Descubrimiento de relaciones paradigmáticas

Salvador López Mendoza

Mayo de 2018

Contexto de palabras

Se define un *seudo-documento*.

Al considerar las siguientes oraciones:

Mi gato come pez los sábados.
Su gato come pavo los martes.

Se tiene:

$$Izq(gato) = \{mi, su, el, un, \ldots\}$$

$$Der(gato) = \{come, es, tiene, \ldots\}$$

$$Todo(gato) = \{mi, su, el, un, come, es, tiene, pez, pavo, martes, \ldots\}$$

Seudo documento

Contexto es un seudo documento.

El seudo documento se maneja como una *bolsa de palabras*.

En una bolsa las palabras no tienen orden.

El contexto puede contener palabras que son adyacentes, pero también palabras que no lo son.

Similaridad de contextos

Medir la similaridad de contextos:

$$\begin{aligned}\text{Sim}(\text{perro, gato}) = &\text{ Sim}(\text{Izq}(\text{perro}), \text{Izq}(\text{gato})) \\ &+ \text{Sim}(\text{Der}(\text{perro}), \text{Der}(\text{gato})) \\ &\ldots \\ &+ \text{Sim}(\text{Todo}(\text{perro}), \text{Todo}(\text{gato}))\end{aligned}$$

Un valor alto en $Sim(palabra_1, palabra_2)$ quiere decir que $palabra_1$ y $palabra_2$ están *relacionadas paradigmáticamente*.

Modelo de espacio vectorial (VSM)

La bolsa de palabras se puede representar como un *espacio vectorial*.

La bolsa de palabras correspondiente a *gato* es el conjunto de palabras que aparecen en las oraciones que contienen la palabra *gato*.

Sea N la cantidad de palabras distintas (vocabulario).

d_1 representa a la bolsa de palabras correspondiente a *gato*.

$d_1 = (x_1, x_2, \ldots, x_N)$

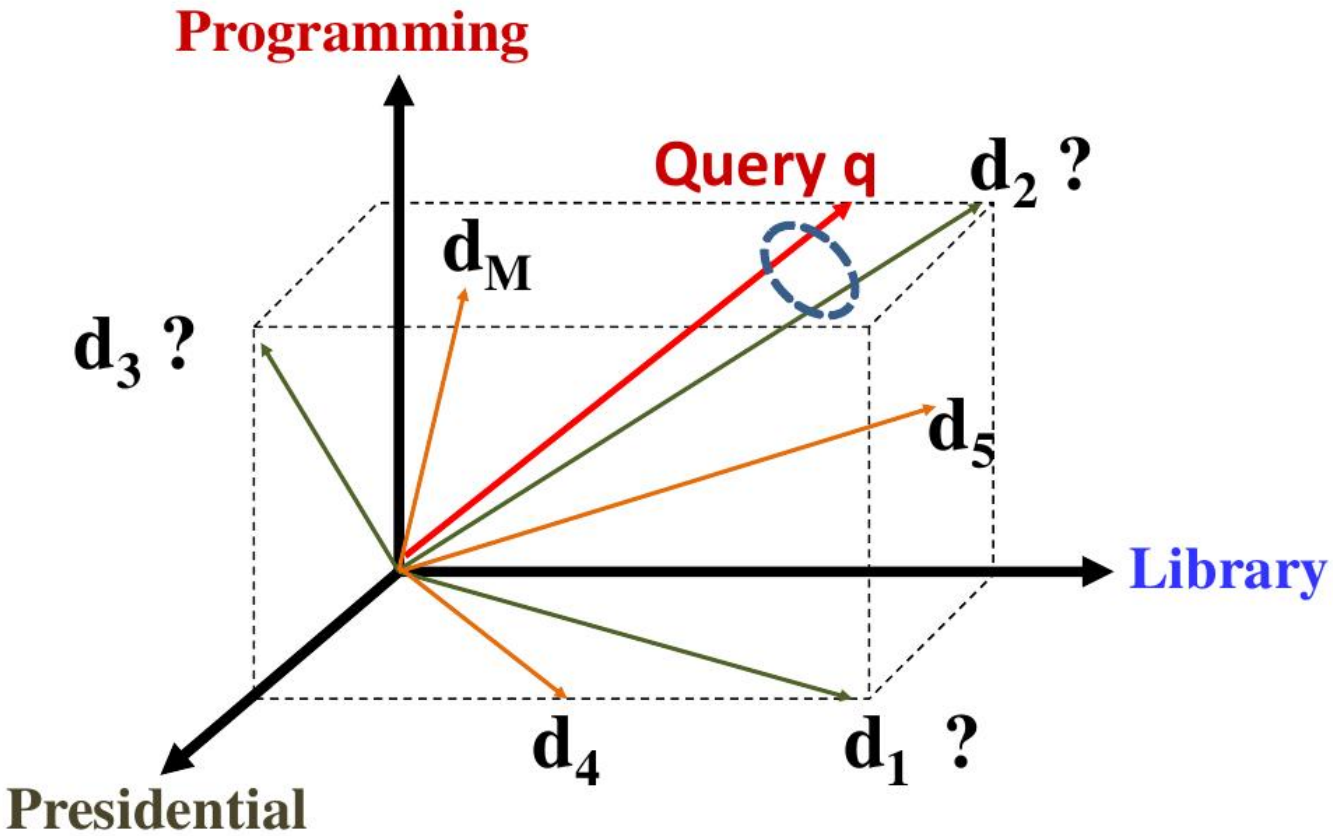
Cada x_i indica la cantidad de veces que aparece esa palabra en la bolsa de palabras.

