Universidad del País Vasco
Matemática Aplicada y Estadística

Álgebra lineal tras los buscadores de Internet

Juan-Miguel Gracia

Extracto: Se analizan dos aplicaciones del álgebra lineal a la construcción de buscadores de Internet: el valor asignado a cada página web por Google, y el análisis semántico latente.

juanmiguel.gracia@ehu.es

11 de noviembre de 2002 Versión final

Índice General

1. **Introducción**
   - 1.1. **Google**
   - 1.2. **Análisis semántico latente**
2. **¿Cómo medir la importancia de una página web?**
   - 2.1. **Modelos de navegación**
   - 2.2. **Nota media "verdadera"**
     - Planteamiento • Relación con la teoría de Perron y Frobenius
3. **Análisis semántico latente**
   - 3.1. **Buscadores**
   - 3.2. **Búsqueda semántica**
     - Universo semántico • Análisis semántico oculto
   - 3.3. **Descomposición de valores singulares**
   - 3.4. **Alternativa**
   - 3.5. **Búsquedas semánticas**
   - 3.6. **Palabras o términos**
   - 3.7. **Ejemplo**
     - Búsqueda literal

3.8. Software “on line”
4. Elaboración automática de tesauros
4.1. Polisemia de una palabra
5. Matemáticas suscitadas por Internet
6. Acreditaciones
6.1. Agradecimientos

1. Introducción

1.1. Google

Al ver cómo los correctores ortográficos detectaban palabras "vecinas" para remediar errores, o cómo buscaban sinónimos, nos preguntábamos cuál sería el algoritmo que buscaba ese "entorno" de cada palabra, correcta o no.

Al aparecer Internet en nuestras vidas, nos trajo los buscadores (Yahoo, Altavista, Lycos, etc.). Éstos escrutaban la Red y, veloces como el rayo, nos devolvían ordenados enlaces a ficheros `html, htm` que podían estar relacionados con nuestra consulta. Extrañamente, muchas veces acertaban con nuestros deseos e incluso nos sorprendían gratamente. La cuestión de cómo lo hacían volvía a bullir en nuestras mentes. De hecho, al convertirse la Red en un fenómeno de masas, se acrecentó el interés por estas cuestiones; interés reflejado en numerosas publicaciones.

En estas cavilaciones estábamos cuando un compañero nos habló del buscador **Google**. Era inimaginable: ¡Qué rapidez! ¡Qué puntería! ¿Cómo lo harían? Además, este buscador encontraba también enlaces a ficheros `ps, pdf, rtf, doc, asp, txt,` etc. Entre otros procedimientos concu-

rrentes, Google utiliza una función que asigna un valor numérico a cada página web: su "PageRank". Este valor es asignado sin intervención humana y mide la calidad e importancia de la página. ¿Cómo lo hace? De una manera democrática. Cada enlace a una página es un voto para ésta. Google también analiza la página que emite el voto y, cuanto más importante sea, más importante le hará a la página destino del enlace. Véase [9]. De hecho, este método recuerda un método para evaluar la cualificación de un conjunto de empleados para realizar una tarea, basado en las opiniones de todos sobre todos al respecto; opiniones que son expresadas a través de una calificación numérica. También aquí es claro que se deben ponderar más las calificaciones emitidas por los que reciban mejor nota media. Véase [2]. El tema de esta subsección está desarrollado en la Sección 2.

1.2. Análisis semántico latente

Muchos métodos de búsqueda de textos en las páginas web de Internet dependen de un emparejamiento de palabras "al pie de la letra" entre las palabras que busca el usuario y las que existen en las páginas web. Se plantean dos situaciones antitéticas: por un lado, el buscador encuentra

términos homónimos y, por otro, los sinónimos le pasan desapercibidos. En el primer caso, son recuperadas páginas cuyo tema no nos interesa; en el segundo, son ignoradas páginas en las que estaríamos interesados.

Dado que los ordenadores no saben nada de lenguas ni de idiomas, se trata de dotarlos de un método matemático que "aprenda" el significado de las palabras que "viven" en el universo semántico constituido por todas las páginas web que existen en un momento dado, dispersas por el mundo. Es posible lograr este aprendizaje mediante el *análisis semántico latente* (en inglés "latent semantic index" o LSI) que desvela el significado oculto dado a las palabras en las páginas de la Red.

Supongamos que el número total de páginas web es n y que hay m palabras significativas si quitamos artículos, pronombres, adverbios, preposiciones, conjunciones, exclamaciones, interjecciones, etc. Podemos imaginar que toda la información relevante de una base de datos de un buscador de Internet está recogida en la matriz de incidencia $m \times n$, $A = (a_{ij})$ en la que a_{ij} denota el número de veces que la palabra i aparece en la página j. La descomposición de valores singulares de la matriz A permite recuperar información basada en conceptos o significados que están escondidos las páginas web. Un desarrollo de estas ideas se podrá leer en la Sección 3.

Creemos que ambos asuntos, Google y el Análisis Semántico, pueden ser de utilidad a los profesores de Álgebra Lineal para motivar a sus alumnos internautas.

2. ¿Cómo medir la importancia de una página web?

Un método para llevar a cabo la tarea de asignación de un número a cada página web que la valore es el llamado PageRank, [12], como ya dijimos en la introducción. Cada enlace de una página u con destino v le da un voto a v. Pero, si queremos tener en cuenta la importancia de la página u, la que vota, deberemos sopesar los votos que ella ha recibido de otras. En principio, para tratar de sustanciar esta idea, supongamos por ejemplo que una página u que es destino de 100 enlaces (retroenlaces) tiene 100 votos. Si ahora esta página tiene 2 enlaces hacia adelante (ultraenlaces) envía un valor de 50 votos a cada uno.

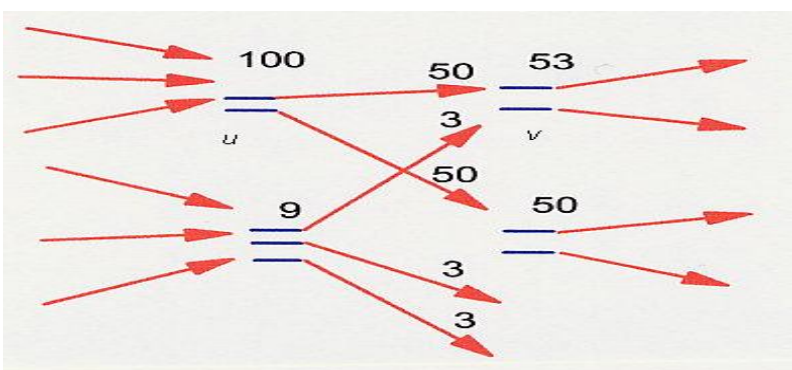


Figura 1: Votos de retro- y ultra- enlaces

Sea v uno de estos 2 destinos; si v recibe también 3 votos de otro retroenlace, obtiene un total de 53 votos. Si v, a su vez, tiene dos ultraenlaces, aporta a cada uno de sus destinos 53/2 votos, y así sucesivamente. Este ejemplo puede verse en la Figura 1.

Sea W el conjunto de todas las páginas web. Sea u una página web; sea F_u el conjunto de páginas web a las que enlaza la página u, y denotemos por B_u el conjunto de páginas que apuntan hacia u. Notemos por $N_u := |F_u|$ el cardinal de F_u. Una versión aproximada del PageRank es una función

$$R : W \to [0, \infty)$$

que satisface la ecuación funcional

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{N_v}, \tag{1}$$

para cada $u \in W$, junto con la condición normalizadora

$$\sum_{u \in W} R(u) = 1. \tag{2}$$

Como la suma total de los valores $R(u)$ es 1, es claro que hay que ponderar

los votos recibidos con un valor entre 0 y 1 en el ejemplo asociado a la Figura 1.

Sea n el número total de páginas web; llamando A a la matriz $n \times n$ cuyo término $a_{uv} := 1/N_v$ si existe un enlace de v a u y $a_{uv} := 0$ si no, se tiene que la ecuación (1) es equivalente la ecuación matricial donde el vector columna $n \times 1, r$ representa a R:

$$r = Ar, \tag{3}$$

con la condición normalizadora $\|r\|_1 = 1$. *Supongamos que toda página web tiene al menos un ultraenlace.* Como los elementos de la matriz A son números no negativos y las columnas de A suman 1, se tiene que la matriz A^{T} es estocástica, donde $^{\mathrm{T}}$ denota la traspuesta. Por lo tanto, el número 1 es valor propio de A; además, el radio espectral $\rho(A)$ es igual a 1. Recordamos que $\rho(A)$ se define como el máximo de los valores absolutos de los valores propios de A. Por la teoría de Perron-Frobenius de matrices no negativas, la matriz A tiene un vector propio no negativo x asociado al valor propio 1. Es claro que el vector

$$r := \frac{x}{\|x\|_1}$$

satisface la ecuación (3) y la condición $\|r\|_1 = 1$. Esto prueba la existencia de una función $R : W \to [0, \infty)$ que cumple las condiciones requeridas. Véase [8, pág. 547 y 543]. Pero ¿estará r (o R) unívocamente determinado? Si la matriz A es irreducible, la respuesta es afirmativa y, además, todos los elementos del vector r son positivos, [8, pág. 536, Theorem 1]. En toda circunstancia en la que A tenga a 1 como valor propio simple, el vector propio r tal que $\|r\|_1 = 1$, estará unívocamente determinado.

La recurrencia

$$p(k+1) = Ap(k), \quad k = 0, 1, 2, \ldots. \tag{4}$$

está asociada a la ecuación (3). La solución de esta ecuación en recurrencias es

$$p(k) = A^k p(0), \quad k = 0, 1, 2, \ldots;$$

si existe el $\lim_{k\to\infty} A^k$, se tendrá que el $\lim_{k\to\infty} p(k)$ existe cualquiera que sea el vector de condiciones iniciales $p(0)$ y el vector

$$\lim_{k\to\infty} p(k)$$

es una solución de la ecuación (3). La condición necesaria y suficiente para que exista el $\lim_{k\to\infty} A^k$ es que la matriz A sea propia. Ésta es la misma

condición para que exista el $\lim_{k\to\infty} p(k)$. Para que éste límite no dependa de $p(0)$ es necesario y suficiente que P sea regular.

Consideremos un segundo ejemplo dado por un conjunto W de tres páginas A,B y C con los enlaces mostrados en la Figura 2.

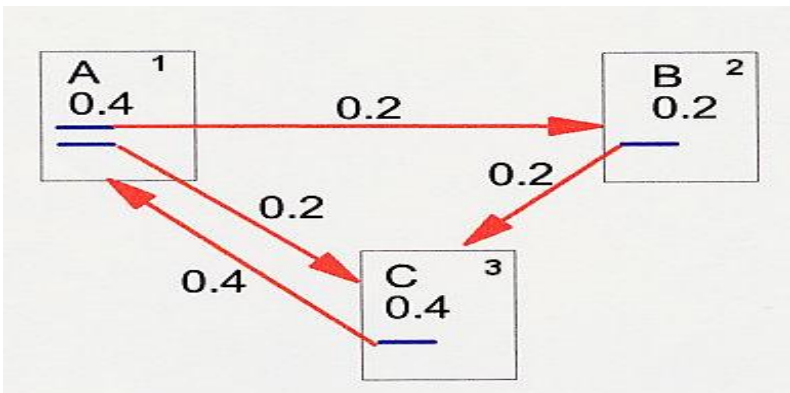


Figura 2: $R(A) = 0,4 \quad R(B) = 0,2 \quad R(C) = 0,4$

Llamemos por simplicidad 1,2 y 3 a A,B y C, respectivamente. Dado que la página 3 sólo tiene un enlace hacia la 1, transmite a ésta todo su valor; la página 2 recibe la mitad del valor de la 1; y la página 3 recibe la mitad del valor de la 1 y todo el valor de la 2. Las ecuaciones (1) y (2) para

este ejemplo se concretan en

$$\begin{cases} R(1) = R(3), \\ R(2) = R(1)/2, \\ R(3) = R(1)/2 + R(2), \end{cases} \qquad R(1) + R(2) + R(3) = 1,$$

que resueltas dan

$$R(A) = 0,4 \quad R(B) = 0,2 \quad R(C) = 0,4.$$

La recurrencia asociada a este ejemplo es

$$\begin{cases} p_1(k+1) = p_3(k), \\ p_2(k+1) = p_1(k)/2, \\ p_3(k+1) = p_1(k)/2 + p_2(k). \end{cases}$$

Cualquiera que sea el vector inicial $\boldsymbol{p}(0) := (p_1(0), p_2(0), p_3(0))^{\mathrm{T}}$ no negativo tal que $p_1(0) + p_2(0) + p_3(0) = 1$, la solución de esta recurrencia

converge a $\boldsymbol{\ell} := (0{,}4,\ 0{,}2,\ 0{,}4)^{\mathrm{T}}$. Esto es así pues la matriz de los coeficientes

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{pmatrix}$$

es irreducible primitiva de radio espectral 1; por tanto, $\boldsymbol{\ell} > 0$ y es independiente de $\boldsymbol{p}(0)$.

