Evaluación Final de PLN
Modelo de Espacio Vectorial

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Mayo 2020

Índice

1 Motivación y definición del MEV

1.1 Motivación

Una colección de documentos (p. ej. páginas web) no está preparada para realizar directamente operaciones de R.I. Necesitamos crear estructuras persistentes que permitan acceder eficientemente a los datos ya procesados.

1.2 Definición del MEV

En un modelo de espacio vectorial (MEV), los documentos se almacenan como vectores de términos y pueden encontrarse en un espacio vectorial de n dimensiones, en grupos formando colecciones de documentos de acuerdo con la relevancia para una misma clase de necesidad de información. Es decir, cuando queremos acceder a cierta información usando la consulta, esa consulta llega a ser comparada con cada elemento de la colección, si se obtiene un alto grado de coincidencia como consecuencia el documento tendrá más probabilidades de ser relevante, y se nos devolverá lo que buscamos.

El proceso para construir un MEV comienza con la extracción de los términos de indización, es decir, los términos que van a ser utilizados para describir el contenido del documento. La posibilidad más simple consiste en considerar todas las palabras aisladas que aparecen en el texto como los términos de indización. Habitualmente se eliminan algunas palabras (las denominadas palabras vacías, entre las que suelen figurar números, preposiciones, conjunciones, etc.). Una vez extraídas las palabras del texto, se ordenan por orden alfabético y se guardan en un matriz, junto con la referencia del documento de donde proceden (normalmente un número de documento asignado previamente por el sistema). Si se repite este proceso con todos los documentos de la colección, obtenemos finalmente una matriz que almacena los siguientes datos:

- En primer lugar, los términos de indización (las palabras) que aparecen en toda la colección (ya sean los propios textos, los resúmenes de los textos del fondo y/o los títulos).
- En segundo lugar, cada uno de estos términos (palabras) incorpora una lista con los números de los documentos en los que aparece.

Por ejemplo, el objetivo principal de una colección de n documentos indexados por m términos que se puede representar por una matriz A de dimensión n x m, donde cada elemento a_{ij} es definido por una frecuencia ponderada del término i en el documento j es mejorar el rendimiento en la habilidad de recuperar información relevante y descartar información irrelevante.

La siguiente tabla (ver Tabla 1) muestra la matriz A, donde cada fila representa un término en la colección, cada columna un documento y cada celda o elemento de la matriz la ocurrencia del término en el documento.

En ella podemos ver que el término 1 aparece en el documento 1, pero no en los documentos 2 y 3, y así se puede demostrar que cada fila de la matriz de 3×3 puede ser representada en un espacio

	D1	D2	D3
T1	1	0	0
T2	0	0	1
T3	1	1	1

Table 1: Matriz A.

de tres dimensiones.

Más formalmente, cada elemento a_{ij} de la matriz A queda definido como:

$$a_{ij} = l_{ij} * g_i * d_j^-1$$

donde l_{ij} es el peso local del término i en el documento j, el cual mide la importancia de dicho término en el documento, g_i es el peso global del término i en la colección de documentos y d_j es el factor de normalización para el j-ésimo documento.

2 Tipos de MEV

2.1 Modelo Booleano

En la tabla de ejemplo (ver Tabla 1) de la sección anterior, se muestra un tipo de MEV: el modelo booleano. Este modelo constituye el primer modelo teórico de RI, el más antiguo, empleado para establecer un subconjunto de documentos relevantes. Al mismo tiempo es, sin duda, uno de los modelos más sencillos tanto desde un punto de vista teórico como práctico, por una parte, al basarse en la teoría de conjuntos y en el álgebra de Boole, y por otra parte al ser fácil de diseñar e implementar.

Una propiedad importante del modelo booleano es que no puede efectuar ningún proceso de ordenación con los documentos resultantes de las consultas. Esta característica suele denominarse **equiparación exacta**, impidiendo que se pueda situar en primer lugar aquel documento posiblemente más útil o relevante para el usuario y relegando a las últimas posiciones a aquellos otros documentos con menos probabilidades de ser relevantes en relación a la consulta.

También se denomina modelo binario por su carácter binario porqué considera exclusivamente la presencia (con el número 1) y la ausencia (con el número 0), los términos en los documentos. Pero ese carácter es el principal responsable de la equiparación exacta, siendo considerada la principal desventaja del modelo.

