre la percepción humana y el contenido de la imagen.

Basado en la problemática anterior, este trabajo propone una aplicación Web 2.0 que use técnicas de Inteligencia Colectiva para la búsqueda de imágenes basadas en contenido. Para comprender mejor la importancia de la propuesta, se describe en la sección 2 los trabajos relacionados, los principales conceptos que aborda este trabajo en la sección 3, la arquitectura del sistema y el caso de estudio en las secciones 4 y 5 respectivamente y finalmente se presenta en las secciones 6 y 7 el trabajo a futuro y conclusiones.

2. Trabajos relacionados

Dada la importancia que tiene la Inteligencia Colectiva en la actualidad, existen múltiples investigaciones dentro de los diferentes campos en los que se utiliza la misma. La recopilación y análisis de artículos relacionados con la propuesta permite obtener información relevante y adquirir conocimiento acerca de técnicas de desarrollo y enfoques en los que en la actualidad se trabaja. En [3] se propone el diseño de nuevos modelos de negocio dentro del comercio electrónico, para los integradores de servicios utilizando los principios de Inteligencia de Enjambre (*Swarm Intelligence* en inglés) para proveer medios que aprovechen la Inteligencia Colectiva de los clientes, de esta forma se desarrollan nuevos productos y servicios basándose en las preferencias de sus clientes maximizando las posibilidades de éxito, tal como actualmente lo hacen empresas como Adidas, Lego Factory y Amazon. En [4] presenta el uso de IC para la colaboración intercultural. Esta investigación se enfoca en el obstáculo que todavía representa el lenguaje en la comprensión mutua entre sociedades. En los sistemas de traducción automática la traducción no es perfecta, debido al desajuste que existe entre los significados de las oraciones y/o palabras dentro de las comunidades. Proponen la producción colaborativa de materiales de traducción personalizadas, tomando en cuenta a la propia comunidad nativa para que colabore en el desarrollo del material de traducción de su lengua a otras. Dentro del enfoque de CBIR, en [5] proponen que mediante la agrupación de documentos etiquetados en una folksonomía, es posible extraer los conjuntos de etiquetas relacionadas con los diferentes contextos en donde existen etiquetas ambiguas. Con su propuesta intentan

solucionar dicha ambigüedad construyendo clasificadores que usan datos recolectados desde sistemas de etiquetado colaborativo. Como pruebas, realizaron experimentos en búsquedas de Google, en donde los resultados muestran que el método es capaz de clasificar los documentos devueltos con alta precisión. Utilizan el algoritmo de k-vecino más cercano para clasificar los documentos devueltos por el motor de búsqueda. En [6] se enfocan en la extracción de elementos de contenido de bajo nivel. Se extraen las características que representan el contenido de las imágenes en vectores y para medir su similitud utilizan la distancia euclidiana. Dentro de este mismo enfoque proponen una prueba de correlación en la distancia convencional que corresponde al mecanismo de recuperación de imágenes basado en contenido. En [7] se propone un sistema de recuperación de imágenes basada en una búsqueda combinada de textos y contenidos. La idea es utilizar el texto presente en el título, descripción y etiquetas de las imágenes para la mejora de los resultados obtenidos de la búsqueda. La búsqueda basada en texto proporciona resultados con similitud semántica, mientras que la búsqueda basada en contenido proporciona resultados con similitud gráfica. Debido a la independencia entre estos enfoques, su combinación mejora el rendimiento de un sistema de búsqueda por el beneficio de ambos enfoques. Finalmente, en [8] se presenta un sistema de CBIR basado en componentes configurables. La principal novedad reside en su contenido basado en búsqueda de imágenes de componentes (CBISC) que apoya las consultas sobre las colecciones de imágenes. CBISC se basa en la Iniciativa de Archivos Abiertos (OAI), permitiendo así consultas planteadas a través de peticiones HTTP y las respuestas codificadas en XML.

3. Web 2.0, Inteligencia Colectiva y CBIR

La Web ha experimentado una evolución a través del tiempo [9] para cubrir las necesidades de los usuarios, en sus orígenes con la Web 1.0 con páginas HTML estáticas, pasando por la Web 1.5 hasta ahora con la Web 2.0, en donde las aplicaciones Web se centran en los usuarios invitándolos a participar activamente interactuando, mostrando su punto de vista y preferencias, de tal forma que convierten a los usuarios en colaboradores, generando en contenido que ellos mismos consumen. Así, la Web 2.0 son todas aquellas utilidades y servicios de Internet que se sustentan en una base de datos, la cual es modificada por los usuarios del servicio, ya sea en su contenido (añadiendo, asociando, cambiando o borrando información), en la forma de presentarlos, o en ambos [9]. Pero, surge una pregunta ¿Qué hacer con todos los datos recolectados por la participación e interacción de los usuarios? Esta información se pierde si no se convierte en inteligencia y se canaliza hacia la mejora de la aplicación.