2.1. Modelos de navegación

Un internauta suele navegar a través de las páginas web de la manera siguiente. Comienza en una página cuya dirección URL ha visto en otro documento y la ha escrito a mano en el campo Dirección del navegador. De ahí salta a nuevos enlaces que hay en esa página, y así sucesivamente, yendo siempre hacia delante.

Otras veces el internauta acude primero a la página principal de un buscador (Google, Altavista, etc.), plantea allí sus palabras de búsqueda e inicia los saltos hacia páginas seleccionadas, y desde éstas hacia delante.

En ocasiones llega a una página sin enlaces (página "colgada"); en cuyo caso, echa marcha atrás con la flecha de retorno del navegador, y vuelve a la página anterior, y suele ir marcha atrás hasta alguna página anterior. Con mucha frecuencia, vuelve hasta la página de resultados retornados por el buscador, y empieza allí de nuevo desde otra de las páginas seleccionadas por la búsqueda. Muchas veces le sale la página con el "error 404" de página no encontrada, en cuyo caso, la vuelta hacia atrás es inmediatamente ejecutada.

Debemos observar dos cosas: primera, que el internauta puede/suele

retroceder y segunda, que el internauta no puede saltar a una página elegida al azar en la totalidad de la Red. Siempre debe partir de algún lugar URL conocido.

Un modelo de este comportamiento ha sido propuesto de la siguiente manera: la Red puede ser vista como un grafo dirigido (o digrafo) G en el que el conjunto de los vértices es W, el conjunto de las páginas web, y las flechas (o aristas) son los enlaces a otras páginas. Habitualmente, estos enlaces van hacia páginas situadas en otro servidor. Cuando son hacia páginas del mismo directorio (y servidor) suelen estar más adentro en la estructura de subdirectorios; aunque también pueden estar más afuera, siendo el caso frecuente de páginas `HomePage.html` o `index.html`.

Sea $p, 0 < p < 1$, un número dado. Supongamos que el internauta describe un paseo aleatorio por el digrafo G que podemos imaginar de la manera siguiente: en cada página web u decide si marchará o bien hacia delante siguiendo un ultraenlace, elegido uniformemente al azar, o bien hacia una página elegida al azar en toda la Red (con probabilidad p). Si en u hay N_u ultraenlaces, con probabilidad $(1-p)/N_u$ el usuario elige hacer clic en uno de ellos, o saltar a una página cualquiera de W con probabilidad p. Procediendo de este modo, cada página web es visitada con más o menos

frecuencia. Si nuestro nauta aleatorio hace en total un número muy grande T de visitas, y de éstas pasa T_v veces por la página v, entonces la frecuencia

$$\frac{T_v}{T}$$

es una medida de la importancia de v. Precisemos algo más este modelo. Sea n el número total de páginas web existentes. Denotemos por $p_v(k)$ la probabilidad de que el internauta visite la página v en su k-ésima visita. Llamando $\boldsymbol{p}(k) := (p_1(k), \ldots, p_n(k))^{\mathrm{T}}$ al vector no negativo de estas probabilidades; por tanto, $\|\boldsymbol{p}(k)\|_1 = 1$, se tiene que dicho vector satisface la recurrencia

$$p_v(k) = \sum_{u=1}^{n} p_u(k-1)p_{uv} \tag{5}$$

para cada $u \in W$, y donde $P = (p_{uv})$ es la matriz $n \times n$ definida así: cada elemento p_{uv} representa la probabilidad de ir de la página u a la v; primero, si en u no hay ultraenlaces hacia v, dicha probabilidad es p/n; segundo, si hay ultraenlaces, la probabilidad es

$$\frac{p}{n} + \frac{1-p}{N_u},$$

por lo que las filas de P suman 1. De donde, la matriz P es estocástica. Siguiendo a Brin y Page, tomaremos el límite

$$\pi_v := \lim_{k\to\infty} p_v(k), \tag{6}$$

como una medida de la importancia de la página v. Pero, ¿existe siempre este límite? En el Teorema 1 de la pág. 4 de [13], se ha probado que sí existe bajo la hipótesis adicional de que cada página tiene al menos un ultraenlace, i.e. que no existan páginas "colgadas". El modelo de cadena de Markov exige que no se contemple la existencia de vértices "colgados", pues la correspondiente fila en la matriz de transición de probabilidades sería cero. Lo que es imposible en una matriz estocástica.

La idea de la demostración que allí se hace es la siguiente. Como la matriz P es positiva, P es irreducible y primitiva (aperiódica). Por lo tanto, existe el límite $\lim_{k\to\infty} P^k$. En consecuencia, para todo vector no negativo $\boldsymbol{p}(0) = (p_1(0), \ldots, p_n(0))^{\mathrm{T}}$ tal que $\|\boldsymbol{p}(0)\|_1 = 1$, se sigue que existe el límite

$$\boldsymbol{\pi} = \lim_{k\to\infty} \boldsymbol{p}(0)^{\mathrm{T}} P^k,$$

donde $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)$; y este límite no depende del vector de distribución inicial $\boldsymbol{p}(0)$.

2.2. Nota media "verdadera"

El siguiente problema tiene analogía con la cuestión de cómo valorar cada página web. Supongamos que una empresa tiene n empleados y que desea obtener una valoración de su personal en orden a llevar a cabo una determinada tarea por los más aptos para ella. Esta valoración se hace por medio de una nota entre 0 y 10. Esta nota es obtenida mediante la opinión de cada empleado sobre sus compañeros y él mismo. Cada empleado da una nota entre 0 y 10 a todos los empleados y a sí mismo. Después la nota media es asignada a cada empleado. Pero esta calificación es *provisional* y debe ser mejorada teniendo en cuenta *las notas medias de los empleados* que han puesto las notas. Esto tiene cierta lógica si pensamos que las opiniones de los mejores trabajadores merecen más atención que las de los peores. A continuación obtenemos la nota media de las notas recibidas por cada empleado, pero esta vez se trata de la nota media **ponderada** con los pesos las notas medias primeras.

De esta forma se obtiene unas segundas notas medias que son más ajustadas. Luego con estas notas se obtienen unas terceras notas medias ponderadas, que son mucho más precisas. Se itera de nuevo el proceso y puede

esperarse que, eventualmente, la nota media recibida por cada empleado se estabilizará alrededor de la nota "verdadera".

- **Planteamiento**

Sea a_{ij} la nota dada por el empleado i al empleado j, para todos $i, j = 1, 2, \ldots, n$. De esta forma se obtiene una matriz no negativa

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{7}$$

que tiene en la columna j las diversas notas que ha recibido el empleado j de sus compañeros y de sí mismo. La media de estas notas es

$$x_j^{(1)} = \frac{a_{1j} + \cdots + a_{nj}}{n}, \quad (j = 1, \ldots, n).$$

Así pues, la primera nota media recibida por j es $x_j^{(1)}$. La segunda nota media ponderada que se asigna a j es

$$x_j^{(2)} = \frac{x_1^{(1)} a_{1j} + x_2^{(1)} a_{2j} + \cdots + x_n^{(1)} a_{nj}}{x_1^{(1)} + x_2^{(1)} + \cdots + x_n^{(1)}}, \quad (j = 1, \ldots, n).$$

Tras estas segundas notas medias ponderadas, se obtiene la tercera $x_j^{(3)}$ y así sucesivamente.

Resumiendo, definamos la p-ésima nota media ponderada asignada a j, $x_j^{(p)}$, por recurrencia de la manera siguiente.

$$\begin{aligned} x_j(0) &:= 1, \\ x_j^{(p+1)} &:= \frac{x_1^{(p)} a_{1j} + x_2^{(p)} a_{2j} + \cdots + x_n^{(p)} a_{nj}}{x_1^{(p)} + x_2^{(p)} + \cdots + x_n^{(p)}}, \end{aligned} \tag{8}$$

$$j = 1, \ldots, n, \quad p = 0, 1, 2, \ldots$$

Surgen de esta manera tres cuestiones:

- ¿Existe el límite de $x_j^{(p)}$ cuando p tiende a infinito?

- Si existe este límite, ¿cuál es su valor?
- Si no, ¿oscila la sucesión? ¿Cuáles son sus valores de adherencia?

Intentaremos responder a estas tres cuestiones en los siguientes apartados.

- **Relación con la teoría de Perron y Frobenius**

Utilizamos las notaciones del artículo satélite *"Matrices no negativas, paseos aleatorios y cadenas de Markov"*. Llamemos

$$x^{(p)} := \begin{pmatrix} x_1^{(p)} \\ \vdots \\ x_n^{(p)} \end{pmatrix}, \quad (p = 0, 1, 2, \ldots)$$

al término p-ésimo de la sucesión de $\mathbb{R}^{n\times 1}$ definida por (8). La recurrencia (8) puede escribirse en términos matriciales de este modo

$$x^{(0)} := [1, 1, \ldots, 1]^{\mathrm{T}}, x^{(p+1)} := \frac{A^{\mathrm{T}} x^{(p)}}{\|x^{(p)}\|_1}, \qquad p = 0, 1, 2, \ldots \tag{9}$$

Utilizando la forma canónica de Jordan de A^{T} y el resultado sobre convergencia de matrices primitivas, se sigue el teorema que ahora enunciaremos.

Teorema 2.1 *Si la matriz $n \times n$, A es positiva, entonces*

(a) La sucesión $\left(x^{(p)}\right)_{p=0,1,2,\ldots}$ converge, digamos a

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{pmatrix},$$

y $\tilde{x}$ es positivo.

(b) $\tilde{x}$ es un vector propio de A^{T} asociado al valor propio $\sum_{i=1}^{n} \tilde{x}_i$.

(c) $\sum_{i=1}^{n} \tilde{x}_i = \varrho(A)$, radio espectral de A.

Sea P una matriz de permutación y sea N la forma normal reducible A. Aquí reducible quiere decir "no necesariamente irreducible". Es decir, que

$$P^{\mathrm{T}} A P = N. \tag{10}$$

Por tanto, haciendo el cambio de variables $y^{(p)} = P^{\mathrm{T}} x^{(p)}$ vemos que el estudio del comportamiento de la sucesión $x^{(p)}$ dada por la recurrencia

$$x^{(p+1)} = \frac{A^{\mathrm{T}} x^{(p)}}{\|x^{(p)}\|_1}, \quad p = 0, 1, 2, \ldots, \quad x^{(0)} = [1, 1, \ldots, 1]^{\mathrm{T}}, \tag{11}$$

queda reducido al de la sucesión $y^{(p)}$ definida por

$$y^{(p+1)} = \frac{N^{\mathrm{T}} y^{(p)}}{\|y^{(p)}\|_1}, \quad p = 0, 1, 2, \ldots, \quad y^{(0)} = [1, 1, \ldots, 1]^{\mathrm{T}}, \tag{12}$$

sin pérdida de generalidad.

Conjetura 2.1 *Sea $\tilde{A}$ la forma normal de la matriz reducible A. Aquí reducible quiere decir "no necesariamente irreducible".*

$$\tilde{A} := \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1r} & A_{1,r+1} & A_{1,r+2} & \cdots & A_{1m} \\ 0 & A_{22} & \cdots & A_{2r} & A_{2,r+1} & A_{2,r+2} & \cdots & A_{2m} \\ \vdots & & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A_{rr} & A_{r,r+1} & A_{r,r+2} & \cdots & A_{rm} \\ 0 & 0 & \cdots & 0 & A_{r+1,r+1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & A_{r+2,r+2} & \cdots & 0 \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & A_{mm} \end{pmatrix}, \quad (13)$$

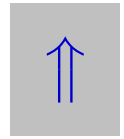
Supongamos que para un solo $i \in \{1, \ldots, m\}$ ***se tiene que*** $\varrho(A_{ii}) = \varrho(A)$. *Esto quiere decir que la clase comunicante $K\{i\}$ cuyos índices corresponden al bloque A_{ii}* en la matriz original A, *es* ***básica***.