De la misma forma se define d_2 para la bolsa de palabras de *perro*.

$d_2 = (y_1, y_2, \ldots, y_N)$

PEUVI

Modelo de espacio vectorial (II)



Similaridad entre documentos

¿Cómo se calcula la similaridad entre documentos?

$Sim(d_1, d_2) = ?$

Ejemplo:

Considerar **Su gato come pavo los martes**.

$d_1 = (1, 1, 1, 1, 1)$

¿Qué tan similar es a **Su perro come pavo los martes**?

$d_2 = (1, 1, 1, 1, 1)$

Comparar:

Su gato come pez los sábados con **Su perro come carne los domingos**.

Se necesita definir la función de similaridad.

¡Hay muchas posibilidades!

PEUVI

Calculando la similaridad

$Sim(d_1, d_2) = d_1 \cdot d_2$

$Sim(d_1, d_2) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$

$Sim(d_1, d_2) = \sum_{i=1}^{N} x_i y_i$

Para el ejemplo anterior:

$Sim(d_1, d_2) = 5$

¿Cuál es la similaridad de las otras dos frases?

PEUVI

Traslape esperado de palabras en el contexto (EOWC)

Intuición: mientras mayor cantidad de palabras se traslapen, la similitud es mayor.

Problemas:

- Se favorece a aquellos elementos que aparezcan con mayor frecuencia.
- Se trata a todas las palabras equitativamente.
  Las coincidencias en palabras como *los* no son tan importantes como las coincidencias en otras palabras, como *come*.

PEUVI

Medidas de similitud

Definiciones:

Sea $V = \{w_1, w_2, \ldots, w_N\}$ el conjunto de palabras (*vocabulario*).

$c(w_i, d)$ es la cantidad de ocasiones en que aparece la palabra w_i en el documento d.

$d_1 = (c(w_1, d_1), c(w_2, d_1), \ldots, c(w_N, d_1))$

Se normaliza el cálculo:

Sea $x_i = c(w_i, d_1)/|d_1|$ y $y_i = c(w_i, d_2)/|d_2|$

$Sim(d_1, d_2) = d_1 \cdot d_2$.

Notas:

x_i es la probabilidad de que al tomar al azar una palabra del documento d_1, la palabra sea w_i.

$Sim(d_1, d_2)$ es la probabilidad de que al tomar al azar una palabra de cada documento, éstas sean idénticas.

PEUVI

Recuperación de textos (TR)

- Existe una colección de textos.
- El usuario hace una consulta para expresar lo que requiere.
- El sistema (*máquina de búsqueda*) regresa los documentos relevantes.

PEUVI

TR vs. bases de datos

- Información.
  El texto no tiene estructura (es texto libre).
  La información es ambigua, no hay una semántica bien definida.
- Consulta.
  En TR es ambigua. En BD hay una semántica bien definida.
  La especificación puede ser incompleta.
- Respuesta.
  Documentos relevantes en TR. En BD son registros que coinciden.

PEUVI

Definición formal de TR

- Vocabulario. $V = \{w_1, w_2, \ldots, w_N\}$
- Consulta. $q = q_1, \ldots, q_m$, en donde $q_i \in V$
- Documento. $d_i = d_{i1}, \ldots, d_{imi}$, en donde $d_{ij} \in V$
- Colección. $C = \{d_1, \ldots, d_M\}$
- Conjunto de documentos relevantes. $R(q) \subset C$
  Es un conjunto desconocido. Depende del usuario.
  La consulta es una *sugerencia* de lo que debe contener un documento en $R(q)$.
- Tarea. Calcular $R'(q)$. Es una aproximación a $R(q)$.

Se usa una *función de clasificación* (ranking function).

Posibilidad: modelos basados en similitud.

