SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Y RECOMENDACIÓN

MÁSTER EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

Sistemas de Recuperación y Ranking de Imágenes basados en Consultas Multi-Atributos

---

Autora
Lidia Sánchez Mérida

Curso 2021 - 2022

Granada, Mayo de 2022

Introducción a los SRI

Un **Sistema de Recuperación de Información** (SRI) es un conjunto de componentes interactivos que realizan labores individuales con un mismo objetivo: facilitar el almacenamiento y recuperación de información bajo demanda. Los datos participantes pueden ser de muy diversa naturaleza, desde información textual como multimedia, considerando imágenes, audio y video. En este trabajo se pretende tratar el caso particular de los **SRI basados en imágenes**. No obstante, independientemente del tipo de información involucrada, la mayoría de SRI persiguen los siguientes objetivos [1]:

1. La primera tarea a efectuar reside en el **análisis y representación** del conjunto de datos que se pretende almacenar en el sistema. Si la naturaleza de la información es visual, existen diversas metodologías que extraen las características de las imágenes para transformarlas en valores más sencillos de procesar con el fin de simplificar las consultas.
2. Uno de los elementos comunes a todos los tipos de SRI son los **índices,** consistentes en estructuras de datos que agilizan las búsquedas de contenido mediante representaciones organizadas de los datos.
3. Presumiblemente un SRI puede disponer de **multitud de motores de búsqueda** con los que llevar a cabo diferentes patrones de consultas con sus respectivas métricas para determinar la relevancia de cada resultado. Generalmente, los sistemas orientados a imágenes suelen tener la capacidad de realizar recuperaciones mediante una serie de imágenes de referencia o un conjunto de términos alfanuméricos [1].

Origen y motivación

La necesidad de almacenar información comienza a incrementar durante la conocida época de la Revolución Industrial en la que se documentaron y publicaron un amplio abanico de inventos de distinta índole. En combinación con la aparición de los ordenadores, varias personalidades de la era comenzaron a plantear mecanismos para guardar y recuperar volúmenes considerables de datos. Ejemplos representativos fueron los de Vannevar Bush, que en 1945 ingenió una versión preliminar de un **motor de búsqueda** o H. P. Luhn que en 1957 propuso un **criterio de consulta** basado en el número de ocurrencias de los términos proporcionados por un usuario [2]. Sin embargo, los avances más relevantes no surgieron hasta la década de los sesenta, cuando la comunidad científica comenzó a prestar interés a los SRI con el objetivo de almacenar ingentes cantidades de **bibliografía** relativa a diferentes ámbitos de investigación. Los primeros sistemas implementados para este propósito organizaban

la información como documentos y únicamente respondían a **consultas *booleanas*** basadas en determinar la existencia de un texto mediante una lista de términos proporcionados por el usuario [1]. Durante las siguientes décadas continuaron los progresos en el desarrollo de SRI implementando diferentes enfoques y diseñando diversas evaluaciones para comprobar la utilidad y eficiencia frente a distintos conjuntos de datos. Esta última característica fue especialmente perseguida en 1992 durante el evento ***Text Retrieval Conference*** (TREC) en el que participaron multitud de agencias gubernamentales americanas con el cometido de comprobar el rendimiento de los SRI sobre conjuntos masivos de datos. Gracias a la celebración de este congreso se optimizaron técnicas existentes y se propusieron otras nuevas con varios fines, como la aplicación de los SRI a datos de distinta naturaleza, la inclusión de filtros de información y la ampliación de estos sistemas a multitud de idiomas [2].

Como consecuencia del auge de los SRI, también comenzaron a aplicarse en recuperación de imágenes pertenecientes a distintos ámbitos, como la medicina o la economía. Uno de los primeros enfoques que surgió en la década de los setenta tiene como objetivo **etiquetar las imágenes empleando sus metadatos** para utilizar esta información intrínseca con la que realizar las consultas de recuperación. Para ello se usaron diversas técnicas de preprocesamiento y análisis de textos que ayudaron a impulsar y publicitar la investigación y desarrollo de esta tecnología [3].

A continuación fueron propuestas nuevas mejoras, como una particular en la que se combina la generación de **términos lingüísticos** manualmente en función de una métrica conocida como ***Content Based Image Retrieval*** (CBIR), que persigue la extracción automática de propiedades visuales basadas en la percepción analítica del ser humano. Las tres características principales son el color, cuya extracción se caracteriza por ser bastante simple, aunque no ocurre lo mismo con la textura y la forma de los objetos o seres representados en las imágenes, pese a encontrarse definido un conjunto de estrategias para su identificación basadas en reglas estadísticas, geométricas y matemáticas [4] [5]. Sin embargo, se presentan dos principales inconvenientes en esta primera aproximación: la excesiva **cantidad de recursos** que se deben invertir en la generación de etiquetas por parte de expertos humanos y la **falta de objetividad** intrínseca a esta actividad.