1. *Si la matriz A_{ii} es primitiva, la sucesión $\left(x^{(p)}\right)_{p=0,1,2,\ldots}$ converge;*
2. *si la matriz A_{ii} no es primitiva, con índice de imprimitividad d, el término $x^{(p)}$ oscila periódicamente acercándose indefinidamente a d*

valores de adherencia. Las componentes $x_j^{(p)}$ *de* $x^{(p)}$ *para* j *accesibles desde la clase* $K\{i\}$*, oscilan con "periodo"* d*.*

En cualquiera de estos dos casos, la componente $x_j^{(p)}$ *tiende a* 0 *cuando* $p \to \infty$ *para todo* j *que: o bien, tiene acceso a la clase* $K\{i\}$*; o bien, no es accesible desde dicha clase.*

Otro planteamiento para valorar la importancia de una página web, que utiliza cuatro matrices estocásticas diferentes puede verse en [11].



3. Análisis semántico latente

Muchos métodos de búsqueda de textos en las páginas web de Internet dependen de un emparejamiento de palabras "al pie de la letra" entre las palabras que busca el usuario y las que existen en las páginas web. La descomposición de valores singulares de una matriz permite recuperar información basada en conceptos o significados que están latentes en una página web.

La evolución de las bibliotecas digitales e Internet han revolucionado el proceso de almacenamiento y recuperación de la información. La idea de una búsqueda *semántica* tiene hondo calado filosófico. Nos enfrentamos nada menos que al proceso de formación de conceptos en la mente humana. Pondremos un ejemplo concreto para que se reflexione sobre la dificultad. ¿Cuál es la definición de **mesa**? Muchas personas dirían que es un objeto plano de madera con cuatro patas ... que sirve para comer,... escribir ... Si acudimos a un diccionario tampoco nos quedamos satisfechos con la definición de mesa que en él se da. La conclusión es que *es imposible definir el concepto de mesa con palabras*. Es decir, no se puede decir una definición de mesa que convenga a las mesas que conocemos y solo a ellas.

Una definición que incluya las mesas sin patas, que excluya las estanterías, ...¿Qué es para cada uno de nosotros una mesa? Como matemático formado de la mano de Bourbaki, diría que *mesa* es cualquier objeto igual o "parecido" a un elemento del conjunto de mesas que he visto en mi vida. Así pues, la idea de mesa *es* un *conjunto* de mesas. En general, cada concepto puede ser asociado con un conjunto de objetos vistos previamente. Pónganse a estas afirmaciones todas las cautelas que se deseen. Lo cierto es que a la formalización de este asunto nos enfrentamos.

El análisis semántico latente intenta que un programa aprenda qué es una mesa a partir de los textos de la Red que dicen algo de mesas. Este análisis puede extenderse al reconocimiento y emparejamiento de imágenes, escenas, sonidos; es decir, de foto, vídeo y audio. Las ideas matemáticas que expondremos pueden trasladarse a estos campos; las dificultades serán mayores pues habrá que definir algorítmicamente cuando dos fotos son casi iguales (¿distancia de Hausdorff entre dos conjuntos del plano?), o cuándo dos pasajes de sonido son casi el mismo. Estas aplicaciones pueden requerir el uso del tratamiento espectral de señales, muestreo, etc.

3.1. Buscadores

Los buscadores utilizan un diccionario de términos "significativos", que contienen **palabras** (excluidas: artículos, proposiciones, conjunciones, adverbios, etc.) y **expresiones** tales como "tren de aterrizaje", "panes de oro", "dorar la píldora", etc.

Ambos, palabras y expresiones, son considerados indistintamente con el nombre de **términos**. Los buscadores tratan de extraer documentos (`html, pdf, ps, doc, rtf,` etc.) que contengan de forma *destacada* los términos que metemos en la petición de búsqueda.

Como ejemplo primero consideremos los términos *coche, automóvil, conductor, elefante*. Vemos que coche y automóvil son **sinónimos**. Observamos que conductor está relacionado con coche y automóvil. Pero, elefante no está relacionado. Si introducimos *automóvil* en la **búsqueda literal**, el programa no recupera *coche*. Nos gustaría que el programa buscará *automóvil* con **búsqueda de significado**, contrapuesta a búsqueda literal. Así, también traería a nuestra atención documentos con el término *coche* e incluso sería preferible que recuperase los que contienen *conductores* en menor extensión.

He aquí unas cuantas palabras asociadas usualmente: *conductor*, *chófer*, *coche*, *automóvil*, *motor*, *vehículo*, *chasis*, *gasolina*, *parabrisas*, *neumático*, *rueda*, *lunas*, *volante*, *Seat*, *Renault*, *etc.* Es fácil poner ejemplos de palabras no relacionadas como *elefante* y *mosca.* Es obvio que la búsqueda **al pie de la letra** tiene la ventaja de una mayor simplicidad del software empleado.

¿Cuáles son los fallos de la búsqueda literal?

En primer lugar la **polisemia u homonimia:** la búsqueda trae documentos con la *palabra* buscada en los que tiene distinto significado. Ejemplo tópico: *Tarifa*, ciudad y *tarifa* de precios.

En segundo lugar la **sinonimia:** la búsqueda ignora documentos con palabras sinónimas. La petición de, por ejemplo, *automóvil* no recupera *coche*. Estando interesados en averiguar cómo se podía escribir un símbolo al revés con el procesador de textos matemáticos LaTeX introdujimos *reverse latex* en el buscador Google.

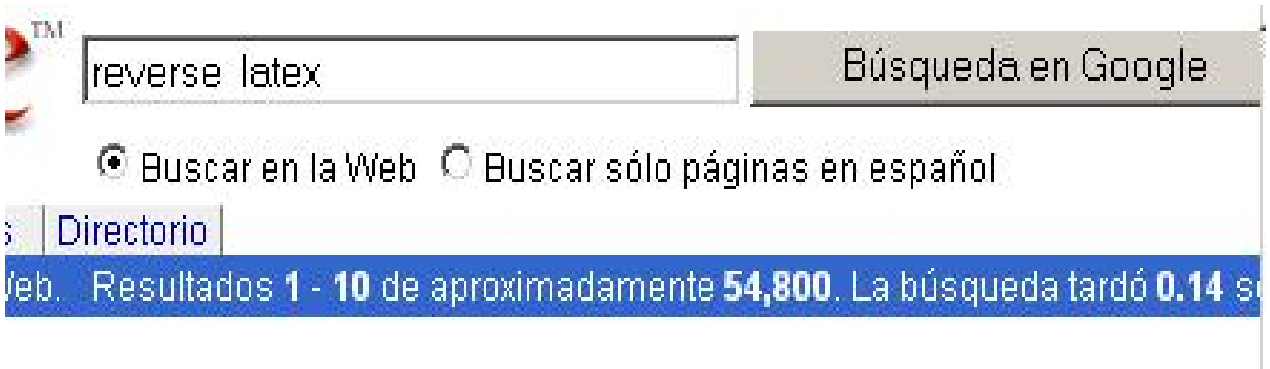


Figura 3: Petición de búsqueda de *reverse latex*.

Obtuvimos las respuestas que se ven en la Figura 4.

Reverse and double-sided printing of **LaTeX** documents - [Traduzca esta página]
... Reverse double-sided printing of **LaTeX** documents. Q:
Can I print a double-sided **LaTeX** document **...**
licensing.mackichan.com/techtalk/v30/30ts65.htm - 27k - En caché - Páginas similares

latex rubber lingerie - [Traduzca esta página]
... latex rubber lingerie, erotic leather lingerie, plus size womens lingerie, lingerie
bridal, Is **latex** rubber lingerie kinky lingerie **reverse latex** lingirie. **...**
lingeriepics.playtoyz.com/latex-rubber-lingerie.html - 5k - En caché - Páginas similares

Re: **Reverse** BiBTeX - [Traduzca esta página]
... From: German Poo Caaman~o; Subject: Re: **Reverse** BiBTeX; Date: Fri, 21 Jul 2000 12:44 **...** asking
on comp.text.tex for a **LaTeX** soltion > would be a better idea. I asked **...**
www.mail-archive.com/lyx-users@lists.lyx.org/msg06139.html - 5k - En caché - Páginas similares

Shopping Items Index - [Traduzca esta página]
... Satin on **reverse**. **Latex** Double Duvet Cover, Black, £170.00,
Latex Double Fitted Sheet, Black, £102.00, **...**
www.bedlinens.co.uk/basket/back-up/page3.htm - 13k - En caché - Páginas similares

Figura 4: Varias acepciones de *reverse latex*.

Por un lado aparecían selecciones relacionadas con el procesador LaTeX, y, por otro, páginas relacionadas con el látex, que es un fluido lechoso que se extrae de los árboles de caucho, con el que se elaboran objetos de lencería, guantes de limpieza, cubrecamas, etc. Esta es una mala pasada que nos juega la polisemia de latex; palabra rara donde las haya.

3.2. Búsqueda semántica

¿Cómo podemos asociar *automóvil* con *coche*? Por las palabras que ambos términos tienen comunes en las páginas web: *motor, vehículo, chasis, gasolina, parabrisas, neumático, rueda, lunas, volante, Seat, Renault, etc.*

Un método bizarro para hacer esto sería el siguiente:

- Se selecciona una página con *automóvil*,
- se buscan otras páginas que compartan con ella 10 palabras.
- ...

¿Funcionará esta idea?

- **Universo semántico**

Nuestro universo semántico puede estar constituido por el conjunto de textos en las páginas web que hay en la Base de Datos del buscador. Dichas páginas son continuamente renovadas por los robots del buscador. Es un universo dinámico, pero por ahora ignoraremos esta dificultad y lo consideraremos fijo. Los textos de las páginas web tienen *ruido* lingüístico. Las palabras tienen diversos significados; incluso significados especiales atribuidos por quienes las escriben. ¿Qué cabe hacer? El ordenador no conoce los idiomas. ¿Hay algún método matemático? Sí, hay varios métodos. Véanse los trabajos de Berry y otros [3] y [4].

Pero este es un asunto en el que aún queda mucha tela por cortar. Aquí expondremos someramente el análisis semántico latente u oculto, que utiliza la descomposición de valores y vectores singulares de matrices rectangulares. Seguimos de cerca la exposición descrita en el excelente libro de Meyer [10, p. 419–421]. También revisaremos los conceptos matemáticos utilizados.

- **Análisis semántico oculto**

Supongamos que la base de datos de nuestro buscador contiene un total de m términos diferentes y n documentos o páginas. Consideremos la matriz $\boldsymbol{A} = (a_{ij}), m \times n$ que muestra las incidencias de las términos en cada documento.

$$\begin{array}{c} \\ \text{términos} \end{array} \begin{array}{c} \text{documentos} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \end{array} \qquad d_j = \begin{array}{c} \text{página } D_j \\ \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} \end{array}$$

por tanto, a_{ij} := es el número de veces que el término T_i aparece en la página web D_j. Cada columna de $\boldsymbol{A}$ corresponde a una página web y cada fila esta asociada a un término (palabra o "expresión idiomática"). En septiembre de 2001 una aproximación para m y n podía ser $m \approx 300.000$ términos y $n \approx$1.600.000.000 páginas.

Un ejemplo ficticio para un par de páginas que contienen *automóvil* una y *coche* la otra, pero no ambas palabras a la vez, podría ser el siguiente.

⋮	⋮	⋮	⋮
automóvil	2	0	⋮
⋮	⋮	⋮	⋮
coche	0	5	⋮
⋮	⋮	⋮	⋮
motor	1	3	⋮
vehículo	5	3	⋮
chasis	1	2	⋮
conductor	3	4	⋮
rueda	3	3	⋮
⋮	⋮	⋮	⋮
	$\boldsymbol{d}_j$	$\boldsymbol{d}_k$	

Llamando $\boldsymbol{d}_j$ a la página primera y $\boldsymbol{d}_k$ a la página segunda, vemos que contienen 2 y 5 veces, respectivamente, las palabras automóvil y coche. Una manera de relacionar ambas páginas que saque a relucir los términos con componentes **no nulas** que comparten (*motor*,...,*rueda*,...) es haciendo el producto escalar ordinario de los vectores $\boldsymbol{d}_j$ y $\boldsymbol{d}_k$ de $\mathbb{R}^{m\times 1}$

$$\boldsymbol{d}_j \cdot \boldsymbol{d}_k = \boldsymbol{d}_j^{\mathrm{T}} \boldsymbol{d}_k \neq 0.$$

donde $^{\mathrm{T}}$ denota matriz traspuesta.

Denotemos por $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n$ las columnas de la matriz $\boldsymbol{A}$,

$$\boldsymbol{A} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n];$$

si dos columnas son ortogonales, i.e. su producto escalar es cero, se tendrá que no comparten ningún término; tal cosa, podría ocurrir para una página que contiene a *mandril* y otra página que contiene *estocástico*. Podríamos tomar una medida de distancia entre dos páginas como el valor de su producto escalar. Pero, esto equivaldría a ponderar no solo la existencia de términos comunes, sino también cuántas veces aparecen éstos. Por eso, una distancia que ignore este último aspecto es el ángulo, o su coseno, entre ambos vectores columnas. Así pues, modularemos el valor del producto escalar

dividiéndolo por el producto de las normas de los vectores.

