***Modelo Vectorial***

El modelo de espacio vectorial se basa en el grado de similaridad de una consulta dada por el usuario con respecto a los documentos de la colección cuyos términos fueron ponderados mediante TF-IDF. El modelo vectorial fue presentado por Salton en 1975 y posteriormente asentado en 1983 junto con Mc Gill y se basa en tres principios esenciales (MARTÍNEZ COMECHE, J.A. 2006):

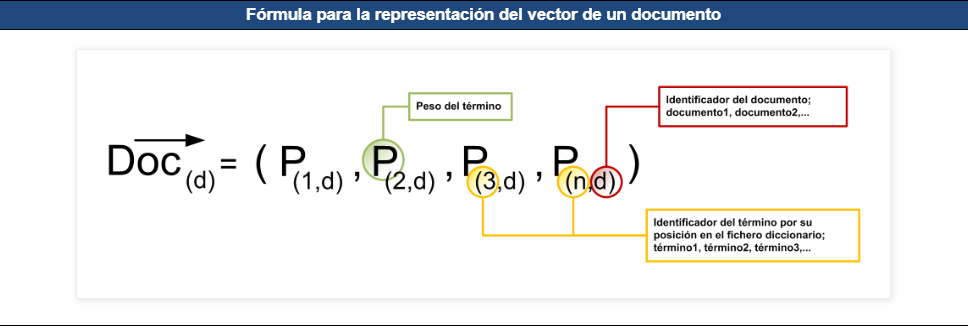
* *La equiparación parcial, esto es, la capacidad del sistema para ordenar los resultados de una búsqueda, basado en el grado de similaridad entre cada documento de la colección y la consulta.*
* *La ponderación de los términos en los documentos, no limitándose a señalar la presencia o ausencia de los mismos, sino adscribiendo a cada término en cada documento un número real que refleje su importancia en el documento.*
* *La ponderación de los términos en la consulta, de manera que el usuario puede asignar pesos a los términos de la consulta que reflejen la importancia de los mismos en relación a su necesidad informativa.*

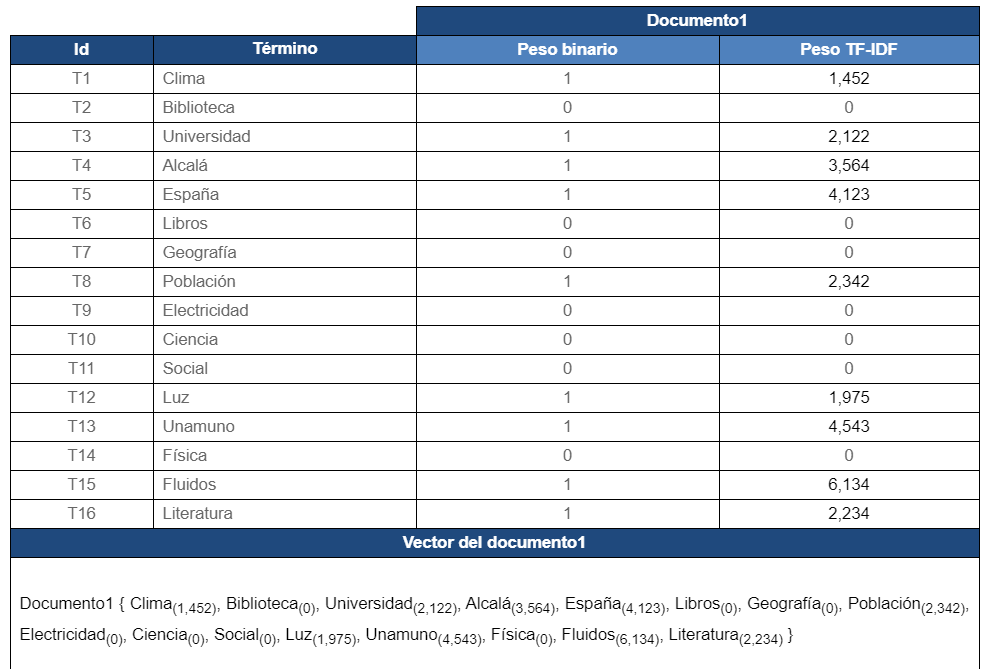
Si bien en el modelo booleano un documento de la colección puede ser representado por la presencia o ausencia de los términos indexados en el fichero diccionario de la siguiente forma...

Documento1 { 1,0,1,1,1,0,0,1,0,0,0,1,1,0,1,1 }

...en el modelo de espacio vectorial se emplea el peso de los términos para cada documento, que refleja la relevancia de los términos del documento de cara a su representatividad en la colección, adquiriendo una forma como la que sigue...