A pesar de esta desventaja, todavía hoy sigue constituyendo el modelo más habitual. Muchos motores de búsqueda en la web se basan en este modelo, por ser de desarrollo sencillo (como hemos visto, en su versión básica solamente involucra el empleo de una matriz y una interfaz de consulta que permita realizar consultas expresadas mediante palabras o expresiones booleanas), fácil de utilizar por parte de un usuario medio (basta introducir palabras relativas a la necesidad informativa), y bastante eficaz en los resultados obtenidos (en gran parte debido al volumen ingente de documentación presente en la red, lo que provoca que la reducción de la respuesta a los documentos que satisfagan estrictamente las condiciones de la consulta.

Observando mejor el ejemplo de la sección anterior (ver Tabla 1) vemos que de la consulta de la matriz A únicamente se puede saber si un determinado término de indización está presente (en

cuyo caso se simboliza con el número 1) o no lo está (en cuyo caso se simboliza con el número 0) en cada uno de los documentos de la colección. De manera que la matriz A se puede generalizar por la siguiente tabla cuyos datos básicos son:

	D1	D2	...	D_n
T1	1	0	...	1
T2	0	1	...	0
...	...	...	...	...
T_n	0	1	...	1

Table 2: Matriz A generalizada.

donde T1, T2, ..., T_n son los términos de indización empleados en la colección de documentos D1, D2, ..., D_n y donde el número 1 significa que el término correspondiente aparece en ese documento concreto, mientras que el número 0 significa que el término no aparece en dicho documento. Esto implica que no se tiene en cuenta la frecuencia de aparición de los términos en los documentos. Se entiendo entonces porqué se denomina modelo binario, pues únicamente se juega con dos posibilidades: la aparición y la no aparición de los términos en los documentos.

Si observamos la tabla anterior (ver Tabla 2), podemos deducir de ella las dos representaciones empleadas al manejar un modelo binario. Por una parte, cada término de indización se representa por la lista de documentos en los que aparece, lo que implica la observación de la tabla por filas:

$$\begin{aligned} \text{T1} &= \{\text{D1, ..., D}_n\} \\ \text{T2} &= \{\text{D2, ..., D}_n\} \\ &\vdots \\ \text{T}_n &= \{\text{D2, ..., D}_n\} \end{aligned}$$

Por otra parte, cada documento se representa por una lista de ceros y sus correspondientes términos de indización que contiene, lo que implica la observación de la tabla por columnas:

$$\begin{aligned} \text{D1} &= \{1, 0, ..., 0\} \\ \text{D2} &= \{0, 1, ..., 1\} \\ &\vdots \\ \text{D}_n &= \{1, 0, ..., 1\} \end{aligned}$$

Se comprende bien ahora por qué se dice que en el modelo binario todo documento se representa mediante una serie ordenada de ceros y unos, tantos como términos se empleen en una colección: des del primer número que siempre corresponderá a T1, el segundo número corresponderá a T2, y así sucesivamente hasta llegar a T_n, siendo n el número de términos distintos que representan el contenido de esa colección.

2.2 Modelo Probabilístico

El modelo probabilístico fue introducido en la década de los setenta por Robertson y Sparck Jones. También es conocido como modelo de recuperación de independencia binaria (RIB). Este modelo evitando el empleo de fórmulas matemáticas, se basa en las siguientes consideraciones:

- Para caracterizar los documentos de la colección se emplean ciertos términos de indización.

