3 Modelo de espacios vectoriales

Muchas de las tareas de recuperación de información como la búsqueda, agrupamiento o categorización de textos tienen como primer objetivo procesar documentos en lenguaje natural. El problema que surge es que los algoritmos que pretenden resolver estas tareas necesitan representaciones internas explícitas de los documentos. En el área de recuperación de información normalmente se usa una expresión vectorial, donde las dimensiones del vector representan términos, frases o conceptos que aparecen en el documento. En este aspecto la representación más adoptada es la conocida como bolsa de palabras: una colección de documentos compuesta por *n* documentos indexados y *m* términos representados por una matriz documento-término de *n* x *m*. Donde los *n* vectores renglón representan los *n* documentos; y el valor asignado a cada componente refleja la importancia o frecuencia ponderada que produce el término, frase o concepto t_i en la representación semántica del documento *j*.

$$d_j = (w_{1j}, w_{2j} \dots w_{mj})$$

Donde *m* es la cardinalidad del diccionario[10] y $0 \leq w_{ij} \leq 1$ representa la contribución del termino t_i para la representación semántica del documento d_j.

En esta representación vectorial de documentos el éxito o fracaso se basa en la ponderación o peso de los términos. Aunque ha habido mucha investigación sobre técnicas de ponderación de términos, en realidad no hay un consenso sobre cuál método es el mejor [2]. También hay que destacar que el espacio de renglones de la matriz documento-término determinan el contenido semántico de la colección de documentos. Sin embargo, una combinación lineal de dos vectores-documento no representa necesariamente un documento viable de la colección. Más importante aún, mediante el modelo espacio vectorial se pueden explotar las relaciones geométricas entre dos vectores documento (y términos) a fin de expresar las similitudes y diferencias entre términos.

Si bien el rendimiento de un sistema de recuperación de información depende en gran medida de las medidas de similitud entre documentos, la ponderación de términos desempeña un papel fundamental para que esa similitud entre documentos sea más confiable. Así, por ejemplo, mientras que una representación de documentos basada solo en las frecuencias o apariciones de términos no es capaz de representar adecuadamente el contenido semántico de los documentos [1], la representación de términos ponderados[11] hace frente a errores o incertidumbres asociadas a la representación simple de documentos.

[10] Un diccionario es una lista de términos únicos que aparecen en un conjunto de documentos.

[11] Aplicación de métodos de normalización a la matriz documento-término.

3.1 Construcción

Una colección de *n* documentos indexados por *m* términos puede ser representada por una matriz A de dimensión *n* x *m*, donde cada elemento a_{ij} es usualmente definido por una frecuencia ponderada del término *i* en el documento *j* cuyo objetivo principal es mejorar el rendimiento en la recuperación de información; entendiendo como rendimiento la habilidad de recuperar información relevante y descartar información irrelevante. La siguiente figura (ver figura 1) muestra una matriz documento-término simple, donde cada columna representa un término en la colección, cada renglón un documento y cada celda o elemento de la matriz la ocurrencia del término en el documento.

	Término 1	Término 2	Término 3
Documento 1	**1**	**0**	**0**
Documento 2	**0**	**0**	**1**
Documento 3	**1**	**1**	**1**
Documento 4	**0**	**1**	**0**

Figura 1 – Matriz documento-termino simple

En ella podemos ver que el término 1 aparece en el documento 1 y 3, pero no en los otros dos documentos. Se demuestra así que cada renglón de la matriz de 4×3 puede ser representado en un espacio de tres dimensiones.

Siguiendo la nomenclatura usada en [1], cada elemento a_{ij} de la matriz documento-termino 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 el peso global del término *i* en la colección de documentos y $d_j$es el factor de normalización para el *j*-ésimo documento. Los siguientes apartados contienen las fórmulas más populares usadas en sistemas de indexado automático. Usualmente los componentes principales son el factor *término-frecuencia* (TF) y el factor de frecuencia inversa del documento, *inverse document frequency* (IDF).