Con el objetivo de solventar las dificultades mencionadas anteriormente, comenzaron a centrarse los esfuerzos en la búsqueda de algoritmos automáticos que fuesen capaces de aprender a realizar las mismas tareas minimizando los costes temporales, económicos y personales.

Técnicas tradicionales

Una de las primeras técnicas empleadas para la recuperación y *ranking* de imágenes se apoya en el **modelado individual** de un conjunto de términos situados en un dominio concreto para su uso en consultas, en conjunción con funciones heurísticas que combinen los diferentes resultados para obtener los ejemplos coincidentes ordenándolos según el valor resultante. Una situación representativa de esta metodología podría consistir en realizar una búsqueda acerca de *perros de pelaje blanco,* con la que se activaría el entrenamiento de tres clasificadores, cada uno especializado en un término, con el fin de unir sus conocimientos y encontrar ejemplares caninos del color especificado ordenando los resultados de mayor a menor precisión. Sin embargo, este procedimiento conlleva ciertos inconvenientes asociados como los **recursos** requeridos para la construcción de tantos modelos como vocablos componen una consulta, cómo encontrar la **combinación de los clasificadores** más óptima para lograr el objetivo común o no considerar las **dependencias existentes entre las palabras** que, generalmente, pueden aportar información útil a ambas operaciones de recuperación y *ranking* [6] [7].

Posteriormente se diseñó un segundo enfoque bastante exitoso conocido como ***Visual Reranking*** dividido en dos principales etapas. El cometido de la primera reside en la recuperación de imágenes basada únicamente en **etiquetas lingüísticas** relativas al contenido, como los metadatos o términos situados en la fuente de los datos. Mientras que la segunda fase aplica un **filtrado** de las imágenes obtenidas usando un clasificador especializado en características visuales ordenando los resultados en función del grado de coincidencia con los términos buscados. No obstante, al igual que en la técnica anterior, la dificultad más importante es la de construir la lista de conceptos requerida para llevar a la práctica la primera etapa [7].

Con el propósito de resolver el obstáculo generalizado que afecta a las metodologías anteriores se planteó un algoritmo capaz de **producir descripciones textuales** a partir de imágenes y vídeos. Tal y como sucedía en el método previo, la primera tarea se fundamenta en el reconocimiento de los objetos y seres participantes en la escena de una imagen incluyendo las conexiones entre sí que componen su **contexto**. A partir de esta cualidad, el algoritmo calcula un vecindario para cada imagen determinando el grado de similitud mediante la distancia existente entre sus respectivos contextos. En la segunda etapa se hace uso de una variante de la métrica conocida como ***Canonical Contextual Distance*** (CCD) que permite estimar el nivel de semejanza entre el contenido de una imagen y un conjunto de etiquetas lingüísticas. El objetivo que pretende conseguir consiste en asignar el mayor número de conceptos explicativos

acerca de la escena de una imagen, seleccionando aquellos con una similitud más elevada. Finalmente, con la integración de diversas técnicas de reconstrucción de frases se combinan las distintas palabras generadas para componer una breve y concisa descripción del significado de una imagen [7] [8].

Evolución de los sistemas

Tras demostrar la elevada relevancia que adquieren los SRI basados en imágenes y las limitaciones asociadas a las técnicas tradicionales anteriormente descritas, comienzan a surgir nuevos enfoques que combinan los últimos avances de investigación relativos a diferentes áreas como **Minería de Textos y Procesamiento del Lenguaje Natural**. Se plantea un doble objetivo para mejorar la usabilidad de los sistemas de recuperación, por un lado se trata de maximizar su **adaptación al lenguaje natural humano** de modo que se simplifique su uso y se reduzca la curva de aprendizaje para aumentar el número de usuarios finales. Mientras que por otro lado se emplean una mayor cantidad de procedimientos altamente complejos para extraer más conocimiento útil mediante la realización de análisis exhaustivos del conjunto de imágenes disponible.