Lévy [10] sugiere que la Inteligencia Colectiva (IC) es una especie de sociedad anónima en donde cada accionario aporta como capital su conocimiento, sus conversaciones, su capacidad de aprender y enseñar. La suma de inteligencias no se somete ni se limita a las inteligencias individuales, sino por el contrario, las exalta, las hace

fructificar y les abre nuevas potencias, creando una especie de cerebro compartido. Debido a los beneficios que nos brinda la suma de las inteligencias individuales, éstas son aplicadas a la Recuperación de Imágenes Basada en Contenido (CBIR, en inglés Content Based Image Retrieval) para tener un mejor resultado en las búsquedas. CBIR es un sistema que recupera imágenes basadas en sus propiedades visuales. El objetivo de las técnicas de consulta basadas en contenido de las imágenes es encontrar, de forma eficiente, las imágenes en una base de datos que son similares a un criterio de búsqueda. A diferencia de las típicas consultas en bases de datos, en este caso se utiliza un criterio de similitud que no tiene por qué ofrecer una coincidencia exacta. Uno de los elementos más importantes en CBIR es la extracción y la evaluación de la medida de similitud de las características de la imagen. Las características de la imagen son aquellos elementos que describen el contenido de una imagen. El significado de las características depende del contexto en el que se manejen, existen características visuales que se derivan directamente de la imagen, el ejemplo más simple es el valor de intensidad de los píxeles de la imagen, aunque este tipo de características visuales son muy sensibles al ruido, brillo, tono y saturación de los cambios. También hay otras características que son específicas de los dominios de aplicación y requieren cierto conocimiento especial. Una buena característica contiene suficiente poder discriminante para distinguir entre las imágenes similares y diferentes y es invariante a la transformación espacial, como la traslación, rotación y cambios menores relacionados con el medio ambiente de iluminación, donde se captura la imagen.

4. Arquitectura del sistema

Para esta propuesta, se propone el desarrollo de una aplicación Web 2.0 que use técnicas de IC para la recuperación de mamografías digitalizadas basadas en su contenido. En la Figura 1, se presenta la arquitectura basada en capas de la propuesta y se describe cada uno de los componentes que la integran. En esta arquitectura, cada capa tiene una función específica que a continuación se describe:

1. **Capa del Cliente:** Está integrada por los módulos a) Cargar imagen, b) Retroalimentación, c) Solicitud de búsqueda, d) Criterio de búsqueda, e) Sugerencia Semántica y f) Resultado de la consulta. Esta capa brinda la interacción con el usuario con el fin de permitirle acceder a los servicios que ofrece la aplicación.
2. **Capa de Control de Imágenes:** Contiene los módulos a) Análisis de la imagen y b) Controlador de la imagen. Esta capa se refiere a las diferentes operaciones que el sistema lleva a cabo con respecto al análisis de las imágenes, por ejemplo la binarización de las imágenes, obtención de regiones de interés, almacenamiento de imágenes y características, entre otras.
3. **Capa de Etiquetado:** Consta de los módulos a) Etiquetado, b) Asignación del vector de características y c) Actualización de características de la imagen. En esta capa se lleva el control del etiquetado colectivo de las imágenes.
4. **Capa de Recuperación de Imágenes:** Integrada por los módulos a) Análisis de características, b) Algoritmo de similitud, c) Búsqueda de imágenes

con características similares, d) Análisis semántico y e) Imágenes. Esta capa contiene todos los módulos necesarios para la recuperación de imágenes basados en contenido y su análisis semántico.

5. **Capa de Datos:** En esta capa se maneja la persistencia tanto de los datos como de las imágenes.