Documento1 { 1`452, 0, 2`122, 3`564, 4`123, 0, 0, 2`342, 0, 0, 0, 1`975, 4`543, 0, 6`134, 2`234 }

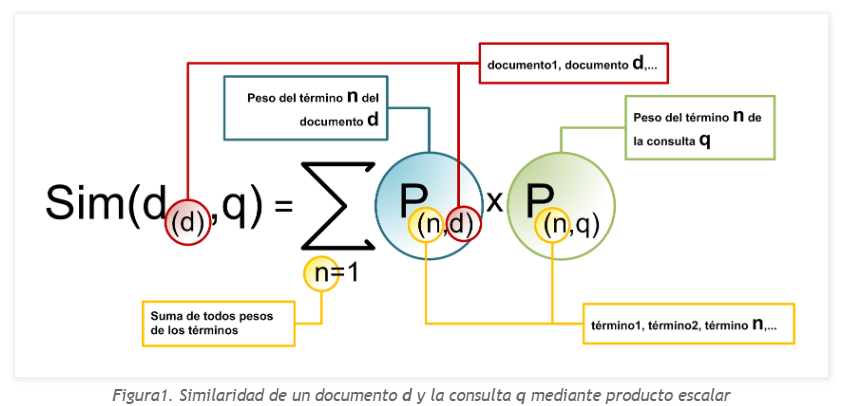
A este conjunto de números reales, que son los pesos, que representan al documento, se les denomina ***Vector del documento***, permitiendo su representación en el espacio vectorial y en  consecuencia, su tratamiento matemático. Por ello la formulación del vector se representa de la  siguiente forma, véase *tabla1*.



Posteriormente, la colección sigue lo que se denomina un ***Proceso de Vectorización*** por el que todos los documentos son representados mediante pesos TF-IDF, la consulta del usuario también requiere de dicho tratamiento. Ello significa que se tiene que ponderar la importancia de los términos de la consulta para poder generar el ***Vector de la consulta del usuario***. Este paso es imprescindible para poder efectuar el ***Proceso de Equiparación de la consulta con los documentos*** de la colección y determinar cuáles de ellos son más relevantes.

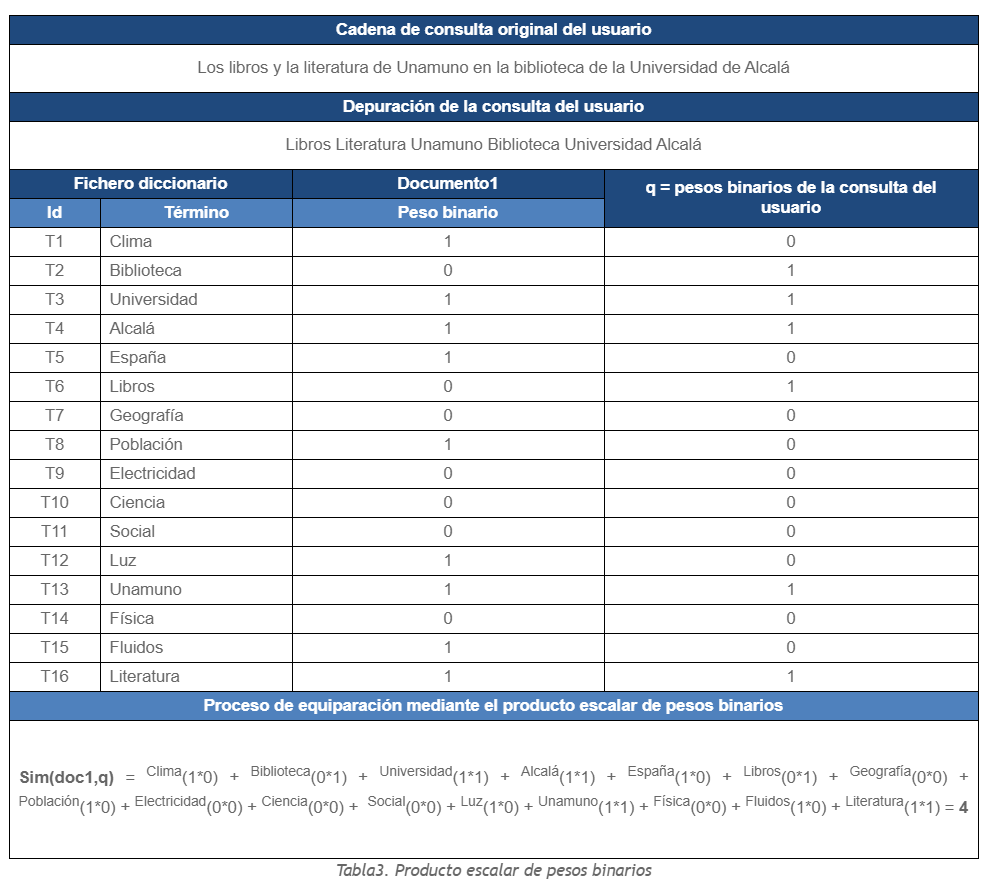
Proceso de equiparación mediante el producto escalar

Los procesos de equiparación de los documentos de la colección con respecto a la consulta del usuario, en el modelo booleano, se efectúan mediante cálculos de similaridad. Existen muchas modalidades de comparación o equiparación mediante similaridad, en este caso se presenta una de las más sencillas por su simplicidad y sistematización inmediata. Se trata del producto escalar de los pesos, véase *figura1*.