- Dada una necesidad informativa del usuario, existe un subconjunto de documentos de la colección que contiene exclusivamente los documentos relevantes con relación a ella.
- Se parte exclusivamente de la presencia o ausencia de los términos en los documentos de la colección. Se trata, pues también, de un modelo binario, como el modelo booleano.
- El usuario no sabe cuáles son los términos de indización que configurarían la consulta ideal. Tampoco sabe en qué medida los términos empleados en la consulta permiten discernir de los documentos relevantes y rechazar simultáneamente los documentos irrelevantes.
- Actúa sobre los términos que configuran la consulta del usuario, ponderándolos, esto es, imponiéndoles un peso a cada uno de ellos, mayor cuanto mejor permita discernir los documentos relevantes de los irrelevantes, y menor en caso contrario. De esta manera se persigue que el modelo efectúe la recuperación de información incidiendo sobre todo en los mejores términos de entre todos los empleados por el usuario en la consulta, minimizando la importancia de aquellos otros términos que, aun figurando en la consulta, son malos términos del conjunto de respuesta ideal.
- Como tampoco se puede saber a priori cuáles, de entre los términos que configuran la consulta, son buenos términos y cuáles no lo son, a este modelo no le queda otro remedio que considerar, para cada uno de los términos empleados en la consulta, la *"probabilidad de ser buen término"* (probabilidad de que el término empleado en la consulta esté presente en un documento del conjunto de documentos relevantes en relación a la consulta) y simultáneamente, para ese mismo término, la *"probabilidad de ser un mal término"* (probabilidad de que ese mismo término esté presente en un documento del conjunto de documentos irrelevantes en relación a la consulta).
- Como las probabilidades nombradas anteriormente son desconocidas inicialmente a la consulta, este modelo se ve en la necesidad de efectuar una hipótesis inicial sobre sus valores. La obligatoriedad de hacer una hipótesis inicial sobre las *"probabilidades de ser buen término o de ser un mal término"* para cada término de la consulta se considera el principal inconveniente de este modelo.

El modelo probabilístico es capaz de calcular el grado de similitud existente entre cada documento de la colección y la consulta ponderada, consiguiendo ordenar los documentos de la colección en orden descendente de probabilidad de relevancia en relación con la consulta. De esta manera el modelo supera el gran inconveniente puesto de manifiesto en el modelo booleano, la equiparación exacta. En efecto, el modelo probabilístico, aun siendo un modelo binario, efectúa **equiparación parcial**, lo que permite ordenar los documentos de la respuesta conforme a su probabilidad de relevancia. Ya que no puede ponderar los términos de la colección ya que es un modelo binario, la equiparación parcial es posible gracias a la ponderación de los términos empleados en la consulta.

Una de las grandes aportaciones del modelo probabilístico a la recuperación de información consiste en el fenómeno denominado retroalimentación por relevancia. Que consiste en la utilización de información generada en procesos de recuperación anteriores o durante el propio proceso de consulta para mejorar los resultados de la recuperación de información solicitando al usuario, tras una respuesta inicial a la consulta, que analice los documentos recuperados y valore cuáles son relevantes. Con esta información se imponen nuevos pesos a las *"probabilidades de ser buen término o de ser un mal término"* para cada término de la consulta, obteniéndose así una nueva respuesta de documentos ordenados por su probabilidad de relevancia, gracias a la información suministrada directamente por el usuario.

Actualmente hay muchos sistemas de recuperación de información que emplean alguna variante de la retroalimentación por relevancia para mejorar y refinar los resultados de las consultas. Quizá la más conocida se base en la sugerencia al usuario de más resultados precedidos de la siguiente advertencia: *"Otros usuarios que adquirieron o preguntaron por ese documento también adquirieron o preguntaron por estos otros"*. Es una manera de emplear la información procedente, en este caso, de procesos de recuperación anteriores.

2.3 Modelo Vectorial

Como se ha observado en las dos secciones anteriores, el modelo probabilístico supera al modelo booleano ya que el probabilístico efectúa equiparación parcial mientras que el modelo booleano efectúa equiparación exacta. Sin embargo, ambos siguen presentado una característica negativa: ni el modelo booleano ni el modelo probabilístico tienen en cuenta la frecuencia con la que aparecen los términos de indización dentro de los documentos, porqué ambos son modelos binarios.

Parece lógico pensar que, si en un documento aparece un término una vez, y en otro documento aparece ese mismo término veinte veces, consideremos que en el primer documento la importancia del término es menor que ese mismo término en el segundo documento. En consecuencia, surge un tercer modelo de recuperación, el modelo vectorial.

El modelo vectorial fue presentado por Salton en 1975 y posteriormente asentado en 1983 junto con McGill. Fue usado por primera vez por el sistema SMART de recuperación de información. Utiliza pesos no binarios para los términos de los documentos para así poder computar el grado de similitud entre documentos y consultas. Esto permite que el conjunto de documentos obtenidos como resultado de una consulta pueda ser ordenado por un ranking de relevancia.

Se basa en los tres principios:

- La equiparación parcial, que es la capacidad de ordenar los resultados de una búsqueda, basándose en el grado de similitud 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 estos, sino ponderando a cada término en cada documento un número real que refleje su importancia en el documento.
- La ponderación de los términos en la consulta, es la manera que el usuario pueda asignar pesos a los términos de la consulta que reflejen la importancia de estos en relación a su necesidad informativa.