PEUVI

Similitud en TR

- Vector de bits.
  $x_i, y_j \in \{0, 1\}$
  Al calcular $Sim(d_1, d_2)$ se cuenta la cantidad de palabras que coinciden.
- Vector de frecuencia de términos.
  Se debe dar mayor valor al hecho de que una palabra se repita varias veces.
  x_i = cantidad de ocasiones en que aparece w_i en la consulta.
  y_i = cantidad de ocasiones en que aparece w_i en el documento.
  Da más peso a las palabras que se repiten con frecuencia.
  Agregar *frecuencia inversa*.
  Ayuda a discriminar palabras que son frecuentes, pero no aportan contenido.
  Ahora $y_i = c(w_i, d) \times IDF(w_i)$

PEUVI

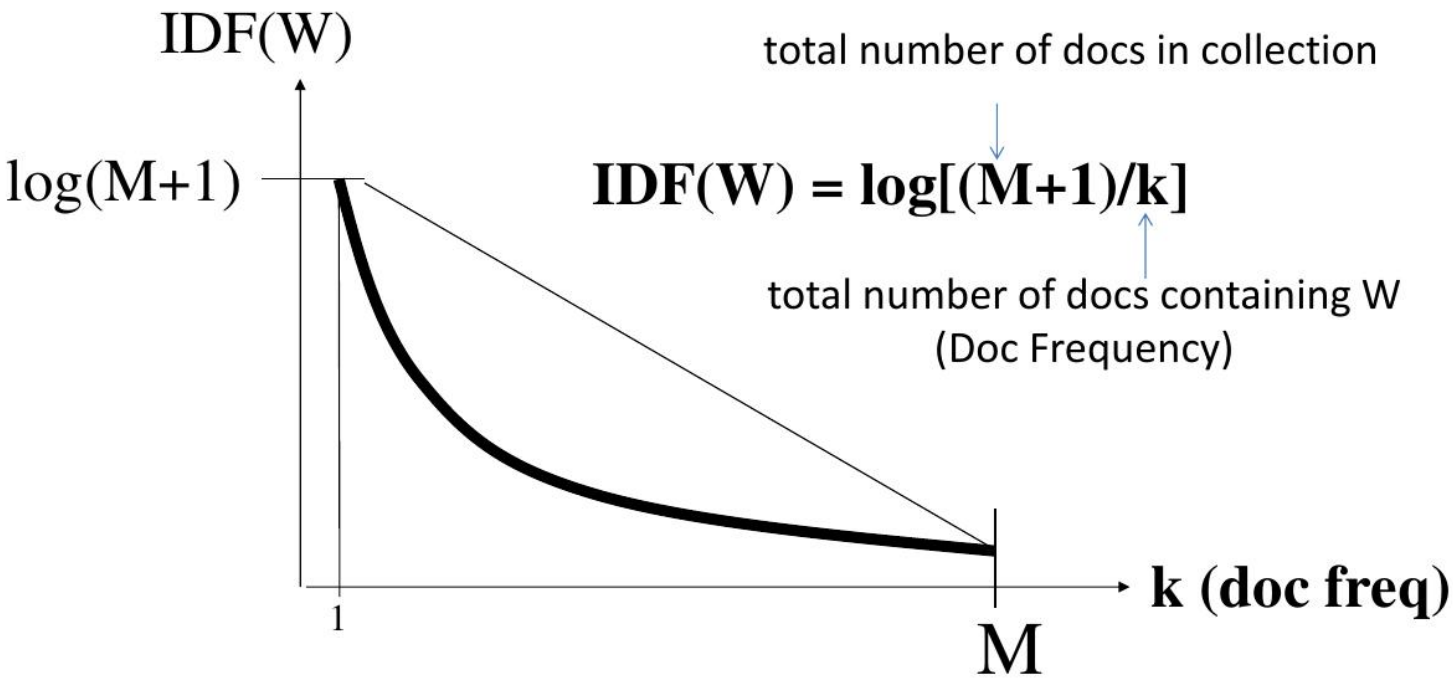
IDF

Penaliza la aparición más frecuente.

Sea M la cantidad de documentos en la colección.

Sea k la cantidad de documentos que contienen la palabra w.

Entonces, $IDF(w) = log[(M + 1)/k]$



IDF (II)

$f(q,d) = \sum_{i=1}^{N} x_i y_i = \sum_{w \in q \cap d} c(w,q) c(w,d) log(M + 1/df(w))$

En vez de usar la cantidad de veces que aparece una palabra en un documento, usar la frecuencia con la que aparece.