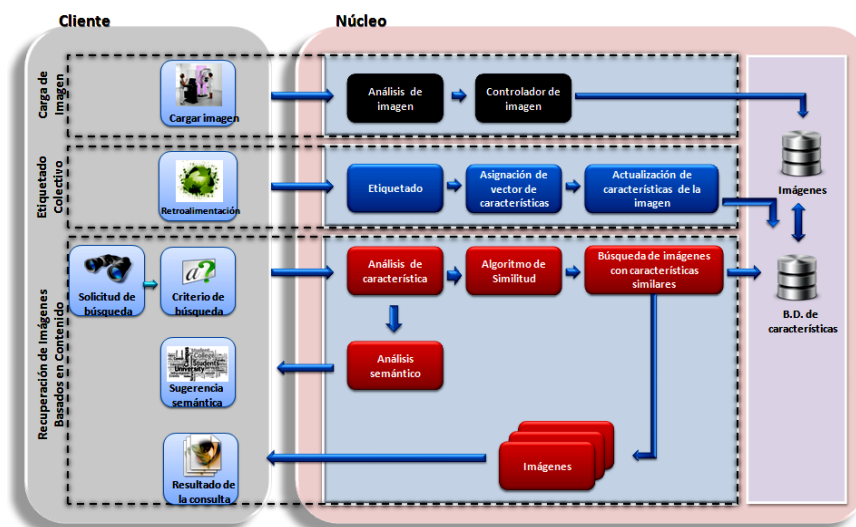


Figura 1. Elementos internos de la arquitectura.

Las capas interactúan entre sí para brindar el funcionamiento total del trabajo propuesto, por medio de diferentes flujos de trabajo. Los flujos de trabajo que contiene el sistema son: carga de imagen, etiquetado colectivo y recuperación de imágenes basados en contenido.

4.1. Flujos de trabajo

A continuación se describe la funcionalidad de los flujos de trabajo que conforman la propuesta.

Carga de imagen. Este flujo de trabajo tiene por objetivo la gestión de las genes. Está integrado por los módulos análisis de la imagen y controlador de la gen. Esta propuesta será empleada para la recuperación de mamografías sadas en contenido. Las imágenes que se emplean son de tipo ICS, que es el Estándar para Imágenes de Citometría (Image Cytometry Standard, en inglés).

formato ICS se propuso por [11] en 1990 y es capaz de almacenar 1) Imágenes bidimensionales y multicanal, 2) Imágenes en 8, 16 o 32 bits enteros, 32 o 64 flotante y datos punto flotante complejos y 3) Parámetros tomados de un mamógrafo acerca de la formación de la imagen. El formato original ICS archiva los datos en archivos independientes: un archivo de encabezado de texto con extensión .ics, otro con los datos de la imagen real, con la extensión .ids. El flujo de trabajo comienza cuando el usuario dentro del módulo Cargar imagen, carga una imagen con formato ICS y la envía al servidor al módulo de Análisis de la imagen en donde, se muestra la

Figura 2. Detección de características, se obtienen las características, por ejemplo micro calcificaciones o tumores; esto se lleva a cabo aplicando el método de Otsu al separar dentro de la imagen el objeto de interés y el fondo, obteniendo así una imagen binarizada, después se aplica un filtro de erosión y otro de dilatación para tener una mejor definición de los contornos de los objetos, por último se analiza cada píxel y sus vecindades para definirlos o no como áreas de interés.

Una vez obtenidas se envían al módulo de Controlador de imagen, que se encarga de almacenarlas en la base de datos de imágenes y características respectivamente. Para manipular las imágenes con formato ICS, se utiliza la librería ImageJ, que es una biblioteca de funciones para el procesamiento de imágenes basado en Java y desarrollado en el Instituto Nacional de Salud de EE.UU. (NIH, National Institutes of Health, en inglés). ImageJ es una plataforma ideal para desarrollar, probar nuevos algoritmos y técnicas de procesamiento de imágenes, también se utiliza en la investigación y desarrollo de aplicaciones en muchos laboratorios del mundo, especialmente en imágenes biológicas y médicas.

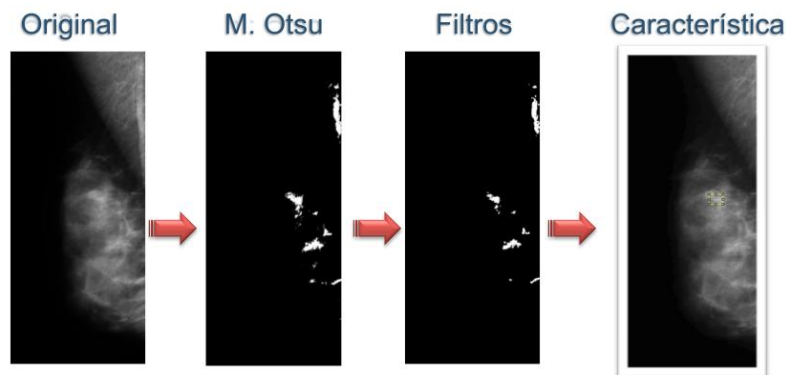


Figura 2. Detección de características

Etiquetado colectivo. Este flujo empieza en el módulo de Retroalimentación en la capa del cliente, la retroalimentación también denominada feedback es el proceso