Esa ponderación de números reales, que son los pesos, que representan al documento, se le denomina vector del documento, permitiendo su representación en el espacio vectorial y en consecuencia, su tratamiento matemático. Es decir, en el modelo vectorial tanto un documento como la consulta se representan mediante un vector de pesos para determinar la representatividad de los documentos de la colección. Por ello la formulación del vector se representa de la siguiente forma, ver Figura 1.

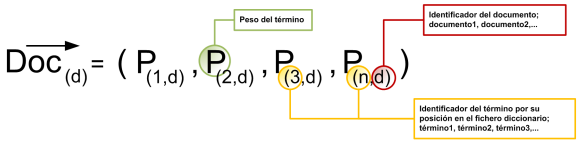

Figure 1: Representación del vector de un documento.

Donde n es el número total de términos considerados en la descripción de la colección y cada peso P es el producto de los valores TF-IDF , que se describen en la sección 3. Formalmente,

$$\mathrm{P}_{i,d} = \mathrm{TF}_{i,d} * \mathrm{IDF}_i$$

Gracias a esta representación, los documentos y las consultas se tratan matemáticamente como vectores en un espacio n dimensional, y da el nombre al modelo.

Veamos un ejemplo. Si consideramos únicamente dos dimensiones (dos únicos términos), concretamente los documentos D1 = (3, 5) y D2 = (4,1) y la consulta Q = (2,1) de un espacio bidimensional, se pueden dibujar tanto los documentos como la consulta en el plano de este documento, como *"flechas"* o vectores que parten del origen de coordenadas, cuyo primer número corresponde al valor del término 1 representado en el eje de abscisas y cuyo segundo número corresponde al valor del término 2 representado en el eje de ordenadas.

En este ejemplo obtendríamos el siguiente gráfico, donde D1 es el vector más próximo al eje de ordenadas, D2 es el vector más próximo al eje de abscisas, y donde la consulta Q es el vector entre D1 y D2:

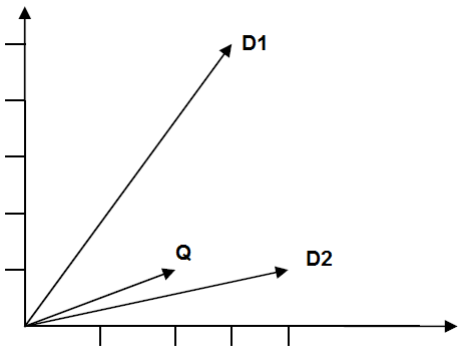

Figure 2: Ejemplo.

Como podemos observar en el gráfico (ver Figura 2), resulta relativamente fácil juzgar cuál de los dos documentos se asemeja más a la consulta. Considerando que el vector de la consulta Q está más próximo a D2, podemos deducir gráficamente que el orden de relevancia de los documentos D1 y D2 con relación a la consulta Q sería en este ejemplo: D2 y posteriormente D1. Sin la ayuda de un gráfico bastaría fijar un criterio de similaridad para poder ordenar fácilmente por orden de relevancia los documentos de una colección con relación a una consulta.

Las ventajas del modelo son:

- Es un modelo simple basado en el álgebra lineal.
- El peso de los términos no es binario.
- Permite aciertos parciales, ya que un documento puede ser considerado relevante, aunque no incluya todos los términos de la consulta.
- La ordenación de los resultados se realiza en base a varios factores: frecuencia de los términos, importancia de los términos y sin primar a los documentos más largos.
- Permite una implementación eficiente para grandes colecciones de documentos.

Las desventajas del modelo son:

- Se pierde parte de la información sintáctica y semántica del documento ya que se basa en la independencia de los términos dentro de un documento.
- Los documentos largos quedan poco representados ya que contienen pocos valores en común
- El orden en el cual los términos aparecen en el documento se pierde en la representación espacio vectorial.

- Las palabras de búsqueda deben coincidir con las palabras del documento, partes de una palabra pueden dar en falsos positivos.
- Sensibilidad semántica, documentos con contextos similares, pero con diferente vocabulario no serán asociados, resultando ser falsos negativos.

Las desventajas pueden ser solucionadas aplicando técnicas matemáticas, como descomposición de valor singular y bases de datos léxicas.