3.2 Peso local

El peso local mide la importancia del término *i* en el documento *j* y sólo depende de las frecuencias en el documento y no de otros documentos.

Nombre	Fórmula para l_{ij}
Binaria	$x(f_{ij})$
Frecuencia del término	f_{ij}
Frecuencia aumentada de términos normalizados	$K * x(f_{ij}) + (1 - K) * \frac{f_{ij}}{max_k(f_{kj})}$
Logaritmo	$\log(f_{ij} + 1)$
Logaritmo alternativo	$x(f_{ij}) * (\log(f_{ij}) + 1)$

Tabla 1 – Fórmulas de pesos locales

3.2.1 Binaria

Da a cada palabra del mismo documento la misma importancia, siendo especialmente útil cuando el número de veces que aparece la palabra no se considera importante [1] [2] [28].

$$x(t) = \begin{cases} 1 & si\ t > 0 \\ 0 & si\ t = 0 \end{cases}$$

3.2.2 Frecuencia del término

Esta fórmula hace un conteo de los términos en el documento, cuantas más veces se produce un término *t* en el documento *j* es más probable que *t* sea relevante para el documento. Es usado principalmente para favorecer a palabras comunes y documentos largos, aunque, por ejemplo, una palabra que aparece diez veces en un documento no quiere decir que sea diez veces más importante que una palabra que aparece una sola vez [1] [2] [28].

Los pesos binarios y frecuencia del término son utilizados frecuentemente en consultas donde los términos aparecen una o dos veces; ninguno de estos dos es mejor que el otro. Por un lado un peso binario no hace diferencia entre los términos que aparecen con frecuencia y por el otro lado el peso frecuencia del término da demasiada importancia a términos que aparecen muy a menudo.

3.2.3 Frecuencia aumentada de términos normalizados

Esta fórmula hace un conteo de los términos que aparecen en el documento y al mismo tiempo da importancia adicional a términos que aparecen con frecuencia. Esta fórmula fue propuesta por [3] y está parametrizada por una variable *K* que toma valores inferiores a 0.5. En este aspecto [3] sugirió que se debe establecer la variable *K* con un valor bajo (comúnmente 0.3) para documentos de gran tamaño y el valor más alto (0.5) para los documentos más breves [4]. Con esta fórmula (ver tabla 1), el valor de salida sólo varía entre 0.5 y 1 para los términos que aparecen en el documento. Al limitar los factores

de *TF* a un valor máximo de 1.0, esta técnica sólo compensa el problema de la presencia de la frecuencia alta de términos para la normalización. Esta fórmula es especialmente útil cuando no se usa con alguna otra fórmula de normalización en documentos muy extensos [1] [2] [28].

3.2.4 Logarítmico

Algunos autores proponen a los logaritmos como los mejores porque minimizan el efecto de la frecuencia [1] [2] [28], partiendo del planteamiento de que un término que aparece diez veces en un documento no necesariamente es diez veces más importante que un término que aparece una vez en ese documento. La literatura propone a la normalización logaritmo y logaritmo alternativo como los más utilizados [28].

3.3 Peso global

Estas fórmulas son utilizadas para medir la importancia de un término y se basan en la dispersión de un término en particular en todos los documentos. En general, el peso global asocia valores más bajos a términos que aparecen muy frecuentemente o en muchos documentos y es por esta razón que en teoría puede eliminar la necesidad de usar una lista de palabras de paro (una lista de palabras funcionales: artículos, preposiciones, verbos auxiliares, etc) para eliminar las palabras más comunes en el lenguaje, aunque en la práctica este proceso no es complicado y se hace en la etapa de pre-procesamiento del texto [1] [2].