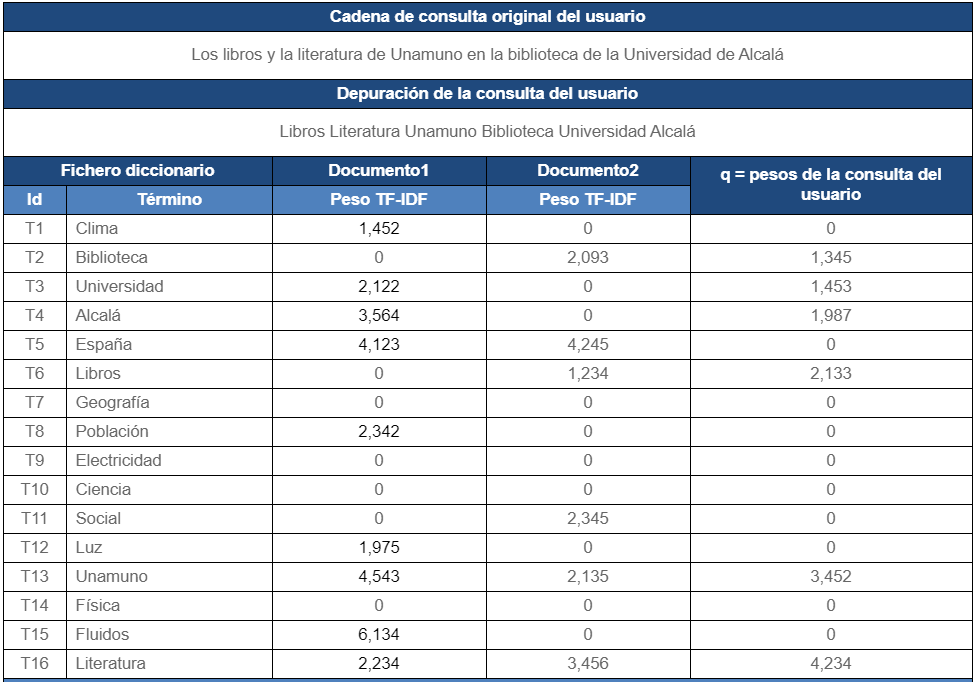


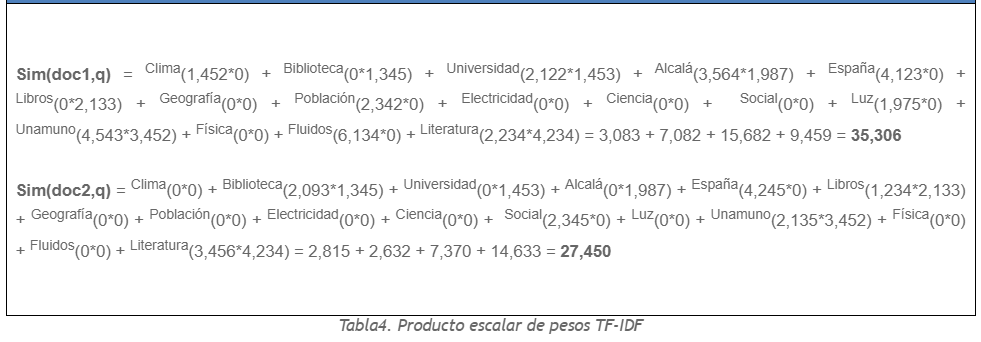
De esta forma, la similaridad de un documento y una consulta, es igual a la suma de los productos de sus pesos. (Y no se debe olvidar que cada peso representa a un término). Este método puede aplicarse tanto a pesos binarios como a pesos TF-IDF.

**Modalidad de pesos binarios** En el caso de la modalidad binaria, la similaridad de un documento con respecto a la consulta es equivalente a la presencia de los términos de la consulta en el documento, véase *tabla3*. Esto quiere decir que la ausencia de un término de la consulta o del documento implica un producto igual a 0 y por lo tanto no tienen incidencia en el cálculo. Por el contrario la presencia de un término dado tanto en la consulta como en el documento siempre tendrá valor 1. Por ello sólo basta con contabilizar el número de términos coincidentes de la consulta en el documento y ése será su valor de similaridad.



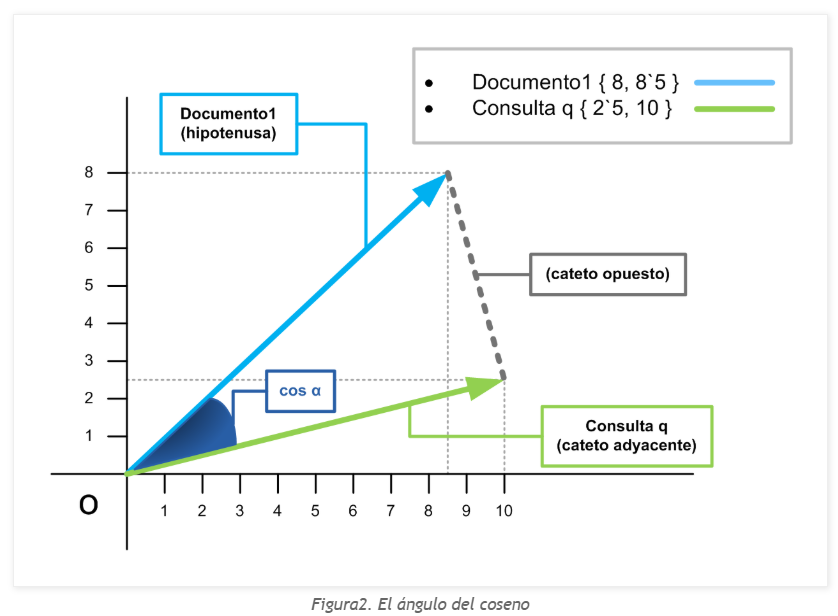
**Modalidad de pesos TF-IDF**

En el caso de la modalidad de pesos binarios, las limitaciones en la definición de la representatividad de los términos de cada documento quedan patentes. Resulta por tanto un resultado bastante limitado y parcial. Por ello el método de la similaridad mediante el producto escalar se aplica habitualmente con pesos TF-IDF, mucho más precisos, véase *tabla4*.