de compartir observaciones y sugerencias, con la intención de recabar información, a nivel individual o colectivo, para mejorar el funcionamiento de un proceso. Dentro del contexto de la propuesta, se utiliza la retroalimentación del usuario al agregar o modificar etiquetas a las imágenes con el fin de describir su contenido.

El objetivo de este flujo de trabajo es realizar un sistema de etiquetado colaborativo, dentro del módulo de Etiquetado y por medio de una interfaz se muestran las mamografías digitalizadas, en donde se permite a los usuarios seleccionar características dentro de la imagen y asociarlas a etiquetas con significado semántico con el fin de describirlas; resultando en una clasificación generada por los usuarios que se conoce comúnmente como folksonomía. El etiquetado colaborativo permite que distintos usuarios describan las imágenes sobre las cuales tienen conocimiento, generando con cada contribución un grado mayor de inteligencia para toda la comunidad. Una vez que el usuario selecciona las características de la imagen, éstas siguen al módulo llamado Asignación de vector de características, en donde se asocian las características obtenidas a la imagen correspondiente para después dentro del módulo de Actualización de características de la imagen, almacenarlas o actualizarlas, según sea el caso, en la base de datos correspondiente.

Recuperación de imágenes basado en contenido. El objetivo de este flujo de trabajo es recuperar un conjunto de imágenes basándose en el contenido de éstas, es decir, en sus características. El inicio de este flujo se define cuando el usuario dentro del módulo de Solicitud de búsqueda en la capa del cliente, solicita una nueva búsqueda, después dentro del módulo de Criterio de búsqueda, el usuario proporciona un criterio, que para una nueva búsqueda, es una palabra clave (mientras la búsqueda se refine éste criterio crecerá en más ítems), la cual envía al módulo de Análisis de características donde se reparten a dos módulos más el módulo de Análisis semántico y el de Algoritmo de similitud. En el Análisis semántico se calcula la semántica de la palabra clave, es decir si la palabra tiene algún problema de polisemia y/o sinonimia, éstos se resuelven y se sugieren los posibles contextos en que el usuario la utilice para delimitar el sentido semántico. Dicha sugerencia se hace al usuario dentro de la aplicación por medio de una nube de etiquetas, para guiar el filtrado de la búsqueda dependiendo el contexto. Mientras que, en el módulo de Algoritmo de similitud, se lleva a cabo la comparación por distancia entre la palabra clave y las etiquetas de características almacenadas en la base de datos con el fin de obtener y mostrar en el módulo de Resultado de consulta aquellas imágenes con contenido similar al criterio de búsqueda propuesto por el usuario.

5. Caso de estudio

Dentro del campo de la medicina existe una gran cantidad de información visual como radiografías, mamografías, resonancias magnéticas, tomografías y muchas más disponibles para análisis médicos. La producción de imágenes crece más rápido

que los métodos para administrar y procesar esa dicha información, imponiéndose un nuevo reto para su eficiente representación, almacenamiento y recuperación.

Para exponer el funcionamiento de la propuesta, se lleva a cabo un caso de estudio dentro del contexto de la medicina, para la recuperación de mamografías digitalizadas con formato ICS, como lo muestra la Figura 3. **Mamografía original digitalizada**

Supóngase que un usuario ha percibido una calcificación en dicha mamografía pero para asegurar su diagnóstico, quiere compararlo con otras mamografías con el mismo diagnóstico. En un caso normal, el usuario busca expedientes de algunas pacientes diagnosticadas de la misma forma, y tomar sus mamografías, con el inconveniente que los expedientes no estén disponibles o que no se trate de una calcificación similar.

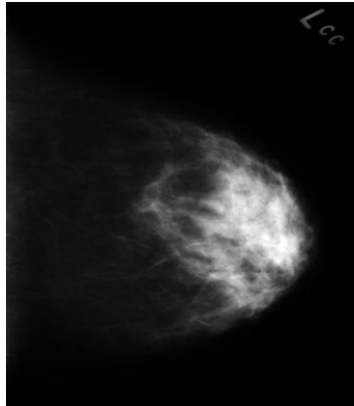


Figura 3. Mamografía original digitalizada

Una alternativa de solución al problema anterior es utilizar la aplicación propuesta en este artículo. Para este caso de estudio se discutirá el modo en que se realiza la asignación de característica y la búsqueda de imágenes similares.