3 Medidas de cálculo de relevancia

Como se ha comentado anteriormente en la sección 2.3 las medidas de cálculo de relevancia típicamente usadas para la generación de los vectores son el factor término-frecuencia (TF) y el factor de frecuencia inversa del documento, inverse document frequency (IDF). Que se usan para ponderar los términos en los documentos de la colección y que consiste en multiplicar estos factores que reflejan la importancia de los términos.

El primer factor, TF (abreviatura de Term Frequency), pretende reflejar la importancia de los términos en los documentos, concediendo mayor importancia a los términos cuantas más veces aparezcan en los documentos. La versión más sencilla de este factor lo representa numéricamente mediante la frecuencia de aparición de cada término en cada documento de la colección. A mayor frecuencia de un término en un documento, mayor importancia.

$$\mathrm{TF}_{d,i} = \frac{f_i}{max_j f_{d,j}}, \text{ donde } f_{d,j} \text{ es la frecuencia de } t_j \text{ en } d$$

El segundo factor, IDF (abreviatura de Inverse Document Frequency), o inverso de la frecuencia de documentos, pretende reflejar la importancia de los términos en la colección, primando la precisión y el poder discriminatorio de los mismos. Así, dará mayor importancia a un término cuanto menor sea el número de documentos de la colección en los que aparezca dicho término. Por el contrario, si un término aparece en todos los documentos de la colección, su precisión y poder discriminatorio (capacidad para discernir los documentos relevantes de los irrelevantes ante una consulta) es nulo (tal término aparecerá necesariamente tanto en todos los documentos relevantes como en todos los documentos irrelevantes), de manera que se le otorgará una importancia mínima en esa colección en concreto (puede que en otra colección ese mismo término posea una gran importancia, porque aparece en muy pocos documentos). Suele representarse numéricamente de manera proporcional al logaritmo neperiano del inverso del número de documentos de la colección en los que aparece dicho término.

Los términos más infrecuentes en la colección son más importantes, pues discriminan antes. Definimos la *"rareza"* de un término como su frecuencia inversa documental, o IDF:

$$\mathrm{IDF}_i = log_2 \frac{D}{d_{f,i}},$$
donde D es igual al número de documentos y
$d_{f,i}$ es igual al número de documentos que contienen el termino t_i

4 Cálculo de similitud entre vectores de un MEV

El modelo vectorial propone evaluar el grado de similitud entre los documentos de una colección y las consultas mediante criterios que muestran la mayor o menor cercanía entre los vectores

correspondientes a los documentos y el vector correspondiente a la consulta.

Una de las maneras más habituales de cuantificar el nivel de cercanía entre vectores es mediante el coseno del ángulo que forman, pues presenta la propiedad de ser un número mayor cuanto más cercanos estén entre sí ambos vectores, mientras que es un número menor cuanto más alejados estén entre sí.

A continuación, se muestran los criterios para evaluar el grado de similitud entre vectores de un MEV más conocidos.

4.1 Mediante el producto escalar

Como veremos, existen muchas modalidades de comparación o equiparación mediante el grado de similitud. Una de las más sencillas por su simplicidad y sistematización inmediata es mediante el producto escalar de los pesos.

En el cálculo de similitud mediante el producto escalar, la similitud de un documento y una consulta, seran igual a la suma de los productos de sus pesos, sin olvidar que cada peso representa a un término. Este método puede aplicarse tanto a pesos binarios como a pesos TF-IDF.

4.1.1 Modalidad de pesos binarios

En el caso de la modalidad de pesos binarios (ver Figura 3), la similitud de un documento con respecto a la consulta es equivalente a la presencia de los términos de la consulta en el documento. 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 tanto en la consulta como en el documento siempre tendrá valor 1. Por ello sólo bastará contabilizar el número de términos coincidentes de la consulta en el documento y ése será su valor de similitud.

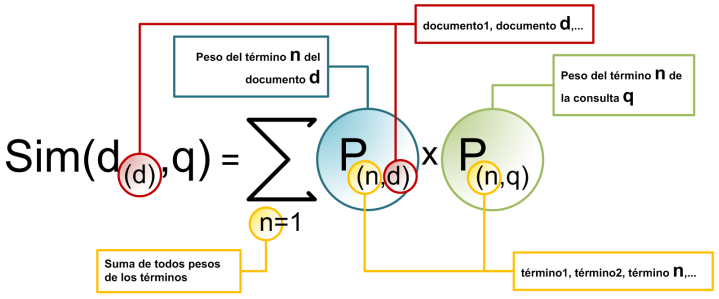

Figure 3: Similitud de un documento d y la consulta q mediante producto escalar.