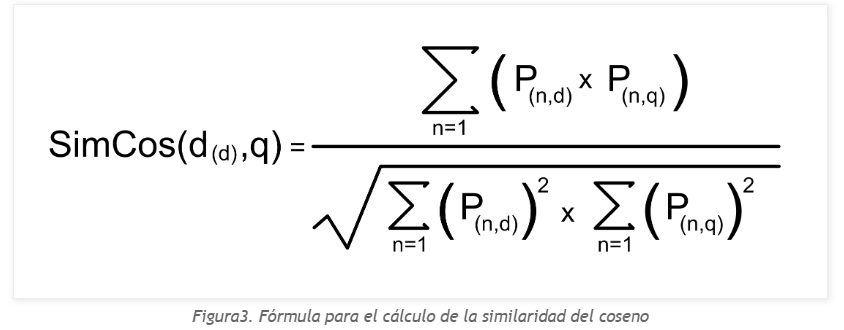


El cálculo de la similaridad se aplica a cada uno de los documentos de la colección siguiendo el patrón expuesto en la *tabla4*. Para el **documento1** la similaridad con respecto a la consulta del usuario **q**, será diferente que para el **documento2**. Obsérvese que al igual que ocurria con los pesos binarios, sólo tienen incidencia aquellos términos presentes tanto en la consulta como en el documento, pues sus pesos se multiplican y se suman sucesivamente al resto. En este caso, la similaridad del documento1 (35,306) es superior a la del documento2 (27,450), siendo éstas unas cifras mucho más precisas que un simple número entero.  
  
Proceso de equiparación mediante la fórmula del coseno

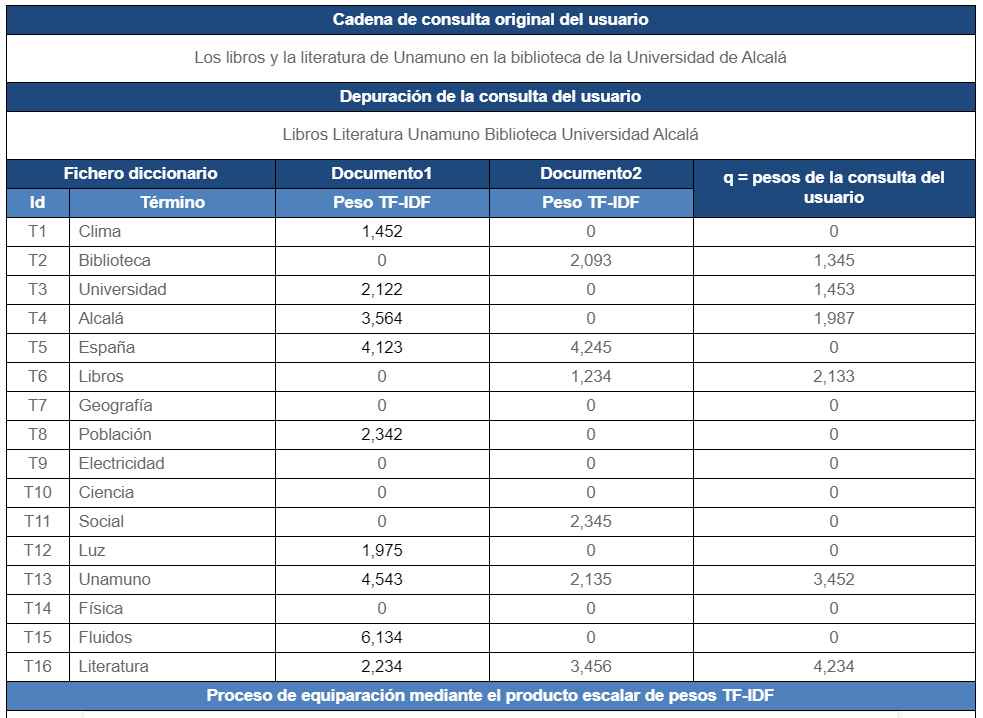
Tal como se ha explicado en la fórmula del producto escalar, el proceso de equiparación es posible cuando en el vector de la consulta y en el del documento existen términos coincidentes. Pero este enfoque no supone la representación del vector de la consulta y del documento. De hecho una de las claves del modelo de espacio vectorial es precisamente la posibilidad de determinar el ángulo que forman los vectores del documento y de la consulta que se está comparando, véase *figura2*.