Asignación de características. El usuario carga la imagen en la aplicación, la cual se envía al servidor, en donde por medio de un proceso de binarización de la imagen y del análisis de los píxeles, se realiza una sugerencia de los sectores en donde posiblemente se encuentran calcificaciones y/o tumores, tal y como lo muestra la Figura 4. **Proceso de binarización y obtención de áreas de interés.** Permitiendo de esta forma, que el experto corrobore la información modificando la selección y/o reemplazándola como crea conveniente; y agregando la etiqueta correspondiente que describa la característica seleccionada.

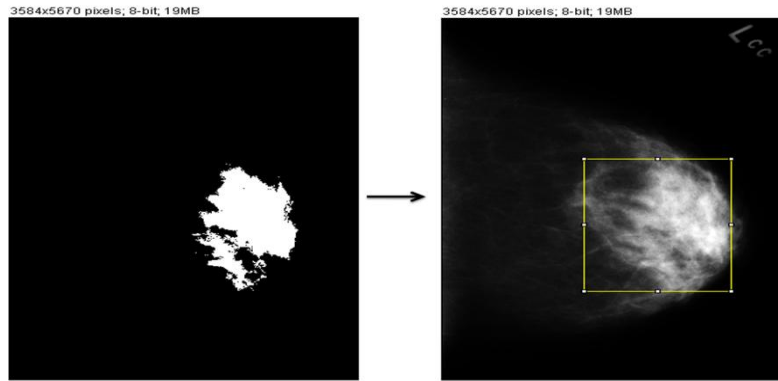


Figura 4. Proceso de binarización y obtención de áreas de interés.

Más adelante otros usuarios si así lo desean, modifican, agregan o eliminan las diferentes características que presenten las imágenes, de acuerdo con su conocimiento para beneficio de la comunidad. La Figura 5. **Etiquetado de características** muestra el proceso de etiquetado, en donde existen opciones de exponer diferentes aspectos de la imagen, en este ejemplo se muestra su contorno. Así las consultas posteriores serán cada vez más exactas y completas debido a la suma de conocimiento de los usuarios participantes en el etiquetado.

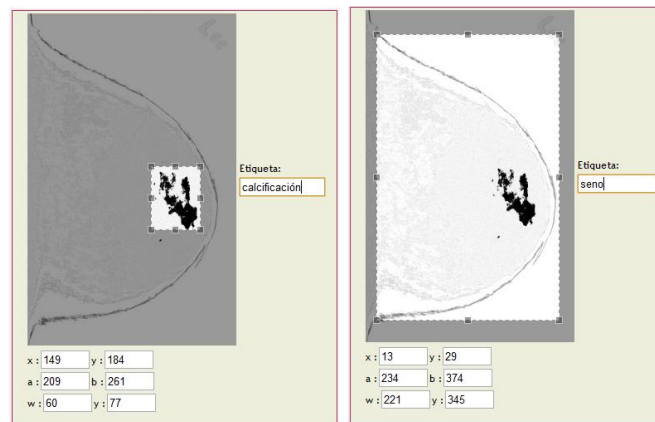


Figura 5. Etiquetado de características

Búsqueda de imágenes similares. Para este caso si el usuario necesita consultar todas aquellas imágenes que contengan *mastitis* como característica, el usuario pro-

porciona la palabra clave en la aplicación, para que, a partir del cálculo de la distancia euclidiana se obtenga la distancia entre la palabra clave y las etiquetas almacenadas en la base de datos con el fin de obtener aquellas imágenes con contenido similar a la palabra clave propuesta por el usuario. Además la aplicación sugiere, a partir de la palabra clave, los posibles contextos en que el usuario la utilice para delimitar el sentido semántico.

Este caso de estudio muestra la utilidad de la aplicación propuesta como una alternativa para la búsqueda de imágenes médicas basadas en contenido, utilizando Inteligencia Colectiva dentro de la aplicación Web 2.0. Es útil para usuarios que deseen recuperar imágenes con significado semántico y visual similar, basándose en el conocimiento y experiencia de otros usuarios.