Cuando alguien introduce una petición de búsqueda en el buscador, podemos considerar que ha dado un **vector de búsqueda**

$$\boldsymbol{q}^{\mathrm{T}} = (q_1, q_2, \ldots, q_m) \in \mathbb{R}^{1\times m},$$

definido por

$$q_i = \begin{cases} 1 & \text{si el término } T_i \text{ aparece,} \\ 0 & \text{en otro caso,} \end{cases}$$

y luego hay que hallar el ángulo que forma $\boldsymbol{q}$ con cada una de las páginas $\boldsymbol{d}_j$ de la base de datos:

$$\cos\theta_j = \frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{d}_j}{\|\boldsymbol{q}\|\|\boldsymbol{d}_j\|} = \frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{e}_j}{\|\boldsymbol{q}\|\|\boldsymbol{A}\boldsymbol{e}_j\|},$$

donde denotamos por $\boldsymbol{e}_j$,

$$\boldsymbol{e}_j^{\mathrm{T}} := (0, \ldots, 1, \ldots, 0),$$

con un 1 en el lugar j-ésimo, $j \in \{1, \ldots, n\}$, el j-ésimo vector de la base canónica de $\mathbb{R}^{n\times 1}$.

Es equivalente decir que el ángulo θ_j es casi 0 ó que su coseno $\cos\theta_j$ es casi 1. Adoptamos el número $\tau, 0 < \tau < 1$, próximo a 1 como el **umbral de tolerancia**. Este valor debe ser determinado heurísticamente, mientras no avancen nuestros conocimientos teóricos al respecto. Si $\cos\theta_j \geq \tau$, hacemos que el buscador traiga la página j a la atención del peticionario.

3.3. Descomposición de valores singulares

Los valores singulares de una matriz $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ se definen como las raíces cuadradas no negativas de los valores propios de la matriz simétrica $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, que es definida no negativa (o semidefinida positiva). Los valores singulares de $\boldsymbol{A}$ se ordenan en sentido decreciente

$$\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_p(\boldsymbol{A}),$$

donde $p = \mín(m, n)$. Si el contexto lo permite, se omite la matriz $\boldsymbol{A}$ de su expresión y se denotan simplemente por

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p.$$

Un importante teorema [6, p. 71–73] nos dice que existen matrices ortogonales

$$\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m] \in \mathbb{R}^{m \times m}, \boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{n \times n},$$

tales que

$$\boldsymbol{U}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{V} = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{pmatrix} \in \mathbb{R}^{m \times n}. \tag{14}$$

La forma de la matriz del segundo miembro no tiene que ser cuadrada necesariamente. Sin decir esto, la notación podría resultar confusa.

Los vectores columnas de $\boldsymbol{U}$ y $\boldsymbol{V}$ vectores singulares por la izquierda y la derecha, respectivamente, de $\boldsymbol{A}$; a diferencia de los valores singulares, no están unívocamente determinados. Se sobrentenderá en adelante que las afirmaciones que se hagan sobre ellos son independientes de los vectores elegidos. Pero, como muy bien dice Meyer [10, p. 554-555], **¡precaución!** Los vectores singulares por la derecha $\boldsymbol{v}_i$ de $\boldsymbol{A}$ son vectores propios de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ y los vectores singulares por la izquierda $\boldsymbol{u}_i$ de $\boldsymbol{A}$ son vectores propios de

$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$; pero esto no significa que cualesquiera bases ortonormales de vectores propios de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ y $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ puedan ser utilizadas como vectores singulares por la derecha e izquierda, respectivamente, de $\boldsymbol{A}$. Las columnas $\boldsymbol{v}_i$ de cualquier matriz ortogonal $\boldsymbol{V}$ que diagonalice a $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ pueden servir como vectores singulares por la derecha de $\boldsymbol{A}$, pero los vectores singulares por la izquierda correspondientes $\boldsymbol{u}_i$ vienen dados por las fórmulas:

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i, \quad i = 1, 2, \ldots, r \Longrightarrow \boldsymbol{u}_i = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\sigma_i} = \frac{\boldsymbol{A}\boldsymbol{v}_i}{\|\boldsymbol{A}\boldsymbol{v}_i\|}, \quad i = 1, 2, \ldots, r;$$

para $\{\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_m\}$ se elige cualquier base ortonormal de $\operatorname{Ker} \boldsymbol{A}^{\mathrm{T}}$.

También se cumple

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i, i = 1, \ldots, r.$$

Si $r = \operatorname{rg} \boldsymbol{A}$ entonces

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0,$$

además, para los subespacios núcleo e imagen de $\boldsymbol{A}$ se tiene que

$$\operatorname{Ker} \boldsymbol{A} = \langle \boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n \rangle, \quad \operatorname{Im} \boldsymbol{A} = \langle \boldsymbol{u}_1, \ldots, \boldsymbol{u}_r \rangle,$$

donde $\langle \cdots \rangle$ denota el subespacio engendrado por los vectores contenidos entre los delimitadores angulares.

Con las notaciones

$$\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{m\times k}, \boldsymbol{V}_k = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{n\times k},$$

se tiene

$$\operatorname{Im} \boldsymbol{A} = \operatorname{Im} \boldsymbol{U}_r, \quad \operatorname{Im} \boldsymbol{A}^{\mathrm{T}} = \operatorname{Im} \boldsymbol{V}_r$$

y las relaciones importantes

$$\operatorname{Ker} \boldsymbol{A} \oplus \operatorname{Im} \boldsymbol{A}^{\mathrm{T}} = \mathbb{R}^{n\times 1}, \quad \operatorname{Im} \boldsymbol{A} \oplus \operatorname{Ker} \boldsymbol{A}^{\mathrm{T}} = \mathbb{R}^{m\times 1},$$

donde $\oplus$ denota suma directa ortogonal.

De la descomposición (14) resulta que es posible expresar $\boldsymbol{A}$ como la suma de r matrices de rango 1:

Descomposición diádica

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}.$$

Se llama diada a toda matriz de rango 1. Esta terminología proviene de la

Física, donde se habla de tensores diádicos.

Usaremos la norma espectral de matrices $\|\boldsymbol{A}\|$, i.e. la norma inducida por las normas euclídeas en $\mathbb{R}^{m\times 1}$ y $\mathbb{R}^{n\times 1}$

$$\|\boldsymbol{A}\| := \max_{\boldsymbol{x}\in\mathbb{R}^{n\times 1}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2};$$

entonces, $\|\boldsymbol{A}\| = \sigma_1(\boldsymbol{A})$.

Teorema de Eckart y Young

Dado un entero $k, 0 \leq k < r$, sea

$$\boldsymbol{A}_k := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}},$$

entonces

$$\min_{\operatorname{rg}\boldsymbol{X}\leq k} \|\boldsymbol{X} - \boldsymbol{A}\| = \|\boldsymbol{A}_k - \boldsymbol{A}\| = \sigma_{k+1}(\boldsymbol{A}).$$

En particular,

$$\min_{\operatorname{rg} \boldsymbol{X} < r} \|\boldsymbol{X} - \boldsymbol{A}\| = \|\boldsymbol{A}_{r-1} - \boldsymbol{A}\| = \sigma_r(\boldsymbol{A}).$$

Llamemos

$$\boldsymbol{\Sigma}_k := \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{pmatrix} \in \mathbb{R}^{k \times k}$$

y

$$[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n] := \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}}, \qquad k = 1, \ldots, p.$$

Se tiene que

$$\boldsymbol{A}_k \boldsymbol{e}_j = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}} \boldsymbol{e}_j = \boldsymbol{U}_k \boldsymbol{s}_j;$$

dado que

$$\boldsymbol{U}_k^{\mathrm{T}} \boldsymbol{U}_k = (\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{u}_j) = I_k,$$

se sigue que

$$\|\boldsymbol{U}_k \boldsymbol{s}_j\| = \|\boldsymbol{s}_j\|.$$

De la definición de $\boldsymbol{A}_k$ deducimos que

$$\boldsymbol{A}_k = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}};$$

por tanto,

$$\operatorname{Im} \boldsymbol{A}_k = \operatorname{Im} \boldsymbol{U}_k, \quad \operatorname{Im} \boldsymbol{A}_k^{\mathrm{T}} = \operatorname{Im} \boldsymbol{V}_k.$$

Por la fórmula (5.13.4), pág. 430, de Meyer [10] se tiene que

$$\boldsymbol{P}_{\operatorname{Im} \boldsymbol{A}_k} = \boldsymbol{U}_k \boldsymbol{U}_k^{\mathrm{T}}.$$

Volvamos a nuestra pesquisa principal. En vez de hallar el coseno

$$\cos \theta_j = \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{e}_j}{\|\boldsymbol{q}\| \|\boldsymbol{A} \boldsymbol{e}_j\|},$$

aproximando la matriz $\boldsymbol{A}$ por una matriz de rango k, menor que $r := \operatorname{rg} \boldsymbol{A}$, podemos considerar el coseno

$$\cos \varphi_j := \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{e}_j}{\|\boldsymbol{q}\| \|\boldsymbol{A}_k \boldsymbol{e}_j\|}$$

como pseudo-coseno de $\boldsymbol{q}$ con $\boldsymbol{d}_j$. Utilizando fórmulas previas,

$$\cos \varphi_j = \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{U}_k \boldsymbol{s}_j}{\|\boldsymbol{q}\| \|\boldsymbol{s}_j\|}.$$

La matriz $\boldsymbol{U}_k$ y los vectores $\boldsymbol{s}_j, j \in \{1, \ldots, n\}$ solo hay que calcularlos una sola vez. Lo único que cambiaría sería el vector de búsqueda $\boldsymbol{q}$. Aquí k se toma *considerablemente* menor que r.

Podemos llamar **espacio vectorial de los documentos** al subespacio $\operatorname{Im} \boldsymbol{A}$ de $\mathbb{R}^{m\times 1}$, y **espacio vectorial de términos** al subespacio $\operatorname{Im} \boldsymbol{A}^{\mathrm{T}}$ de $\mathbb{R}^{n\times 1}$. Aproximamos estos espacios por los espacios de pseudo-documentos, $\operatorname{Im} \boldsymbol{A}_k$, y de pseudo-términos $\operatorname{Im} \boldsymbol{A}_k^{\mathrm{T}}$.

¿Es válida esta aproximación? Dicho de otro modo, ¿por qué la aproximación de $\boldsymbol{A}$ por $\boldsymbol{A}_k$ funciona desde un punto de vista semántico? Los diversos autores hablan aquí de que "la variación en el uso del vocabulario y la ambigüedad de muchas palabras producen *ruido* significativo en $\boldsymbol{A}$". Al tomar $\boldsymbol{A}_k$ en lugar de $\boldsymbol{A}$, por un lado, "se captura lo suficiente de la estructura que asocia términos y documentos para retener su significado *oculto*", y, por otro, "se consigue quitar *ruido*". El autor de este escrito no acaba de entender el porqué de este fenómeno. Pensemos en nuestra búsqueda de documentos con la palabra *coche*. Queremos que el buscador nos traiga documentos significativos aunque no contengan dicha palabra. Ya hemos dicho que esto sólo puede hacerce a través de las palabras comunes entre documentos. Las palabras *chófer*, *ruedas*, *motor*, *gasolina*, etc.

estarán en la intersección evocada. Pero, el término *chófer* está vinculado a *conductor*, y éste a *líder, comunicador, ... religión, televisión, ...* Claro que estos **campos semánticos** son muy distintos y es poco probable que un documento sobre **presentadores de televisión** contenga las palabras *chófer, ruedas, motor, gasolina,* etc. Pero puede haberlos.

Una buena explicación sobre el uso de los valores singulares para filtrar el ruido de unos datos puede leerse en el Ejemplo 5.12.3, "Filtering Noisy Data" en el libro de Meyer [10, p. 418].

Al considerar $\cos \varphi_j$ removemos parte del ruido lingüístico, y son traídos a nuestra atención más documentos relacionados con la petición de búsqueda.

3.4. Alternativa

Sea $\boldsymbol{P}_{\operatorname{Im}\boldsymbol{A}}$ la proyección ortogonal de $\mathbb{R}^m$ sobre $\operatorname{Im}\boldsymbol{A}$. Es conocido [10, p. 435, (5.13.12)] que la matriz de $\boldsymbol{P}_{\operatorname{Im}\boldsymbol{A}}$ en la base canónica es igual a $\boldsymbol{U}_r\boldsymbol{U}_r^{\mathrm{T}}$.