Múltiples párrafos en consultas

Uno de los enfoques que consigue ampliar la aplicación de consultas multi atributos en SRI persigue el objetivo de recuperar conjuntos de imágenes relacionados con textos de mayor volumen como **párrafos**. La inspiración de este trabajo, tanto para su planteamiento como para su resolución, reside en la idea de permitir que usuarios, procedentes de diferentes aplicaciones, puedan visualizar las fotografías más representativas de las opiniones publicadas. Como datos de entrada disponen de un primer conjunto de *posts* que permite la construcción de un modelo predictivo capaz de **relacionar la lista de imágenes contenidas en las publicaciones con sus comentarios** asociados. En la fase de validación se emplean dos conjuntos adicionales, siendo una colección de fotografías secuenciales tomadas por el mismo usuario en la misma fecha, y una serie de comentarios redactados por otros usuarios de la aplicación.

Con el fin de determinar el grado de asociación entre imágenes y comentarios, en primer lugar se emplea una técnica conocida como **segmentación de textos**, con la que dividir la entidad completa en diferentes grupos de frases que facilite la conexión con una única imagen del conjunto. Para implementar este procedimiento se han combinado diversos métodos pertenecientes al área denominada **Procesamiento del**

Lenguaje Natural (NLP), como un ***tokenizador*** basado en expresiones regulares con el que identificar las distintas temáticas o eventos por los que puede ser dividido un párrafo. Adicionalmente, mediante un **análisis semántico latente** (LSA) se procede a obtener el listado de términos más representativos de cada una de las frases generadas en la etapa anterior. Mientras que con el algoritmo ***LexRank*** se recopilan los conceptos claves de cada documento generando un grafo cuyos nodos representan las sentencias procedentes del tokenizador, y cuyos enlaces interconectan aquellas más parecidas en función de la similitud del coseno basado en la métrica **TF-IDF**. Su cometido consiste en extraer las características de los documentos con las que elegir el listado de términos más representativos del significado de cada texto. Unificando los resultados proporcionados por estas dos técnicas es posible llevar a cabo un filtrado del contenido de los documentos con el fin de seleccionar una única sentencia que **maximice la representación semántica** para cada párrafo disponible. Tras construir un *corpus* de comentarios preprocesados, a continuación se procede al empleo de ciertas técnicas de tratamiento de textos, como la normalización de términos, eliminación de palabras de parada o *stopwords*, además de la generación de bolsas de palabras que simplifiquen las estructuras de datos que contienen los términos de los documentos.

La tercera etapa consiste en realizar un procedimiento homólogo al anterior fundamentado en la descripción de textos, aunque utilizando un conjunto de imágenes. Para ello se colocan **rejillas divisorias** sobre cada una de las fotografías con el cometido de extraer propiedades e histogramas visuales de las porciones resultantes. Como el volumen de información es considerablemente elevado, mediante el uso de **algoritmos de *clustering*** se seleccionan un subconjunto de los atributos identificados en cada región con los que a continuación definir el descriptor global de cada imagen.

Finalizando esta sección detallada acerca de la formulación del problema que presenta este trabajo, se establece el procedimiento de incrustación de texto en imágenes mediante dos enfoques principales. El primero se encuentra fundamentado en el ***Normalized Canonical Correlation Analysis*** que pretende obtener información común a partir de dos entidades representadas matricialmente, mientras que por otro lado se emplea el algoritmo de los **K vecinos más cercanos** (KNN) almacenando un listado de parejas compuestas por imágenes y textos con las que vincular futuras muestras de sendas entidades [9].

Consultas estructuradas y grafos de escenas

El pilar fundamental del SRI más novedoso recopilado en este trabajo se fundamenta en el empleo de **consultas estructuradas**. Se trata de estructuras de datos adaptadas al

almacenamiento de las relaciones existentes entre un conjunto de objetos o seres vivos que participan en la escena de una imagen. Su formulación reside en el esquema tradicional utilizado en análisis sintácticos de oraciones, siendo sus tres componentes principales el **sujeto, predicado y objeto**. Para generar esta información se hace referencia al uso de **grafos de escenas** que contienen las propiedades gráficas relativas al contenido de una imagen. Los nodos representan las entidades protagonistas que aparecen en la fotografía que pueden atribuirse los roles de sujetos u objetos, mientras que los enlaces reflejan las interconexiones presentes entre ellas jugando el papel de predicados. Este enfoque presenta ciertas ventajas sobre las metodologías detalladas en secciones previas, ya que es capaz de adaptarse por completo al lenguaje natural utilizado por los humanos, a la vez que representa las características visuales más destacables de una imagen con las que generar una descripción altamente precisa [10].