6. Trabajo futuro

Como trabajo futuro está el desarrollo de la interfaz para el flujo de trabajo Recuperación de imágenes basado en contenido y sus correspondientes módulos para agregarlos a la aplicación Web, de tal forma que se solucione los posibles problemas de semántica para las palabras clave y además se aplique el algoritmo de distancia entre la palabra clave y las características almacenadas en la base de datos, para obtener las imágenes con contenido similar. Así también, la implementación de un sistema de reputación para los usuarios que participan en el etiquetado de las imágenes, de tal forma que sea posible conocer que tan confiables son sus opiniones para etiquetar las imágenes. Dicha reputación se adquiere de las calificaciones de una encuesta realizada a otros usuarios, quienes realizan las búsquedas y expresan su opinión acerca de la certeza de dichas etiquetas.

7. Conclusiones

El factor determinante en el interés por las técnicas CBIR es el rápido incremento del tamaño de colecciones de imágenes digitales, así como en la importancia y disponibilidad de las imágenes en diversos campos aplicación como la medicina. Por lo que existe una necesidad de gestionar datos de imagen digital, dado que la digitalización en sí no facilita la gestión, aunque posibilita derivar automáticamente información de las imágenes en sí mismas. Esta propuesta intenta resolver problemas semánticos y de contenido en la realización de búsquedas, específicamente de mamografías digitalizadas, mediante el uso de Inteligencia Colectiva y al implementar un algoritmo CBIR, permitiendo que los usuarios sean las personas encargadas de administrar y conformar la información existente dentro de la aplicación, obteniendo así un mayor grado de conocimiento para la comunidad. Los beneficios que presenta la propuesta dentro del contexto de la medicina en donde los diagnósticos se hacen frecuentemente diferenciales, son gran ayuda debido al uso de la inteligencia colectiva de los especialistas.

Agradecimientos

Este trabajo fue apoyado por la Dirección General de Educación Superior Tecnológica de México (DGEST). Además, este trabajo fue patrocinado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) y la Secretaría de Educación Pública a través de PROMEP.

Referencias

1. Polisemia. Enciclopedia Microsoft Encarta Online 2009 <http://mx.encarta.msn.com>. 1997-2009. Microsoft Corporation. Reservados todos los derechos.[Consulta: sábado, 31 de noviembre de 2009]
2. Revista Española de Lingüística. "Sinonimia y teoría semántica en diccionarios de sinónimos de los siglos XVII y XIX". <http://www.uned.es/sel/pdf/ene-jun-94/24-1-Gonzalez.pdf>. 1994. [Consulta: miércoles, 23 de diciembre de 2009]
3. Ulrike, B., Sandro, G., Henrik, I., Reinhard, J.: Design of new business models for service integrators by creating information-driven value webs based on customers collective intelligence. In: Proceedings of the 42nd Hawaii International Conference on System Sciences. 2009.
4. Ishida T.: Service-Oriented Collective Intelligence for Intercultural Collaboration. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2008.
5. Ching-man, Yeung A., Gibbins, N., Shadbolt, N.: A k-Nearest-Neighbour Method for Classifying Web Search Results with Data in Folksonomies. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2008.
6. Jian, C., Ming, Y.: A New Similarity Measurement Based on Distance and Correlation Test for Content-Based Images Retrieval. In: Congress on Image and Signal Processing. IEEE. 2008.
7. Barrios, J.M., Díaz-Espinoza, D., and Bustos, B.: Text-Based and Content-Based Image Retrieval on Flickr: DEMO. In: Second International Workshop on Similarity Search and Applications. IEEE. 2009.
8. Da S-Torres, R., Bauzer-Medeiros, C., Goncalves M. A., Fox, E.: An OAI Compliant Content-Based Image Search Component. In: Join ACM/IEEE Conference on Digital Libraries. ACM. 2004.
9. Ribes Xavier. "La Web 2.0. El valor de los metadatos y de la inteligencia colectiva". <http://www.telos.es/articuloperspectiva.asp?idarticulo=2&rev=73>. 2008. [Consulta: jueves, 03 de septiembre de 2009]
10. Cobo Romani Cristobal, Pardo Kuklinski Hugo. "Planeta Web 2.0. Inteligencia colectiva o medios fast food". Grup de Recerca d'Interaccions Digitals, Universitat de Vic. Flacso México. Barcelona / México DF. 2007.
11. Dean, P., Mascio, L., Ow, D., Sudar, D., Mullikin, J.: Propsed standard for image cytometry data files, Cytometry. In: Wiley-Liss, Inc. n.11, pp. 561-569, 1990.