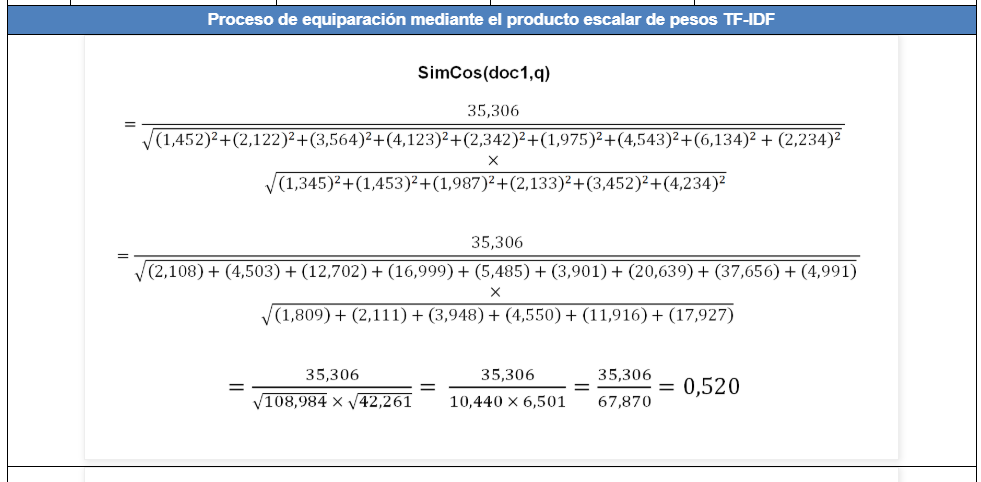
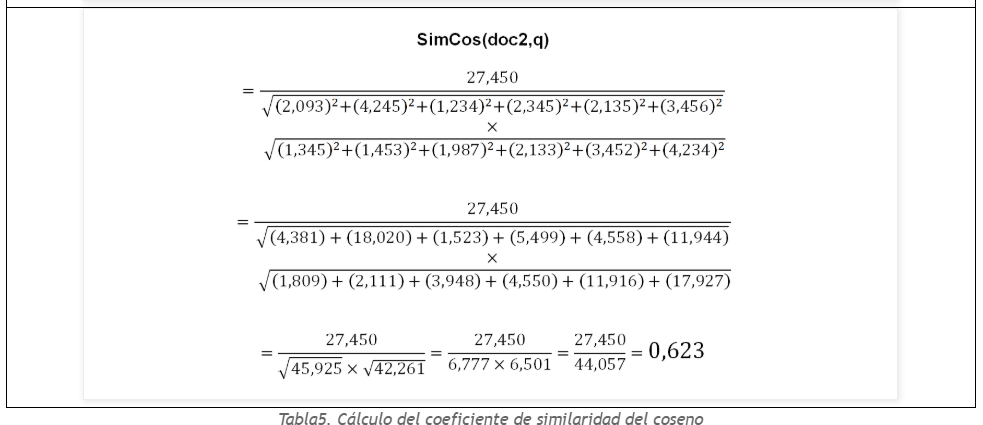


Es posible medir cuál es la desviación de un documento con respecto a una consulta, por el número de grados del ángulo que forman. Esto es posible porque crean una estructura triangular a la que se aplica el cálculo del ángulo que forma la hipotenusa (en este caso el vector del **documento1**) y el adyacente (el vector **q** de la consulta dada por el usuario) que resulta ser el coseno del triángulo. En el caso de la *figura2*, se comprueba visualmente cierta distancia del vector de la consulta con respecto al documento1; cuando ambos vectores se muestran tan próximos como para superponerse, implicará que el ángulo que forman será menor y que su nivel de coincidencia será superior. De hecho, un coseno de 0º implicaría una similaridad máxima.



Por lo tanto, la fórmula aplicada para calcular el coeficiente de similaridad del coseno entre un documento y una consulta es aquella que permite poner en relación los vectores de la consulta y del documento. De hecho el coseno de alfa de un triángulo cualquiera siempre es igual al cateto adyacente entre la hipotenusa. Tomando como clave esa idea, la *figura3* muestra la misma relación pero esta vez con los pesos que forman los vectores del documento y la consulta. De hecho el numerador no deja de ser un producto escalar entre los pesos del documento y la consulta; y el denominador la raíz cuadrada del producto del sumatorio de los pesos del documento y la consulta al cuadrado. La formulación del denominador con raíz cuadrada y cálculo de cuadrados, se diseñó para conseguir un resultado final de la división, inferior a 1, de tal manera que el coeficiente fuera de fácil manejo y lectura. La similaridad del coseno aplicada al ejemplo que se viene utilizando, tendría la forma que sigue a continuación en la *tabla5*.





Como se puede observar en los resultados del coeficiente de similaridad del coseno para el documento1 y 2 en la *tabla5*, son diametralmente distintos a los obtenidos en la *tabla4*. Esto significa que los pesos de los términos del documento2, lo convierten en más representativo y probablemente más relevante que el documento1, dando por lo tanto una mayor precisión que el cálculo del producto escalar. El máximo valor del coeficiente de similaridad del coseno es 1, que equivaldría a un ángulo de 0º entre los vectores del documento y la consulta.

Ventajas e inconvenientes del modelo vectorial

**Ventajas**

* El modelo vectorial es muy versátil y eficiente a la hora de generar rankings de precisión en colecciones de gran tamaño, lo que le hace idóneo para determinar la equiparación parcial de los documentos.
* Tiene en cuenta los pesos TF-IDF para determinar la representatividad de los documentos de la colección.