En vez de buscar los vectores $\boldsymbol{d}_j$ que forman un ángulo "pequeño" con el vector de búsqueda $\boldsymbol{q}$, podemos considerar la proyección ortogonal

$$\tilde{\boldsymbol{q}} := \boldsymbol{P}_{\operatorname{Im}\boldsymbol{A}}(\boldsymbol{q})$$

de $\boldsymbol{q}$ sobre el subespacio de documentos $\operatorname{Im}\boldsymbol{A}$. Por el teorema de Pitágoras en el espacio euclídeo m-dimensional, es sabido que $\tilde{\boldsymbol{q}}$ es el vector de $\operatorname{Im}\boldsymbol{A}$ que está más próximo a $\boldsymbol{q}$ en la norma euclídea:

$$\min_{\boldsymbol{x}\in\operatorname{Im}\boldsymbol{A}} \|\boldsymbol{x}-\boldsymbol{q}\|_2 = \|\tilde{\boldsymbol{q}}-\boldsymbol{q}\|_2.$$

Después buscamos los vectores $\boldsymbol{d}_j$ que forman los ángulos más "pequeños" con $\tilde{\boldsymbol{q}}$. Estos vectores son determinados calculando los cosenos

$$\cos\tilde{\theta}_j = \frac{\tilde{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{e}_j}{\|\tilde{\boldsymbol{q}}\|\|\boldsymbol{A}\boldsymbol{e}_j\|}.$$

Como

$$\boldsymbol{q} = \tilde{\boldsymbol{q}} + \boldsymbol{q}_2, \quad \text{con } \tilde{\boldsymbol{q}} \in \operatorname{Im} \boldsymbol{A}, \boldsymbol{q}_2 \in \operatorname{Ker} \boldsymbol{A}^{\mathrm{T}}$$

de manera única, se tiene que $\tilde{\boldsymbol{q}}^{\mathrm{T}} \boldsymbol{A} = \boldsymbol{q}^{\mathrm{T}} \boldsymbol{A}$; por esta razón y la desigualdad $\|\tilde{\boldsymbol{q}}\| \leq \|\boldsymbol{q}\|$, se sigue que $\cos \tilde{\theta}_j \geq \cos \theta_j$. Por tanto, más documentos son considerados. Utilizando la aproximación $\boldsymbol{A}_k \approx \boldsymbol{A}$, proyectaríamos $\boldsymbol{q}$ ortogonalmente sobre el subespacio $\operatorname{Im} \boldsymbol{A}_k$; llamando

$$\boldsymbol{d} := \boldsymbol{P}_{\operatorname{Im} \boldsymbol{A}_K}(\boldsymbol{q}),$$

se deduce que $\boldsymbol{d} = \boldsymbol{U}_k \boldsymbol{U}_k^{\mathrm{T}} \boldsymbol{q}$; en consecuencia, son más fácilmente calculados los pseudo-cosenos

$$\cos \tilde{\varphi}_j = \frac{\boldsymbol{d}^{\mathrm{T}} \boldsymbol{A}_k \boldsymbol{e}_j}{\|\boldsymbol{d}\| \|\boldsymbol{A}_k \boldsymbol{e}_j\|} = \frac{\boldsymbol{U}_k \boldsymbol{U}_k^{\mathrm{T}} \boldsymbol{q} \boldsymbol{U}_k \boldsymbol{s}_j}{\|\boldsymbol{U}_k^{\mathrm{T}} \boldsymbol{q}\| \|\boldsymbol{s}_j\|},$$

una vez conocidos $\boldsymbol{U}_k$ y los $\boldsymbol{s}_j$.

3.5. Búsquedas semánticas

Tomemos un k adecuado, mucho menor que r, y consideremos el espacio de los pseudo-documentos $\operatorname{Im} \boldsymbol{A}_k = \operatorname{Im} \boldsymbol{U}_k$. Como en la sección anterior sea

$\boldsymbol{d}$ la proyección ortogonal de $\boldsymbol{q}$ sobre este subespacio. Su expresión en la base ortonormal del subespacio es

$$\boldsymbol{P}_{\operatorname{Im}\boldsymbol{U}_k}(\boldsymbol{q}) = \sum_{j=1}^{k} (\boldsymbol{q}^{\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j =: \boldsymbol{d}.$$

Llamemos $\hat{\boldsymbol{q}} := \boldsymbol{q}^{\mathrm{T}}\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{-1}$; dado que

$$\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{-1} = \left[\frac{\boldsymbol{u}_1}{\sigma_1}, \ldots, \frac{\boldsymbol{u}_k}{\sigma_k}\right]$$

se tiene que

$$\hat{\boldsymbol{q}} = \boldsymbol{q}^{\mathrm{T}}\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{-1} = \left[\boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_1}{\sigma_1}, \ldots, \boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_k}{\sigma_k}\right].$$

Lo que nos permite hallar $\boldsymbol{d}$ como combinación lineal de $\{\sigma_1\boldsymbol{u}_1, \ldots, \sigma_k\boldsymbol{u}_k\}$, que es otra base del subespacio,

$$\boldsymbol{d} = \sum_{j=1}^{k} (\boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_j}{\sigma_j})\sigma_j\boldsymbol{u}_j.$$

Así pues, las componentes del vector $\hat{\boldsymbol{q}}$ son las **coordenadas ponderadas** de $\boldsymbol{d}$ en esta última base. Cuando $k = 2$ podremos representar el vector $\boldsymbol{d}$ por el punto $\left(\boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_1}{\sigma_1}, \boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_2}{\sigma_2}\right)$ en un plano cartesiano.

Otra medida que puede utilizarse de distancia entre un vector de búsqueda $\boldsymbol{q}$ y un vector de un pseudo-documento $\boldsymbol{\delta}$ es el coseno en $\mathbb{R}^k$ de los vectores

$$\hat{\boldsymbol{q}} = \left[\boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_1}{\sigma_1}, \ldots, \boldsymbol{q}^{\mathrm{T}}\frac{\boldsymbol{u}_k}{\sigma_k}\right] \quad \text{y} \quad \hat{\boldsymbol{\delta}} = \left[\boldsymbol{\delta}^{\mathrm{T}}\frac{\boldsymbol{u}_1}{\sigma_1}, \ldots, \boldsymbol{\delta}^{\mathrm{T}}\frac{\boldsymbol{u}_k}{\sigma_k}\right].$$

Al coseno $\cos(\hat{\boldsymbol{q}}, \hat{\boldsymbol{\delta}})$ le llamaremos **coseno ponderado por columnas** de los vectores $\boldsymbol{q}$ y $\boldsymbol{\delta}$, y será denotado por $\boxed{\mathrm{cospc}(\boldsymbol{q}, \boldsymbol{\delta})}$.

Ejercicio 3.1 ¿Hay alguna relación entre $\mathrm{cospc}(\boldsymbol{q}, \boldsymbol{\delta})$ y $\cos(\boldsymbol{q}, \boldsymbol{\delta})$?

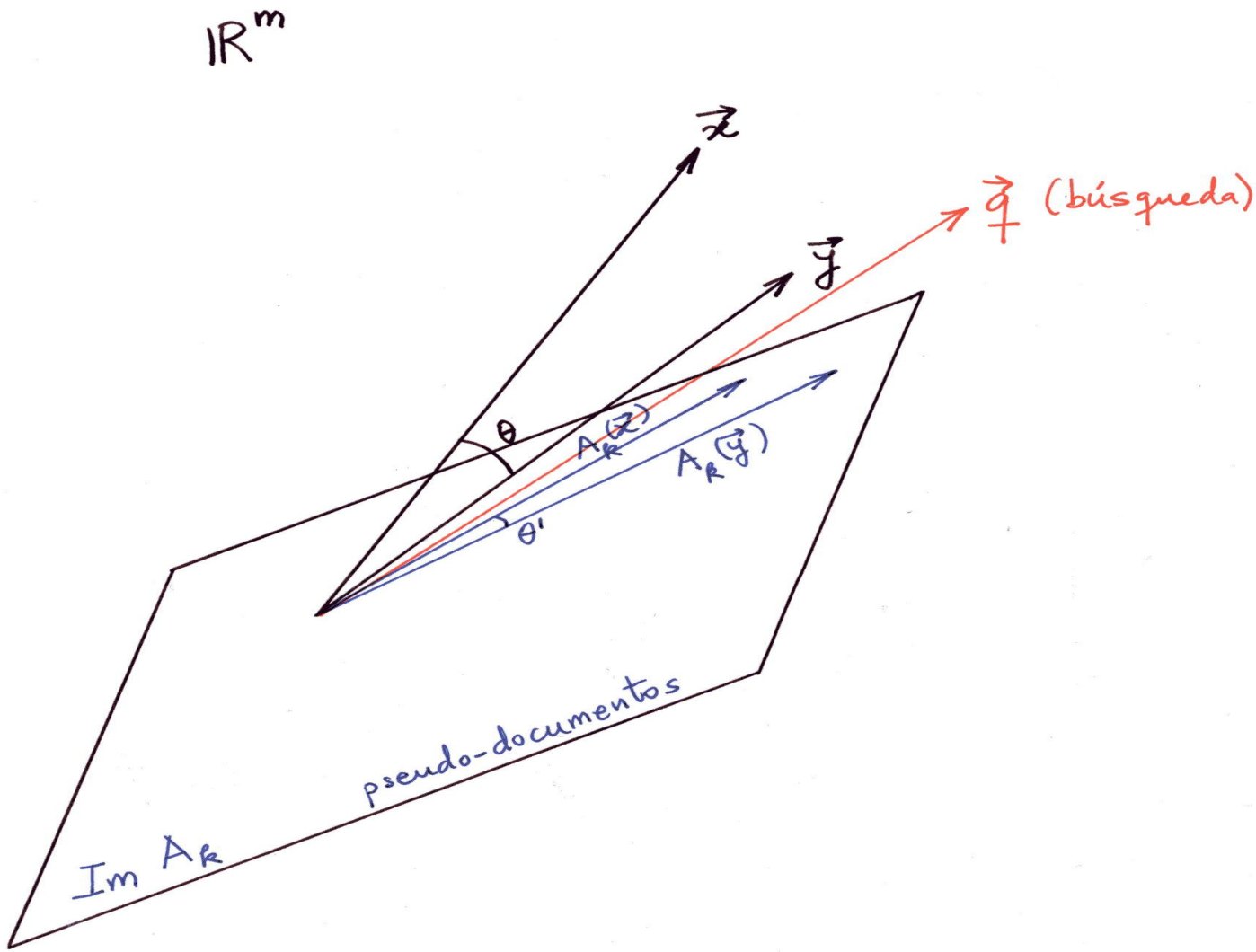


Figura 5: Subespacio Im $\boldsymbol{A}_k$ de pseudo-documentos

3.6. Palabras o términos

Análogas consideraciones cabe hacer cuando se pone el énfasis en los términos en vez de en los documentos. El subespacio $\operatorname{Im} \boldsymbol{A}_k^{\mathrm{T}} = \operatorname{Im} \boldsymbol{V}_k$ es el espacio de los pseudo-términos. Así como un término se asocia con una fila de la matriz de enteros $\boldsymbol{A}$, un pseudo-término será cualquier fila de la matriz $\boldsymbol{A}_k$; es obvio que las componentes de $\boldsymbol{A}_k$ no son números enteros necesariamente, sino números reales no negativos. Sea $\boldsymbol{t}^{\mathrm{T}}$ un vector de $\mathbb{R}^{1\times n}$ (por ejemplo, puede tratarse de una palabra que queramos añadir a la base de datos). Llamemos $\boldsymbol{p}$ a la proyección ortogonal de $\boldsymbol{t}$ sobre el subespacio de los pseudo-términos. Su expresión en la base ortonormal del subespacio es

$$\boldsymbol{P}_{\operatorname{Im} \boldsymbol{V}_k}(\boldsymbol{t}) = \sum_{j=1}^{k} (\boldsymbol{t}^{\mathrm{T}} \boldsymbol{v}_j) \boldsymbol{v}_j =: \boldsymbol{p}.$$

Llamemos $\hat{\boldsymbol{t}} := \boldsymbol{t}^{\mathrm{T}} \boldsymbol{V}_k \boldsymbol{\Sigma}_k^{-1}$; dado que

$$\boldsymbol{V}_k \boldsymbol{\Sigma}_k^{-1} = \left[\frac{\boldsymbol{v}_1}{\sigma_1}, \ldots, \frac{\boldsymbol{v}_k}{\sigma_k} \right]$$

se tiene que

$$\hat{\boldsymbol{t}} = \boldsymbol{t}^{\mathrm{T}} \boldsymbol{V}_k \boldsymbol{\Sigma}_k^{-1} = \left[\boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_1}{\sigma_1}, \ldots, \boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_k}{\sigma_k} \right]$$

Lo que nos permite hallar $\boldsymbol{p}$ como combinación lineal de $\{\sigma_1 \boldsymbol{v}_1, \ldots, \sigma_k \boldsymbol{v}_k\}$, que es otra base del subespacio,

$$\boldsymbol{p} = \sum_{j=1}^{k} (\boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_j}{\sigma_j}) \sigma_j \boldsymbol{v}_j .$$

Así pues, las componentes del vector $\hat{\boldsymbol{t}}$ son las **coordenadas ponderadas** de $\boldsymbol{p}$ en esta última base. Cuando $k = 2$ podremos representar el vector $\boldsymbol{p}$ por el punto $\left(\boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_1}{\sigma_1}, \boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_2}{\sigma_2} \right)$ en un plano cartesiano.