Puesto que los datos de entrada se encuentran formateados como grafos no es posible aplicar algoritmos de Aprendizaje Automático o Deep Learning convencionales por su **falta de estructura** y su incapacidad de aplicar métricas de distancias tradicionales, como la Euclídea, para llevar a cabo el proceso de aprendizaje. Gracias a los avances realizados en el área de procesamiento y clasificación de imágenes se define una nueva arquitectura denominada ***Graph Convolutional Neural Network*** (GCNN), como una variante generalizada de la red *Convolutional Neural Network* (CNN). Para representar la información entrante emplea una **matriz de características** con la que almacenar las propiedades de cada nodo como filas, además de un **vector *one-hot*** que contiene las etiquetas relativas a las muestras del conjunto de entrenamiento. Su composición se apoya principalmente en una de las redes más básicas y eficientes conocidas hasta el momento como es el **Perceptrón Multicapa**, por lo tanto como consecuencia directa podemos determinar que un modelo GCNN dispone de un conjunto de capas ocultas y emplea un mecanismo de propagación de información hacia delante aunque presenta las siguientes diferencias:

- Los valores de los nodos no se corresponden con los originales del grafo de entrada, si no que sufren un proceso de **suavizado** calculando la media aritmética con aquellos vértices que pertenecen a sus respectivos vecindarios locales. Así se reducen los cambios bruscos que puedan existir entre los valores de los nodos más cercanos, a la par que se promueve la predicción de una misma clase entre vértices localmente conectados.
- Si bien las capas ocultas de una arquitectura GCNN también llevan asociados tanto una matriz de pesos como transformaciones lineales para propagar el conocimiento hacia capas posteriores, la **función de activación** empleada presenta una formulación **no lineal** como es el caso de la ReLU.

Finalmente en la última capa se utiliza una función *softmax* con la que calcular la probabilidad de pertenencia de cada nodo para las distintas clases contempladas en el problema de clasificación, al igual que sucede en la arquitectura original de las redes convolutivas.

Tras la explicación de las cualidades esenciales y el desempeño de las etapas integradas para la construcción de clasificadores con GCNN, a continuación procedemos a introducir los detalles del propósito de su uso para el aprendizaje de patrones situados en grafos de escenas. En la Figura 10 se muestra el flujo de cada uno de los procesos que efectúa para cumplir su cometido. Comienza con la aplicación de un procesamiento y análisis del grafo de escena proporcionado para **generar una tripleta** de información embebida en la que se describen las entidades que actúan como sujeto, objeto y predicado. A continuación se hace uso de este conocimiento resultante para producir **máscaras de tripletas** capaces de identificar la localización de cada una de las entidades participantes en la escena de un grafo a partir del rol que juegan como sujeto u objeto. En la tercera fase se plantea la **delimitación del espacio físico** que ocupan los objetos y/o seres protagonistas del grafo de escena mediante la exploración de las máscaras resultantes, después de ordenar la presencia de las entidades y reproducir los vínculos que los interconectan a partir de los predicados identificados en la fase previa. Tal y como se puede observar en la Figura 10, el producto final se puede interpretar como una representación intermedia de un grafo de escena asociado a una imagen con la que generar las consultas estructuradas para recuperar y ordenar las fotografías coincidentes en función de los términos proporcionados.

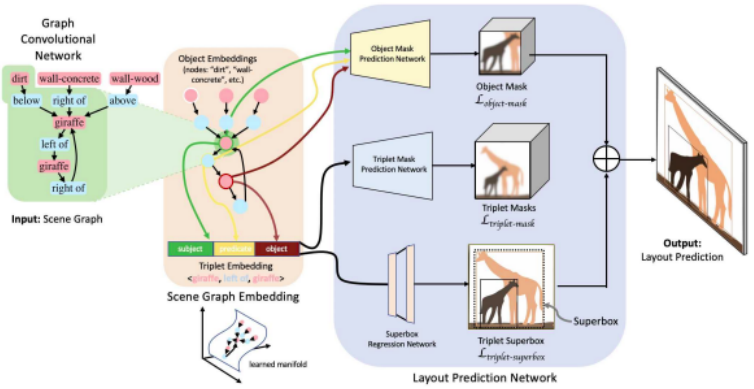

Figura 1. Arquitectura de una Red Neuronal Convolutiva de Grafos basada en generación de *embeddings* a partir de grafos de escenas [10].

Aplicaciones

Una de las primeras aplicaciones conocidas durante la revisión del Estado del Arte para los SRI basados en imágenes se fundamenta en la posibilidad de obtener aquellas **fotografías más representativas de comentarios de usuarios** con respecto a cualquier temática. Particularmente, en el artículo [9] se realizan diversos experimentos demostrativos, de la eficacia asociada a las metodologías propuestas, consistentes en generar un carrusel de imágenes del parque de atracciones Disneyland a partir de las críticas redactadas en varios portales de viajes altamente populares, como TripAdvisor.