**Inconvenientes**

* El modelo vectorial por producto escalar tiene la desventaja de que sólo tiene en cuenta la intersección de los términos del documento con respecto a la consulta, por lo que la gradación de los resultados no es tan precisa como en el caso del cálculo del coseno.
* Necesita de la intersección de los términos de la consulta con los documentos, en caso contrario no se produce la recuperación de información.
* Al ser un modelo estadístico-matemático, no tiene en cuenta la estructura sintáctico-semántica del lenguaje natural.

***Vector space model adaptation and pseudo relevance feedback for content-based image retrieval***

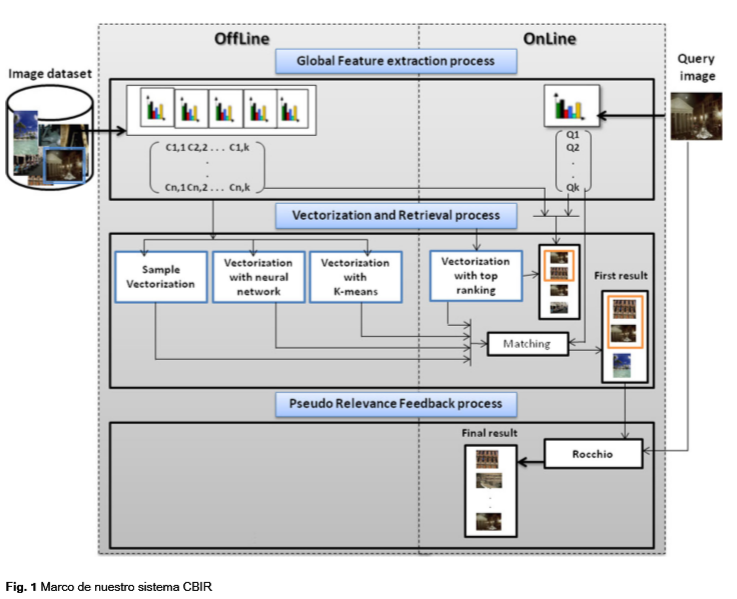
La recuperación de imágenes es un problema importante para los investigadores en los campos de la visión por ordenador y la recuperación de imágenes basada en contenido (CBIR). Durante las últimas décadas, muchos sistemas de recuperación de imágenes se basaron en la representación de imágenes como un conjunto de características de bajo nivel extraídas como el color, la textura y la forma. A continuación, los sistemas calculan las métricas de similitud entre entidades para encontrar imágenes similares a una imagen de consulta. La desventaja de este enfoque es que las imágenes visualmente y semánticamente diferentes pueden ser similares en el espacio de entidades de bajo nivel. Por lo tanto, es necesariodesarrollarherramientasparaoptimizarrecuperarinformación. Los modelos de espacio Integrationofvector es una solución para mejorar el rendimiento de la recuperación de imágenes. En este artículo, presentamos un marco de recuperación eficiente y eficaz que incluye una técnica de vectorización combinada con un modelo de pseudorelevancia. La idea es transformar cualquier modelo de coincidencia de similitud (entre imágenes) en un modelo de espacio vectorial que proporcione una puntuación. Un estudio donde presentan varias metodologías para obtener la vectorización. Se han realizado algunos experimentos en conjuntos de datos de Wang, Oxford5k e Inria Holidays para mostrar el rendimiento de nuestro marco propuesto.

Una mejora significativa en el rendimiento de los sistemas CBIR se puede conseguir mediante la integración de las técnicas de recuperación de información [ 40 ] Tal como el modelo de espacio vectorial (VSM) [ 55 ]. Con la VSM, las técnicas de recuperación de información [ 40 ] Tal como el modelo de espacio vectorial (VSM) [ 55 ]. Con la VSM, las técnicas de recuperación de información [ 40 ] Tal como el modelo de espacio vectorial (VSM) [ 55 ]. Con la VSM, las técnicas de recuperación de información [ 40 ] Tal como el modelo de espacio vectorial (VSM) [ 55 ]. Con la VSM, las técnicas de recuperación de información [ 40 ] Tal como el modelo de espacio vectorial (VSM) [ 55 ]. Con la VSM, las imágenes cambian de (a veces) la representación inexacta de su contenido a una representación más exacta con las puntuaciones. Esta adaptación se llama técnica de vectorización y inspirado en el trabajo de Vincent Claveau et al. Esta adaptación se llama técnica de vectorización.