Transformación $BM25$:

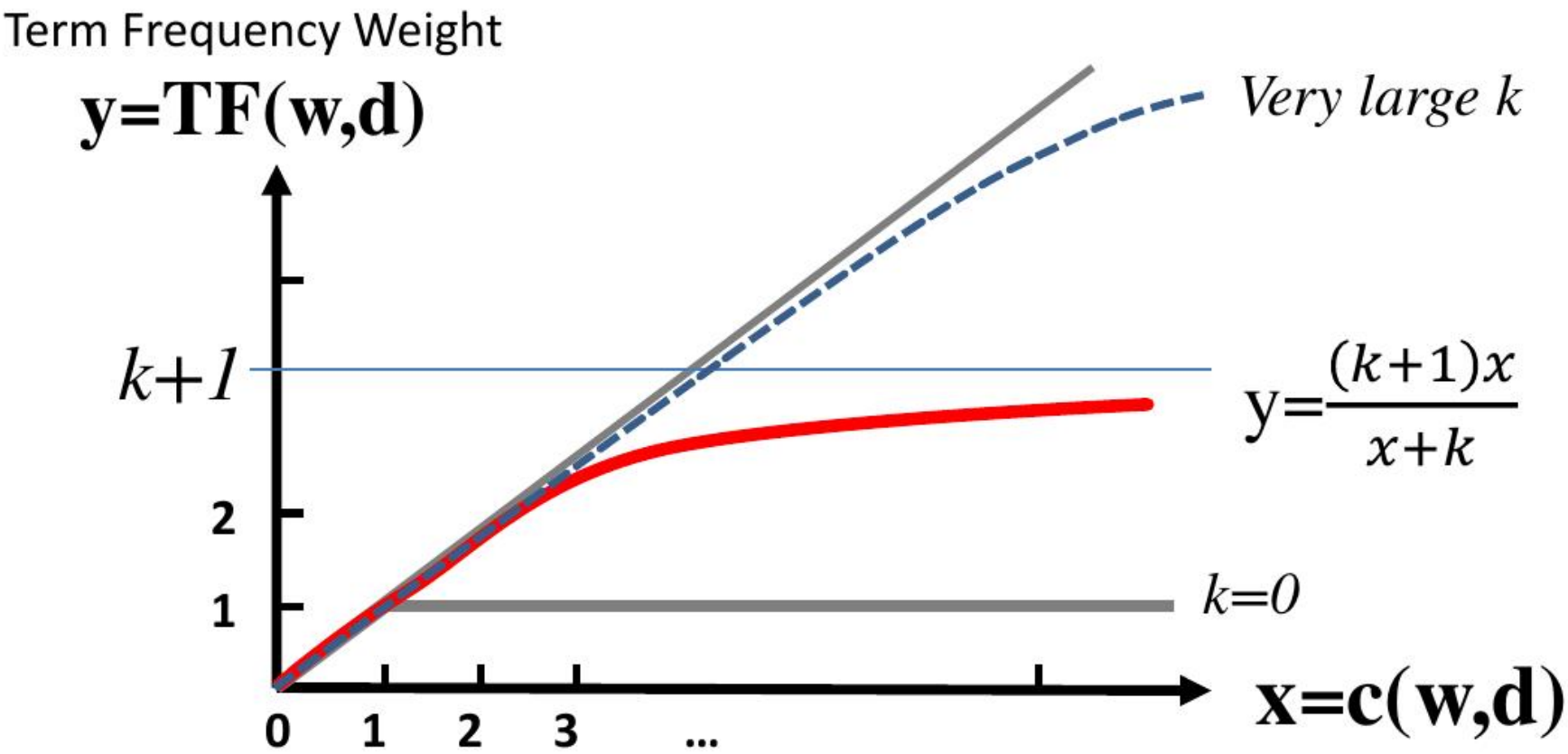
$$y = (k+1)x/x + k \text{ con } x = c(w,d)$$

Función de clasificación:

$$f(q,d) = \sum_{i=1}^{N} x_i y_i$$

$$f(q,d) = \sum_{w \in q \cap d} c(w,q)(k+1)c(w,d)/c(w,d) + klog(M+1)/df(w)$$

BM25


Term Frequency Weight
y=TF(w,d)
Very large k
k+1
y=(k+1)x / x+k
2
1
k=0
0
1
2
3
...
x=c(w,d)

Mejorando EPWC

Se pueden utilizar transformaciones sublineales de la frecuencia de términos (TF).

Hay que recompensar a las ocasiones en que coinciden en palabras *raras*.

Se usa IDF.

Se mejora mucho usando BM25.

PEUVI

BM25 y relaciones paradigmáticas

Sea $d_1 = (x_1, \ldots, x_N)$

$$BM25(w_i, d_1) = (k+1)(c(w_i, d_1)/(c(w_i, d_1) + k(1 - b + b * |d_1|/avd_1))$$

Con $b \in [0, 1]$ y $k \in [0, +\infty)$

Entonces $x_i = BM25(w_i, d_1)/\sum_{j=1}^{N} BM25(w_j, d_1)$

y_i se define de la misma manera.

$$Sim(d_1, d_2) = \sum_{i=1}^{N} IDF(w_i) x_i y_i$$

PEUVI