Nombre	**Fórmula para g_i**
Sin cambios	1
Frecuencia inversa de documentos (IDF)	$\log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)$
Frecuencia cuadrática inversa de documentos	$\log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)^2$
Frecuencia probabilística inversa de documentos	$\log\left(\frac{\mathrm{n}-\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)^2$
Frecuencia global inversa GFIDF	$\frac{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{f_{ik}}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}$
Entropía	$1+\sum_{j=1}^{n} \frac{p_{ij} * \log(p_{ij})}{\log(n)}, p_{ij} = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}$

Tabla 2 – Fórmulas de pesos globales

3.3.1 Sin cambios

Útil cuando solo se desea considerar la frecuencia de los términos, cuando las frecuencias de los términos son muy pequeñas o cuando nos interesa destacar la frecuencia de los términos en un documento [1] [2] [28].

3.3.2 Frecuencia inversa de documentos (IDF)

Frecuencia inversa de documentos (IDF, por sus siglas en inglés) es una medida muy popular para medir la importancia de una palabra [2]. Definido como el logaritmo de la relación del número total de documentos *n* y el número de documentos *k* que contienen un término dado [1] [2] [28]. De esta manera las palabras raras o poco frecuentes tienen un valor IDF alto, mientras que las más comunes, el IDF más bajo. Por ejemplo, para obtener un valor de 0 el término debe aparecer en cada documento y para que aumente el valor IDF el número de documentos donde aparece un término debe disminuir con respecto al número total de documentos. He aquí un ejemplo para una colección de 10 000 documentos:

$$\log\left(\frac{10\,000}{10\,000}\right) = 0\,; \quad \log\left(\frac{10\,000}{20}\right) = 2.698\,; \quad \log\left(\frac{10\,000}{1}\right) = 4$$

3.3.3 Otros esquemas IDF

3.3.3.1 Frecuencia cuadrática inversa de documentos

Se utiliza raramente como una variante del IDF, y no es más que la fórmula de IDF elevada al cuadrado [1].

3.3.3.2 Frecuencia probabilística inversa de documentos

Asigna pesos que van desde $-\infty$ para términos que aparecen en más de un documento y $\log(n-1)$ para términos que aparecen en un solo documento [1] [2].

3.3.3.3 GFIDF

En esta fórmula se calcula la relación del número total de apariciones de un término en el documento con el número de documentos en el que aparece [1] [2] [28].

3.3.4 Entropía normalizada

La entropía se basa en las ideas de la teoría de la información y es el esquema de ponderación más sofisticado. En esta medida se asignan pesos entre 0 y 1. Si un término aparece una vez en cada documento de la colección, entonces a ese término se le da una ponderación de 0. Si un término aparece una vez en un solo documento de la colección, entonces a ese término se le asigna un peso de 1. Cualquier otra combinación de las frecuencias dará un peso en algún punto entre 0 y 1. La entropía es un peso útil porque da mayor peso a los términos que aparecen menos veces en un pequeño número de documentos. Así que esta fórmula toma en cuenta la distribución de los términos sobre los documentos [1] [2] [28].

3.4 Normalización

El tercer componente del esquema de pesos es el factor de normalización, el cual se utiliza para corregir las discrepancias en la longitud de los documentos, para que de esta manera sean recuperados con independencia de su longitud. Ver tabla 3.

Dos razones principales por lo que se requiere el uso de la normalización de pesos son:

- **Frecuencias de términos muy altas:** los documentos largos suelen utilizar los mismos términos en repetidas ocasiones y como resultado los factores término-frecuencia pueden ser grandes para documentos muy extensos [28].

- **Número de términos:** los documentos extensos tienen una gran cantidad de términos y esto aumenta las coincidencias entre una consulta y un documento largo, aumentando las posibilidades de recuperación de documentos largos en preferencia sobre los documentos más breves [28].