Una medida distinta que puede utilizarse de distancia entre dos términos $\boldsymbol{t}$ y $\boldsymbol{s}$ es el coseno en $\mathbb{R}^k$ de los vectores

$$\hat{\boldsymbol{t}} = \left[\boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_1}{\sigma_1}, \ldots, \boldsymbol{t}^{\mathrm{T}} \frac{\boldsymbol{v}_k}{\sigma_k} \right] \quad \text{y} \quad \hat{\boldsymbol{s}} = \left[\boldsymbol{s}^{\mathrm{T}} \frac{\boldsymbol{v}_1}{\sigma_1}, \ldots, \boldsymbol{s}^{\mathrm{T}} \frac{\boldsymbol{v}_k}{\sigma_k} \right]$$

Al coseno $\cos(\hat{\boldsymbol{t}}, \hat{\boldsymbol{s}})$ le llamaremos **coseno ponderado por filas** de los

vectores $\boldsymbol{t}$ y $\boldsymbol{s}$, y será denotado por $\boxed{\text{cospf}(\boldsymbol{t}, \boldsymbol{s})}$.

Ejercicio 3.2 ¿Hay alguna relación entre $\text{cospf}(\boldsymbol{t}, \boldsymbol{s})$ y $\cos(\boldsymbol{t}, \boldsymbol{s})$?

3.7. Ejemplo

Consideremos la lista de libros de Matemáticas del Cuadro 1 como una *base de datos*. Fijemos nuestra atención en los términos **analysis, calculus, canonical, chaos, differential, eigen, equations, factorization, golden, mathematics, matrix, matlab, method, number, ordinary, problem, spectral, statistic, stochastic, theory**, a las que llamaremos palabras "significativas".

Cuadro 1: **Libros**

Núm.	Título
L1	Proofs and Confirmations : The Story of the Alternating Sign Matrix Conjecture
L2	Factorization and Primality Testing
L3	A Radical Approach to Real Analysis
L4	Second Year Calculus : From Celestial Mechanics to Special Relativity
L5	Three Pearls of Number Theory
L6	Matrix Analysis and Applied Linear Algebra
L7	Matrix Algorithms, Volume II: Eigensystems
L8	Matrix Differential Calculus with Applications in Statistics and Econometrics
L9	The Elements of Statistical Learning : Data Mining, Inference, and Prediction
L10	Spectral Methods in **MATLAB**
L11	Ordinary Differential **Equations**
L12	Iterative Methods for Linear and Nonlinear **Equations**
L13	Partial Differential **Equations** : An Introduction
L14	Elementary Differential **Equations** and Boundary Value Problems
L15	What Is Mathematics? : An Elementary Approach to Ideas and Methods
L16	Theory of Matrices
L17	Chaos : A Statistical Perspective

L18	Turbulent Mirror : An Illustrated Guide to Chaos, Theory and the Science of Wholeness
L19	**Matlab** Companion for Multivariable Calculus
L20	Chaos and Fractals : New Frontiers of Science
L21	Brownian Motion and Stochastic Calculus
L22	Stochastic Partial Differential **Equations** and Kolmogorov **Equations** in Infinite Dimensions
L23	The Golden Section
L24	The Golden Ratio and Fibonacci Numbers
L25	Formalized Music : Thought and Mathematics in Composition
L26	Fractals in Music : Introductory Mathematics for Musical Analysis
L27	Statistical Learning Theory
L28	Ordinary Differential **Equations** Using **Matlab**
L29	Fibonacci Fun: Fascinating Activities With Intriguing Numbers
L30	Boundary Value Problems of Mathematical Physics
L31	Introduction to Spectral Analysis
L32	Matrix Groups : An Introduction to Lie Group Theory
L33	Spectral Theory of Canonical Differential Systems : Method of Operator Identities
L34	Convolution Operators and Factorization of Almost Periodic Matrix Functions
L35	Prime Numbers and Computer Methods for Factorization
L36	The Theory of Canonical Moments With Applications in Statistics, Probability, and Analysis
L37	Theory of Stochastic Canonical **Equations**
L38	Random Matrices, Frobenius Eigenvalues, and Monodromy

Fijamos nuestra atención en las palabras que aparecen repetidas dos o más veces. La correspondencia de estas palabras con los libros viene dada en el Cuadro 2.

Cuadro 2: **Palabras repetidas**

Núm.	Palabra	Libros
1	analysis	3, 6, 26, 31, 36
2	calculus	4, 8, 19, 21
3	canonical	33, 36, 37
4	chaos	17, 18, 20
5	differential	8, 11, 13, 14, 22, 28, 33
6	eigen	7, 38
7	**equations**	11, 12, 13, 14, 22, 22, 28, 37
8	factorization	2, 34, 35
9	golden	23, 24
10	mathematics	15, 25, 26
11	**matlab**	10, 19, 28
12	matrix	1, 6, 7, 8, 32, 34
13	method	10, 12, 15, 33, 35
14	number	5, 24, 29, 35
15	ordinary	11, 28
16	problem	14, 30
17	spectral	10, 31, 33
18	statistic	8, 9, 17, 27, 36
19	stochastic	21, 22, 37
20	theory	5, 16, 18, 27, 32, 33, 36, 37

Sea $\boldsymbol{A} = (a_{ij})$ la matriz 20×38 de *términos* $\times$ *libros* definida por a_{ij} := número de veces que el término i aparece en el título del libro Lj.

Así pues, $a_{13} = 1$ ya que el término *analysis* aparece una vez en el libro L3, y $a_{7,22} = 2$ puesto que *equations* aparece dos veces en el libro L22. Sin embargo, $a_{11} = 0$ dado que *analysis* no aparece en el libro L1. No transcribimos la matriz $\boldsymbol{A}$ por sus excesivas dimensiones. Tomemos $k = 2$ y aproximemos $\boldsymbol{A}$ por la matriz $\boldsymbol{A}_2$. Los dos primeros valores singulares de esta matriz son $\sigma_1 = 4{,}1952$ y $\sigma_2 = 3{,}3361$. La matriz $\boldsymbol{U}_2 := [\boldsymbol{u}_1, \boldsymbol{u}_2]$ obtenida es

$$\boldsymbol{U}_2 = \begin{bmatrix} -0{,}0638 & 0{,}2777 \\ -0{,}0797 & 0{,}0462 \\ -0{,}1851 & 0{,}2788 \\ -0{,}0277 & 0{,}1201 \\ -0{,}5382 & -0{,}1712 \\ -0{,}0061 & 0{,}0260 \\ -0{,}6582 & -0{,}3626 \\ -0{,}0201 & 0{,}0620 \\ -0{,}0022 & 0{,}0132 \\ -0{,}0155 & 0{,}0522 \\ -0{,}1167 & -0{,}0375 \\ -0{,}0959 & 0{,}2373 \\ -0{,}1632 & 0{,}1463 \\ -0{,}0343 & 0{,}1207 \\ -0{,}1609 & -0{,}1210 \\ -0{,}0767 & -0{,}0585 \\ -0{,}1035 & 0{,}1574 \\ -0{,}1232 & 0{,}3373 \\ -0{,}2095 & -0{,}0362 \\ -0{,}2808 & 0{,}6393 \end{bmatrix}$$

El libro Lj está aproximado por la columna j-ésima, $\boldsymbol{d}_j^{(2)}$, de $\boldsymbol{A}_2$. Podemos representar el libro Lj en el plano x, y por medio de las coordenadas ponderadas

$$\boldsymbol{d}_j^{(2)\mathrm{T}} \boldsymbol{U}_2 \boldsymbol{\Sigma}_2^{-1} = \left[\boldsymbol{d}_j^{(2)\mathrm{T}} \frac{\boldsymbol{u}_1}{\sigma_1}, \boldsymbol{d}_j^{(2)\mathrm{T}} \frac{\boldsymbol{u}_2}{\sigma_2} \right].$$

del vector $\boldsymbol{d}_j^{(2)}$ en la base $\{\sigma_1 \boldsymbol{u}_1, \sigma_2 \boldsymbol{u}_2\}$ de $\operatorname{Im} \boldsymbol{A}_2$. Así se consigue la representación de todos los libros dada en la Figuras 6 y 7, sin numerar y numerados, respectivamente.

Veamos el efecto de someter una búsqueda a la base de datos de los libros. Supongamos que estamos interesados en los libros cuyo título contiene **equations y matlab**. Escribimos esta consulta en el buscador de la base de datos.

La palabra $\boxed{\mathbf{y}}$ es desechada por no ser una de las 20 palabras que fueron elegidas para definir la matriz $\boldsymbol{A}$. Nos quedan *equations matlab*. Viendo la numeración de estas dos palabras en el Cuadro 2, observamos que corresponden a los números 7 y 11, respectivamente. Sea $\boldsymbol{q} = (q_1, \ldots, q_{20})^{\mathrm{T}}$ el vector de $\mathbb{R}^{20\times 1}$ cuyas componentes son todas cero excepto $q_7 = 1$ y $q_{11} = 1$. El vector $\boldsymbol{q}$ es el vector de búsqueda. Sea $\boldsymbol{d}$ su proyección ortogonal

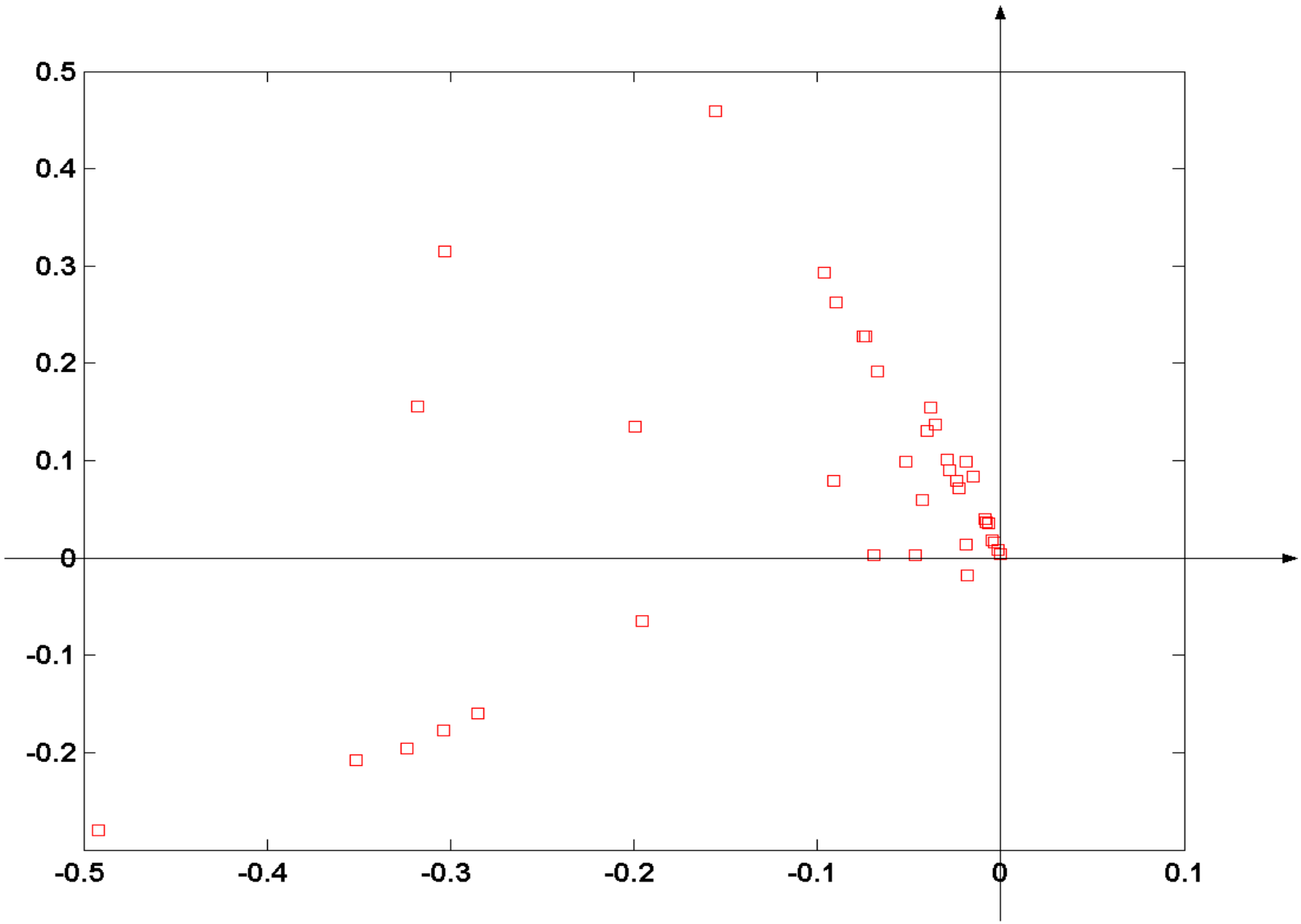


Figura 6: Libros dados por sus coordenadas ponderadas.