4.1.2 Modalidad de pesos TF-IDF

En el caso de la modalidad anterior de pesos binarios, las limitaciones en la definición de la representatividad de los términos de cada documento son patentes. Por tanto, es un resultado bastante limitado y parcial. Por ello el método de la similitud mediante el producto escalar se aplica habitualmente con pesos TF-IDF, mucho más precisos.

El cálculo de la similitud se aplica a cada uno de los documentos de la colección. Por ejemplo, para el primer documento la similitud con respecto a la consulta de un usuario q, será diferente respecto al segundo documento. Al igual que ocurría 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, siendo estos mucho más precisos que un simple número entero.

4.2 Mediante la fórmula del coseno

Mediante el 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, ver Figura 4.

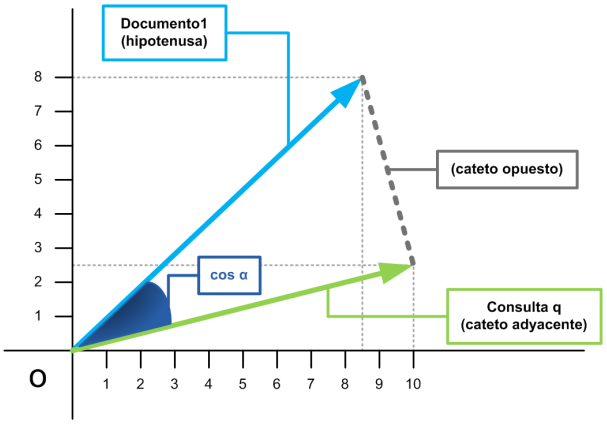

Figure 4: El ángulo del coseno.

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 (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 Figura 4, se comprueba visualmente cierta distancia del vector de la consulta con respecto al Documento1, cuando ambos vectores se muestran tan próximos como para superponerse, implica que el ángulo que forman sea menor y que su nivel de coincidencia sea superior. De hecho, el coseno de alfa de un triángulo cualquiera siempre es igual al cateto adyacente entre la hipotenusa y un coseno de 0^{o} implica una similitud máxima.

Por tanto, la fórmula aplicada para calcular el coeficiente de similitud del coseno (ver Figura5) entre un documento y una consulta es aquella que permite poner en relación los pesos que forman los vectores del documento y la consulta. 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 fácil de calcular.

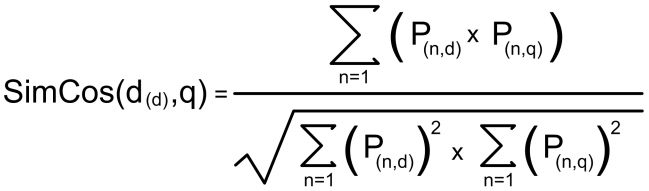

Figure 5: Fórmula para el cálculo de la similitud del coseno.

4.3 Mediante el coeficiente de Dice

El cálculo del coeficiente de similitud según Lee Raymond Dice es una adaptación del cálculo del coeficiente del coseno. La diferencia en la formulación reside en que la cardinalidad del numerador es dos veces la información compartida y el denominador la suma de los pesos al cuadrado del documento y su consulta (ver Figura 6).

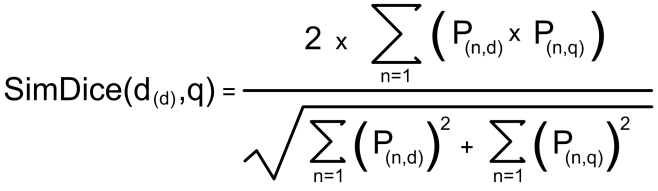

Figure 6: Fórmula para el cálculo del coeficiente de similitud de Dice.

4.4 Mediante el coeficiente de Jaccard

El cálculo del coeficiente de similitud de Jaccard al igual que el de Dice, son una adaptación del coeficiente de similitud del coseno.

La aplicación del coeficiente de similitud de Jaccard, centrada en usos estadísticos, mide la similitud entre conjuntos. Se puede definir como el tamaño de la intersección (numerador) dividido por el tamaño de la unión de la muestra, en este caso la suma de los pesos al cuadrado del documento y la consulta menos la intersección (ver Figura 7).