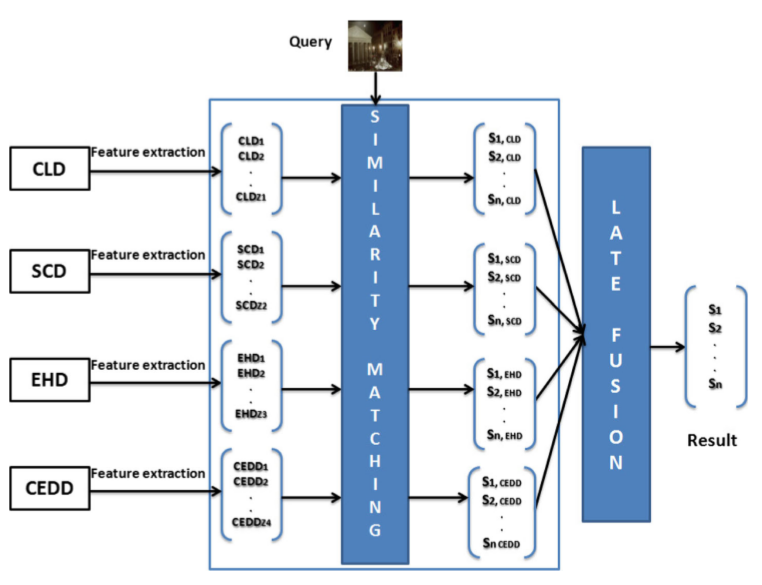
En este trabajo, presentamos dos contribuciones.

* Nuestra primera contribución es proponer un marco global eficiente y eficaz que utiliza características globales y funciona tan bien como marco el uso de características locales. Este marco incluye nuestra adaptación de la VSMmodel que transforma el modelo de adaptación de recuperación de imágenes a un modelo coincidente vectorial. Denotamos esta técnica como una técnica de vectorización en este documento. Nuestra VSMmodel representa cada imagen y consulta por vectores expresados ​​en un espacio en el que cada dimensión corresponde a un descriptor atómica. Cada coordenada en este espacio refleja la importancia del descriptor en la imagen considerada. El procesamiento de consultas se basa entonces en la comparación de los vectores considerados.
* La segunda contribución es incluir en nuestro marco de diferentes procesos previos y posteriores que mejora el rendimiento de nuestra propuesta. Los resultados del experimento destacan el potencial de todas las técnicas de vectorización desarrollados descriptores de mezcla después y antes de la recuperación. Una discusión sobre la complejidad y los resultados muestran nuestro proceso de recuperación propuesto es menos complejo que el proceso original.

***Sistema de marco CBIR***

El diagrama representa los diferentes componentes de nuestro En esta sección, presentamos nuestro marco del sistema (ver Fig. 1 ). El diagrama representa los diferentes componentes de nuestro sistema de investigación. Sus tres componentes principales: la extracción de características y de recuperación, incluyendo el proceso de retroalimentación pseudo-relevancia se detallan individualmente en las siguientes secciones.

* Extracción de características:
  + Extracción de características adecuadas a partir de imágenes y almacenarlas en la base de datos.
  + Cada imagen está representada por un vector de características. Este proceso se lleva a cabo fuera de línea para todas las imágenes de la base de datos.
  + Utilizamos descriptores globales, para la extracción de características, porque buscamos para describir la imagen en su conjunto.
  + El coste de extracción de características globales es inferior a las características locales, también la comparación entre los vectores de características globales es muy simple de los vectores de características locales.
    - Más específicamente se utilizó Diseño de color (EPC) [ 32 ].
    - Scalabe Color (SCD) [ 54 ].
    - Edge Histogram (EHD) [ 45 ].
    - Color y el borde de directividad de descriptores (CEDD) [ 12 ]
* Proceso de recuperación:
  + Una imagen de consulta Q por lo tanto, es un vector diseñado por ( C q, 1, C q, 2 ... C q, f ... C q, Z p) valores de características, es el descriptor individuo considerado donde C q, f es el valor de característica F en la consulta y Z p es el número de características. Para comparar dos imágenes yo yo y yo j, utilizamos la distancia euclidiana [ 28 ].
  + En la fusión tarde, el cálculo de la similitud entre las imágenes se realiza para cada descriptor y fusionada por lo tanto, en nuestro caso. una imagen yo yo está representado por cuatro vectores de características (ver fig. 3 ).
  + Para cada descriptor, buscamos la K imágenes de vecino más próximo, que serán almacenados en una lista de resultados. Esta lista será útil para los algoritmos de fusión finales.



* Proceso de retroalimentación Pseudo-Relevancia:
  + retroalimentación pseudo-relevancia, desarrollado originalmente para la recuperación de información [ 52 ], Se utiliza para mejorar la eficacia de los sistemas de recuperación de información. La idea principal de retroalimentación pseudorelevance es utilizar ejemplos positivos para mejorar el rendimiento del sistema. Para una consulta dada, el sistema obtiene primero una lista de imágenes clasificados de acuerdo con las métricas de similitud predefinidos, que a menudo se define como la distancia entre vectores de características de imágenes.