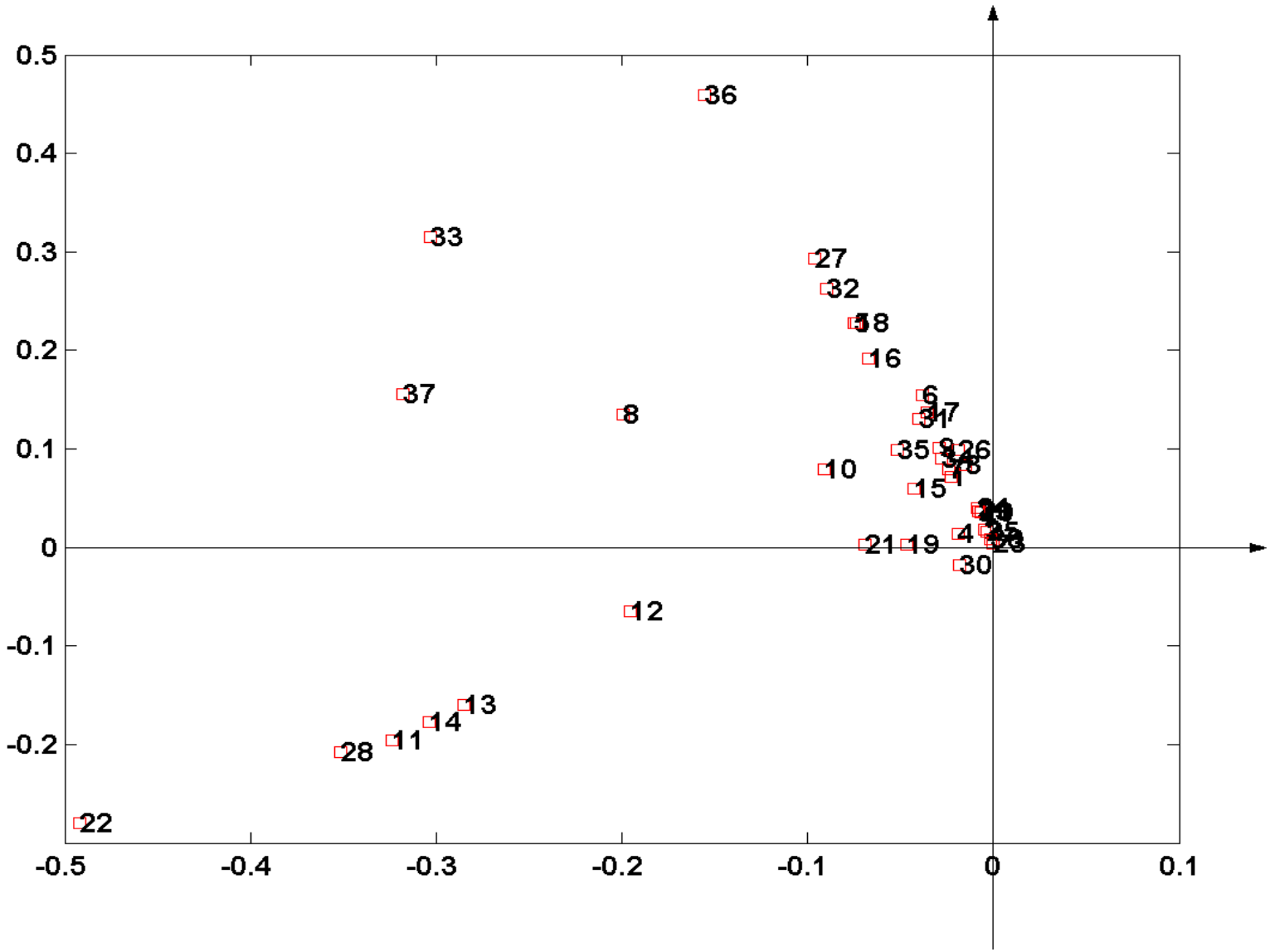


Figura 7: Números de los libros.

sobre $\mathrm{Im}\,\boldsymbol{A}_2$. Representamos esta consulta por un asterisco verde, seguido del signo ?, en la Figura 8. Este punto no coincide con ninguno de los puntos que representan a los 38 libros, pues en sus títulos no están a la vez *equations* y *matlab* como las **únicas** palabras "significativas".

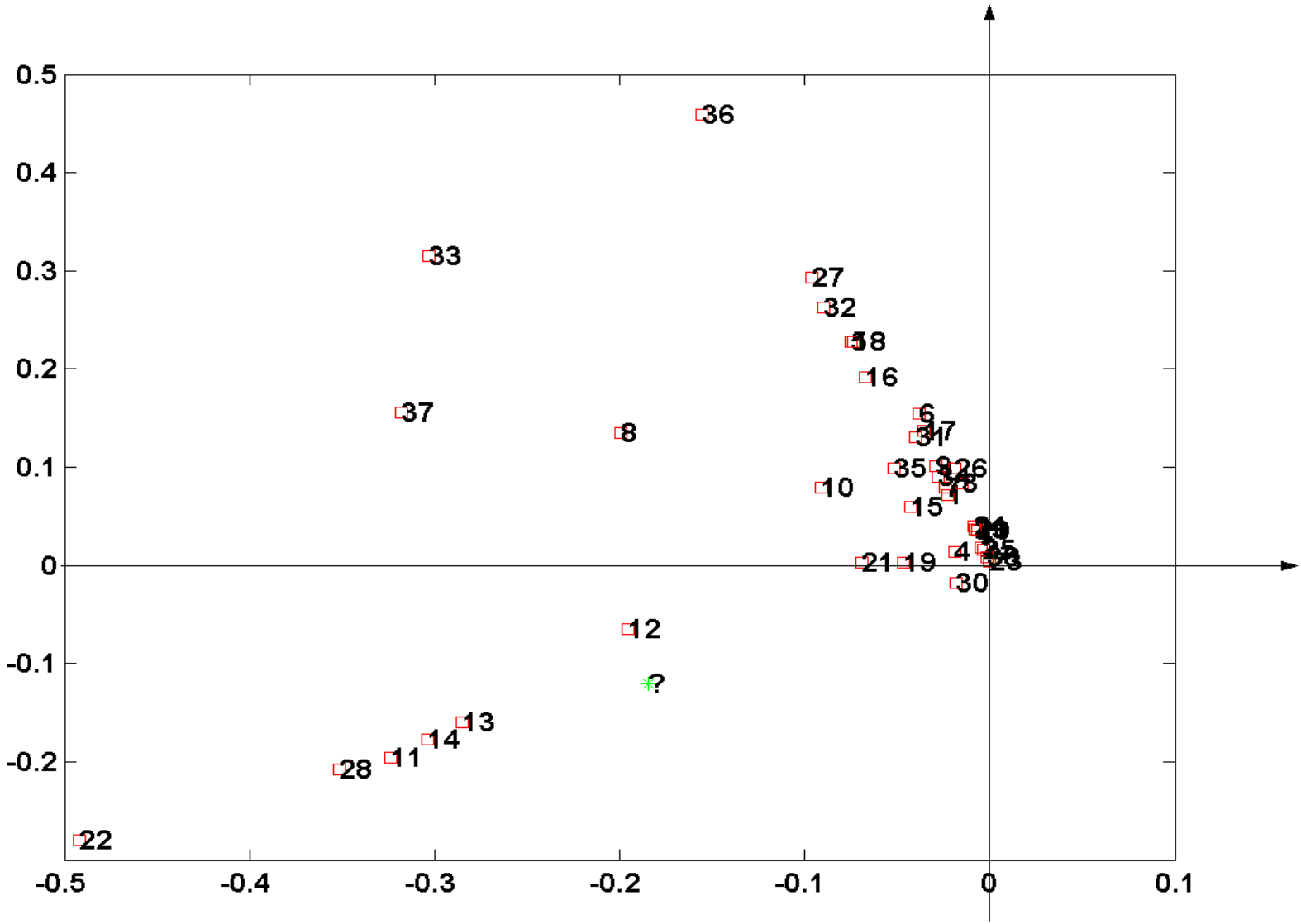


Figura 8: Consulta *equations y matlab*.

Con el umbral $\tau = 0{,}70$, el buscador devuelve los libros $\boldsymbol{d}_j^{(2)}$ tales que

$$\text{cospc}(\boldsymbol{d}, \boldsymbol{d}_j^{(2)}) > 0{,}70.$$

Tales libros están señalados por cuadritos verdes en la Figura 11. Han sido traídos a nuestra atención los libros **L11, L12, L13, L14, L19, L21, L22, L28, L30**. Los libros **L11, L12, L13, L14, L19, L21, L22, L28** contienen *equations* o *matlab*. Los libros **L21** y **L30** no contienen ninguna de estas dos palabras. El libro **L21** ha sido asociado al mismo campo semántico que **equations, matlab**. ¿Cómo? El libro **L21** es capturado por las palabras *stochastic* y *calculus* a través de esta posible vía: *stochastic* conecta con **equations** en **L22** y **L37**; **L8** vincula *differential* con *calculus* y **L19** enlaza *calculus* con **matlab**; también **L11,L13,L22** ligan *differential* con **equations**, y **L28** engancha *differential* con **equations** y **matlab**. Parece haber más vías. Véase el digrafo de la Figura 9.

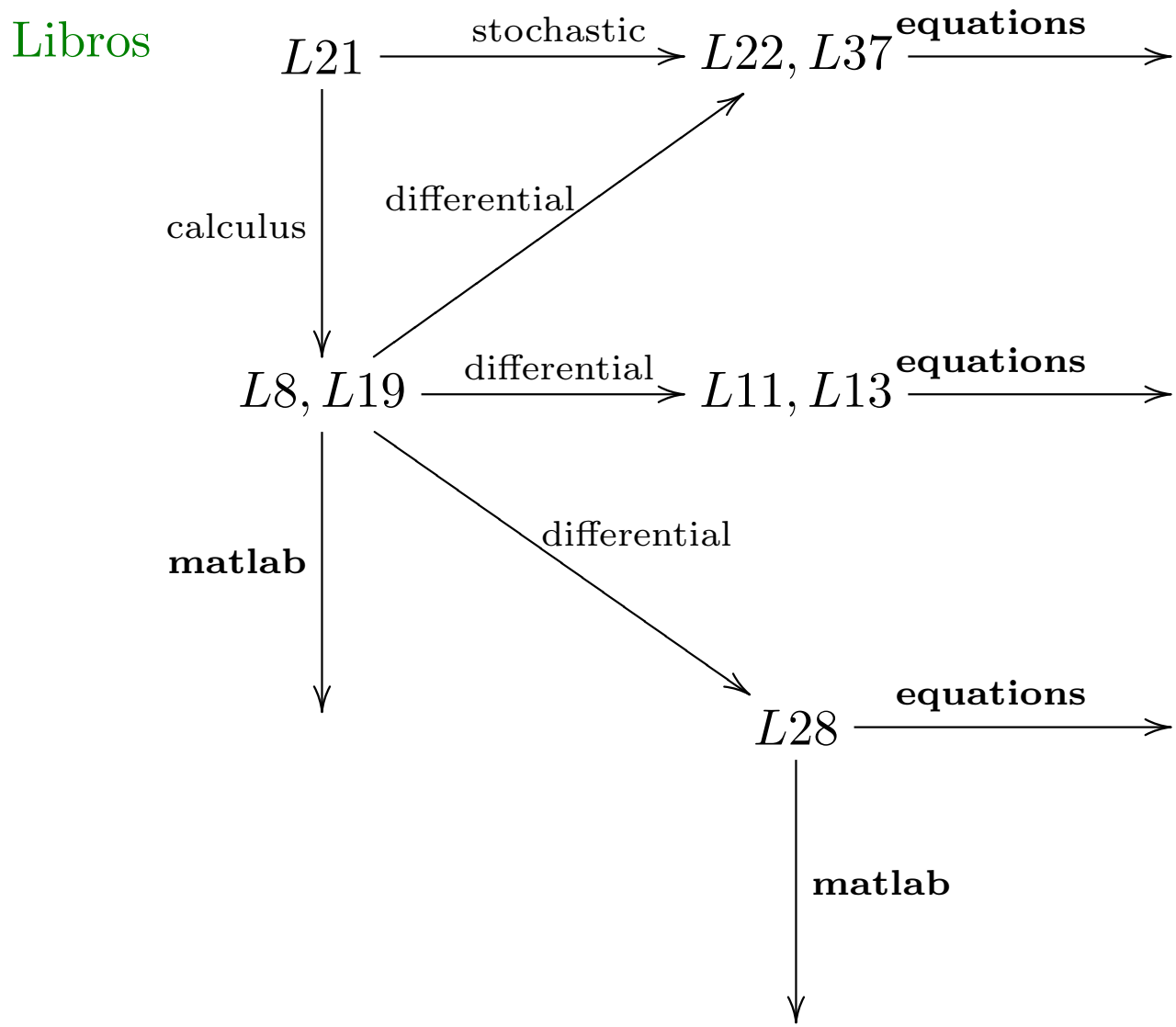


Figura 9: Digrafo L21.