Una vez calculada la similitud entre cada documento de la colección y la consulta, el modelo vectorial es capaz de ordenar todos los documentos de la colección en orden decreciente de su grado de similitud con la consulta, incorporando a los resultados aquellos documentos que satisfacen sólo parcialmente los términos de la consulta. Se efectúa, en consecuencia, equiparación parcial.

5 Espacios vectoriales basados en *word embeddings*

Espacios vectoriales basados en representaciones continuas o densas de palabras (*word embeddings*)

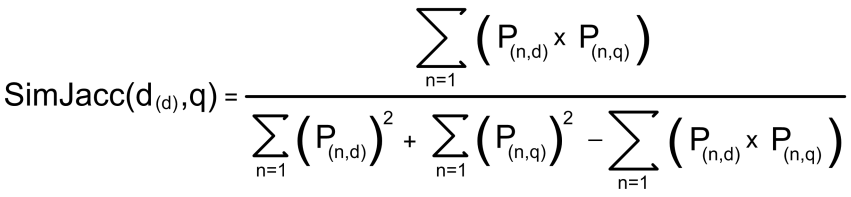

Figure 7: Fórmula para el cálculo del coeficiente de similitud de Jaccard.

6 Aplicaciones de *word embeddings*

Apache Lucene: motor de búsqueda textual de alto rendimiento escrito en Java. Weka : aplicación popular para realizar minería de datos. Karpanta: motor de búsqueda documental, adopta el conocido modelo del espacio vectorial. se apoya en dos módulos: uno de indización, que construye los vectores de los documentos, y otro de consulta, que calcula la similaridad con una consulta dada Sistemas antiplagio. Se utiliza para comparar la similitud entre documentos.A cada documento se le asocia un vector. Sistemas de recomendación: cálculo de similaridad entre vectores (función vectorial del coseno).

7 Reflexión personal sobre la utilidad y futuro de los MEV

Reflexión personal sobre la utilidad y futuro de los MEV para la representación de la semántica, de la información y la generación de conocimiento.

Actualmente la recuperación de información ha cobrado un gran auge debido al crecimiento espectacular de Internet, tratando de facilitar la tarea de discernimiento de los escasos documentos relevantes que puedan existir en la red frente a los millones de documentos irrelevantes en relación a cada consulta formulada en la red. Dado que esta inmensa "colección" carece por completo de organización, la automatización de los procesos de análisis y recuperación de los billones de documentos que configuran la red se ha convertido en una tarea de importancia capital.

Los programas que rastrean la web en busca de páginas y los programas que efectúan el proceso de análisis y tratamiento de tales páginas con el objeto de poder recuperarlas ante las consultas de los usuarios, además de muchos otros programas con un objetivo semejante en cualquier ámbito (desde las bibliotecas hasta el comercio electrónico), se siguen basando en los tres modelos clásicos de recuperación de información creados entre los años sesenta y ochenta del siglo XX: los modelos booleano, probabilístico y vectorial.

Como hemos podido observar, el fenómeno más destacado actualmente en estos sistemas de recuperación de información consiste en el empleo simultáneo de características y algoritmos propios de cada uno de estos modelos. Así, lo más frecuente es que los buscadores de Internet se basen en el modelo booleano, pero efectúen la ordenación de los documentos de las respuestas empleando criterios de similaridad originarios del modelo vectorial clásico. De igual modo, cada vez en mayor medida los SRI emplean una u otra variante de la retroalimentación por relevancia para aumentar la precisión de la respuesta, técnica empleada en sus inicios por el modelo probabilístico. En consecuencia, puede afirmarse que con la popularización de Internet han cobrado nuevo auge los modelos clásicos de recuperación de información, tratando de aunar en un mismo programa de recuperación las ventajas primordiales de cada uno de ellos. La investigación en este área, muy activa en la actualidad, sigue tratando de mejorar la precisión y exhaustividad de los sistemas,

pero tratando ahora de incorporar el usuario real y su punto de vista subjetivo en la evaluación de los sistemas. Sin duda en un futuro –esperemos que no muy lejano- tales avances se incorporarán a los SRI en beneficio de un acceso rápido y eficaz a la información por parte de cualquier habitante de nuestro planeta.