Haciendo una revisión del listado de utilidades que hacen uso de los SRI basados en imágenes y consultas multi-atributos he podido comprobar que la mayoría de trabajos publicados se encuentran orientados al **reconocimiento de personas** a partir de una descripción cualitativa de su aspecto físico. Para ello comienzan empleando diversas técnicas de preprocesamiento de imágenes con las que identificar cada una de las **regiones situadas en el rostro**. Algunas de las más populares consisten en utilizar rejillas divisorias en combinación con mecanismos de detección de partes humanas para etiquetar lingüísticamente el contenido de cada porción [7], o bien haciendo referencia a técnicas de edición de imágenes con las que aumentar el rostro de la persona para simplificar la extracción e identificación de sus características físicas [12]. A continuación producen estructuras de datos que agilizan la recuperación de imágenes a partir de descripciones textuales, como son los **índices** basados en los atributos gráficos obtenidos de cada una de las imágenes.

Relacionada con la temática anterior también existen aproximaciones similares aunque orientadas a otro tipos de entidades, como es el caso de prendas de ropa según el artículo [13]. La formulación presentada en este trabajo es ligeramente diferente del previamente comentado puesto que han experimentado mejores resultados al dividir el conjunto de atributos posibles en dos categorías: **propiedades generales y específicas**, siendo estas más complicadas de aprender, y por ende afectan negativamente al rendimiento de los modelos. Una de las propuestas que resuelven este problema consiste en intentar convertir la representación de los atributos específicos a una descripción general a partir de la propiedad de mayor similitud. El objetivo de la siguiente etapa reside en el reconocimiento de cada una de los elementos que componen las diferentes prendas disponibles. En primer lugar se aplican técnicas de segmentación de colores para **separar el fondo de la imagen de la entidad** en cuestión, mejorando la calidad de la fotografía resultante con el fin de **eliminar huecos y regiones ruidosas** que hayan podido aparecer como consecuencia.

Dependiendo de la ubicación en la que pueda colocarse cada prenda de ropa, existe la posibilidad de generar **límites** que simplifican la identificación de cada una de las partes. Un ejemplo representativo de este procedimiento puede ser la discontinuidad que presentan en el cuello aquellas prendas de ropa que se utilizan para la parte superior, siendo considerablemente similar la situación relativa a la distinción entre las mangas y el torso. El algoritmo que lleva a la práctica esta metodología no actúa en caso de detectar que se trata de una indumentaria orientada a la parte inferior. Tras preprocesar las imágenes disponibles obteniendo sus atributos visuales más característicos, las consultas se realizan **minimizando la diferencia** existente entre los términos especificados en la búsqueda y el conjunto de propiedades lingüísticas asociadas a cada una de las imágenes, siendo prioritarias aquellas cuyo grado de coincidencia sea mayor [13].

Bibliografía

1. INFORMATION RETRIEVAL SYSTEM: CONCEPT AND SCOPE, National Institute of Open Schooling
2. Modern Information Retrieval: A Brief Overview, Amit Singhal.
3. Intelligent Image Retrieval Techniques: A Survey, Mussarat Yasmin, Sajjad Mohsin, Muhammad Sharif
4. Content-based image retrieval system with most relevant features among wavelet and color features, Abdolreza Rashno, Elyas Rashno
5. Image Retrieval Using a Combination of Keywords and Image Features, Praveen Bandikolla & Keshi Reddy Vishwanath Reddy
6. Multi-Attribute Queries: To Merge or Not to Merge?, Mohammad Rastegari, Ali Diba, Devi Parikh
7. Image Ranking and Retrieval based on Multi-Attribute Queries, Behjat Siddiquie, Rogerio S. Feris, Larry S. Davis
8. Automatic Sentence Generation from Images, Yoshitaka Ushiku, Tatsuya Harada, Yasuo Kuniyoshi
9. Ranking and Retrieval of Image Sequences from Multiple Paragraph Queries, Gunhee Kim, Seunghwan Moon, Leonid Sigal
10. Structured Query-Based Image Retrieval Using Scene Graphs, Brigit Schroeder, Subarna Tripathi
11. Simplifying Graph Convolutional Networks, Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, Kilian Q. Weinberger
12. Dynamic multi-attribute priority based face attribute detection for robust face image retrieval system, S. Suchitra, R.J. Poovaraghan

13. Efficient Multi-attribute Similarity Learning Towards Attribute-Based Fashion Search, Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, Ashraf A. Kassim.