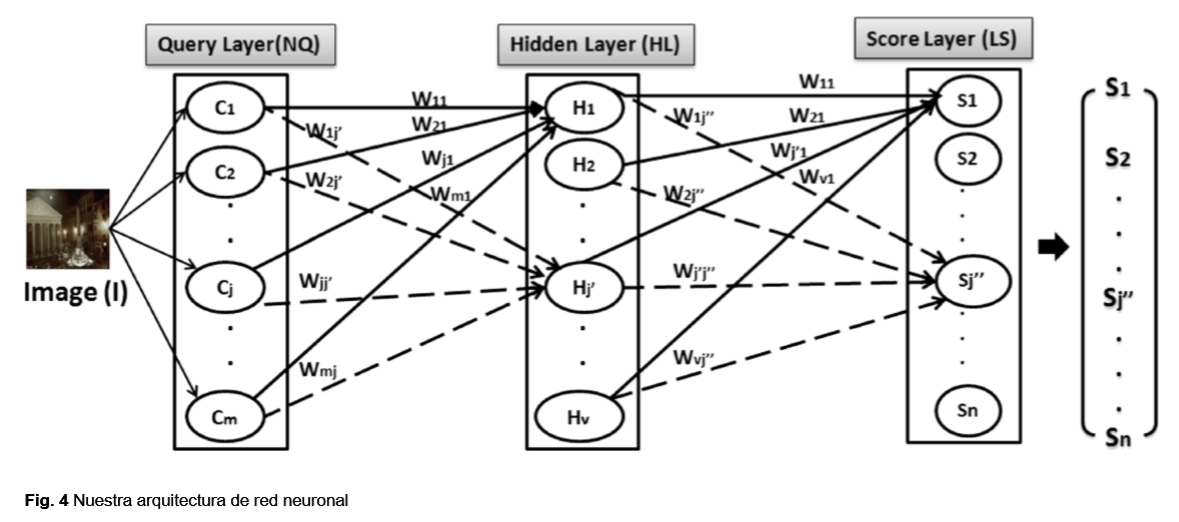
***Proceso de vectorización***

El sistema CBIR propuesto pone a prueba varios métodos de recuperación para la recuperación basada en el contenido dependiendo de las características visuales de la imagen. Se describe en esta sección se aproxima a nuestra vectorización.

* Principio de vectorización
  + La idea de vectorización es utilizar una serie de imágenes R llamadas de Referencia imágenes. Para cada imagen a partir de una base de datos, calculamos su similitud con cada una de estas imágenes de referencia obtenidos con el sistema de recuperación inicial (cada imagen en la colección es considerada como una consulta). la R en valores recogidos son un vector para las imágenes de recopilación de datos que caracterizan a cada imagen. Este proceso es costoso y se realiza fuera de línea, que está en línea durante las consultas.
  + Para aplicar el proceso de vectorización, se calcula la similitud entre las imágenes de referencia y las imágenes de la colección.

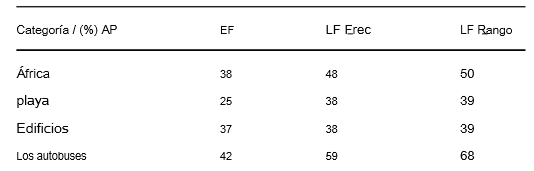
***Vectorización basada en redes neuronales***

En esta sección, se presenta un nuevo método para transformar un proceso de recuperación de imágenes basada en contenido en un simple modelo de espacio vectorial, que construye la conexión entre la imagen de búsqueda y la puntuación resultado a través de una red neuronal. Esto funciona sin vectorización de Referencia imágenes [ 30 , 31 ], Pero con una matriz de peso ( W). Para ello, hemos probado una red neuronal con uno capas ocultas.

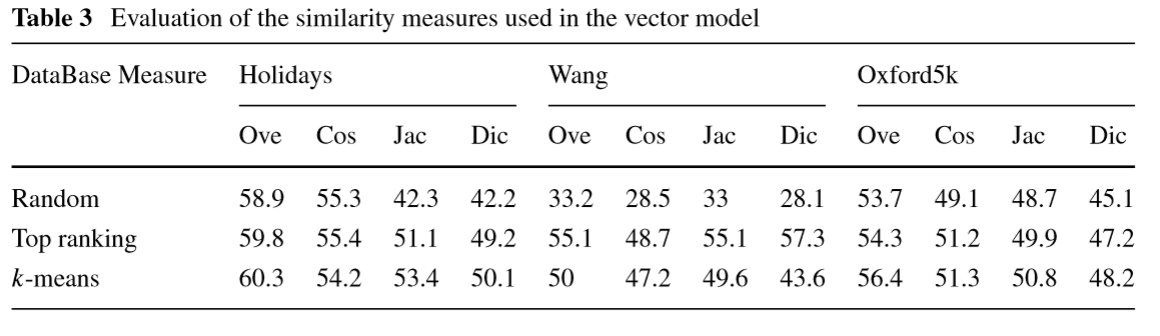


Resultados experimentales:

En esta sección, evaluamos los métodos de fusión de combinación de características. Comparamos los resultados de los tres métodos siguientes: fusión temprana (EF), la fusión a finales de acuerdo con la frecuencia (LF Frec) y la fusión a finales de acuerdo al rango (LF Escala).



En la recuperación de informaion, para medir la similitud entre documentos utilizando el modelo de espacio vectorial, se utilizan cuatro diferentes métricas: coseno (cos) ( 10 ), La superposición (OVE) ( 5 ), Jaccard (JAC) ( 11 ) Y Dados (DIC) ( 12 ). En esta sección, evaluamos estas medidas para encontrar la mejor medida que corresponde a nuestro modelo de espacio vectorial.



Conclusiones:

El trabajo futuro se centra en la aplicación de la vectorización en características de alto nivel. También pretendemos explorar técnicas semánticas para extraer "palabras" de imágenes basadas en un conjunto específico de datos.