Nombre	Fórmula para d_j
Sin cambios	1
Normalización de coseno	$\sqrt{\sum_{k=1}^{n} (g_k * l_{kj})^2}$
Suma de pesos	$\sum_{k=1}^{n} (g_k * l_{kj})$
Cuarta normalización	$\sum_{k=1}^{n} (g_k * l_{kj})^4$
Normalización del peso máximo	$max_{k=1}^{n}(g_k * l_{kj})$
Normalización de pivote único	$\frac{1}{(1 - pendiente) * pivote + (pendiente * \zeta_j)}$

Tabla 3 – Fórmulas de normalización

3.4.1 Sin cambios

Se utiliza cuando se quiere dar énfasis a los documentos largos sobre otros más cortos. Algunas veces es usada en conjunto con la normalización término frecuencia aumentada como peso local.

3.4.2 Normalización de coseno

La normalización coseno cubre las dos necesidades por las cuales es importante implementar la normalización (frecuencias de términos muy altas y el gran número de términos), siendo el método de normalización más popular. Su objetivo es favorecer a documentos pequeños con términos de pesos pequeños contra documentos grandes.

3.4.3 Otros esquemas de normalización de coseno

Tanto la suma de pesos como la cuarta normalización son una variante de la normalización de coseno y rara vez son utilizados [5].

3.4.4 Normalización del peso máximo

No es una normalización real. Aunque asigna pesos entre 0 y 1, esta fórmula no toma en cuenta la distribución de los términos sobre los documentos. Es útil cuando queremos dar importancia a los términos ponderados más relevantes dentro de un documento.

3.4.5 Normalización de pivote único

El problema de la normalización de coseno es que a menudo toma valores muy altos. Cuanto mayor es el valor de normalización, menor es la probabilidad de recuperar ese documento. La normalización de pivote único trata de corregir este problema favoreciendo a los documentos más cortos [28].

En la formula, ζ_j es el número de términos distintos en el documento j y gracias a la sugerencia de [28] la pendiente se encuentra en 0.2 y el pivote se establece como el promedio de términos distintos por documento en una colección.

3.5 Métodos más utilizados

Salton y Buckley [7] confirman que el método más utilizado de ponderación de término documento se obtiene con el producto de las funciones frecuencia del término (*tf*) y la frecuencia inversa de documentos (*idf*) con una normalización coseno.

$$a_{ij} = \frac{f_{ij} * \log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)}{\sqrt{\sum_{k=1}^{n}\left(f_{ij} * \log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)\right)^2}}$$

Salton y Buckley también proponen la combinación Frecuencia aumentada de los términos normalizados en conjunto con IDF y normalizados por coseno como el mejor esquema de ponderación de términos.

$$a_{ij} = \frac{\left(0.5 * x(f_{ij}) + (0.5) * \frac{f_{ij}}{max_k(f_{kj})}\right) * \log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)}{\sqrt{\sum_{k=1}^{n}\left(\left(0.5 * x(f_{ij}) + (0.5) * \frac{f_{ij}}{max_k(f_{kj})}\right) * \log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)\right)^2}}$$

Por el contrario, Basilio [8] asegura que da mejores resultados la combinación de peso local logaritmo y entropía como peso global.

$$a_{ij} = \log(f_{ij} + 1) * \left(1 + \sum_{j=1}^{n} \frac{\left(\frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}\right) * \log\left(\frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}\right)}{\log(n)}\right)$$

Singal [28] propone los pesos de frecuencia del término y frecuencia inversa de documentos normalizados utilizando la normalización de pivote único.

$$a_{ij} = \frac{f_{ij} * \log\left(\frac{\mathrm{n}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{X}(\mathrm{f_{ik}})}\right)}{(1 - 0.2) * pivote + (0.2 * \zeta_j)}$$

Donde ζ_j es el número de términos distintos en el documento *j*, el pivote es el promedio de términos distintos por documento en una colección y la pendiente toma el valor de 0.2. [28] mostró la debilidad de la función coseno como método de normalización en documentos muy largos y propone la normalización de pivote único con estos valores para la mayoría de las aplicaciones.