La única palabra "significativa" en **L30** es *problem*; a través de **L14** que contiene *differential,equations* y *problem*s, el análisis semántico latente ha podido hacer esta asociación de **L30** con **L28**; éste libro conecta a **L30** con **equations** y **matlab**. Existen otros caminos. Véase el digrafo de la Figura 10.

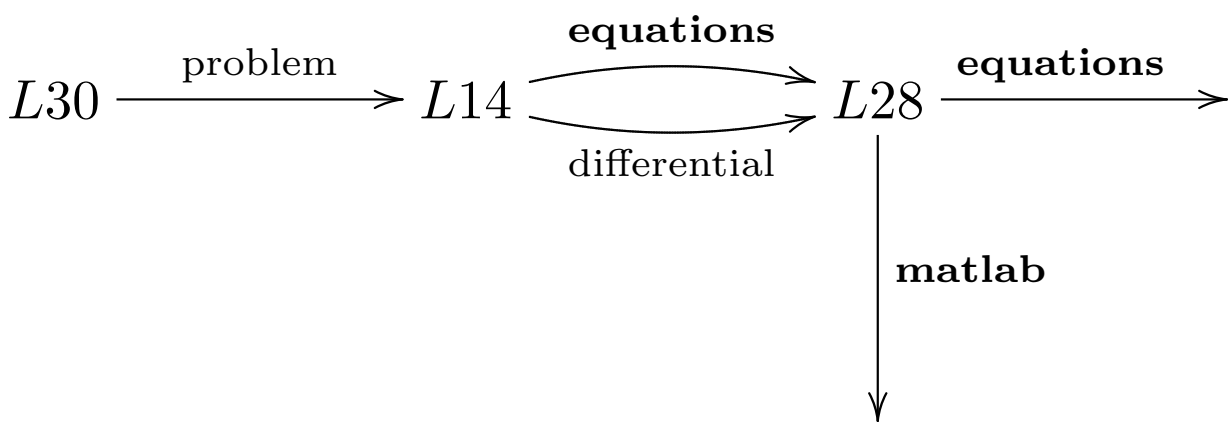


Figura 10: Digrafo L30.

- **Búsqueda literal**

Si en el ejemplo que precede hubiéramos hecho una búsqueda literal, solo habría sido devuelto el libro **L28**. Esto es así pues ninguno de los 37 libros

restantes contiene a la vez las palabras *equations* y *matlab*. La búsqueda semántica ponderada con umbral 0,70 devuelve **L28** y ocho libros más.

Las 20 palabras o términos claves están representadas en el plano x, y, por sus coordenadas ponderadas en la base $\{\sigma_1 \boldsymbol{v}_1, \sigma_2 \boldsymbol{v}_2\}$, del subespacio $\operatorname{Im} \boldsymbol{A}_2^{\mathrm{T}}$, en las Figuras 12 y 13.

3.8. Software "on line"

Pueden hacerse pruebas en directo del Análisis Semántico Latente con el sistema RUCIO en la dirección URL http://193.146.10.159/rucio/index.html.

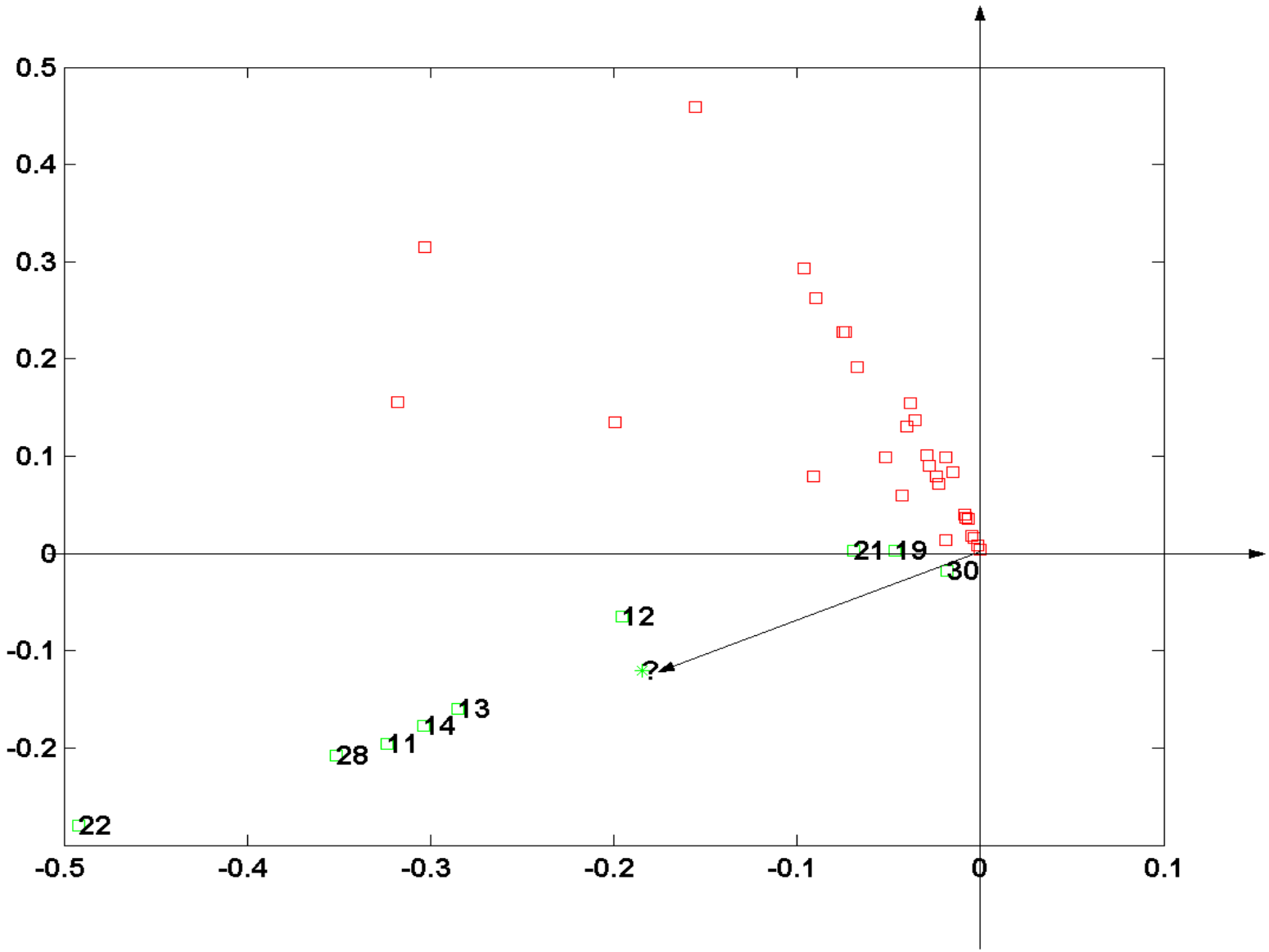


Figura 11: 0,70-respuesta ponderada a la consulta *equations y matlab*.

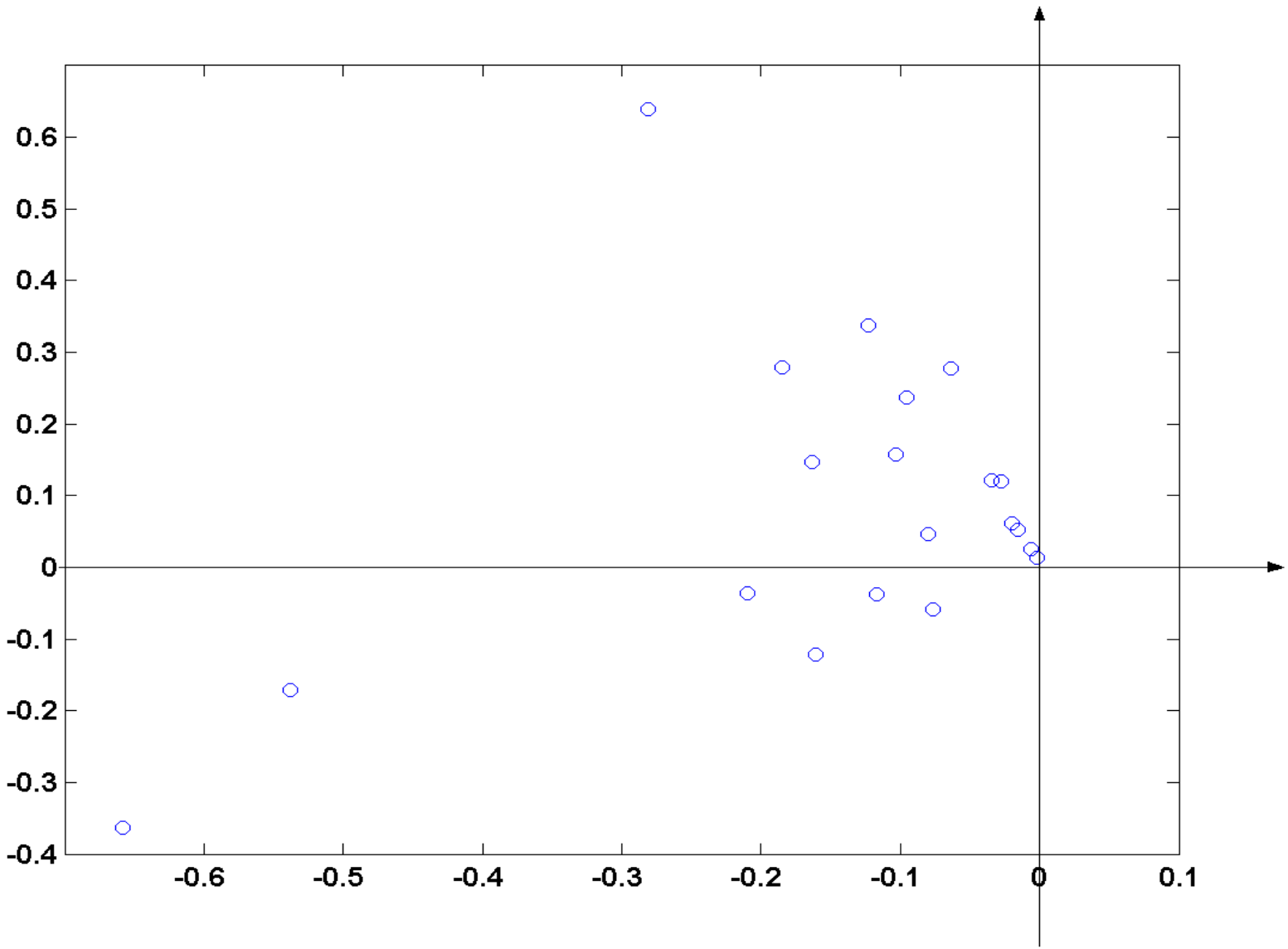


Figura 12: Términos dados por sus coordenadas ponderadas.

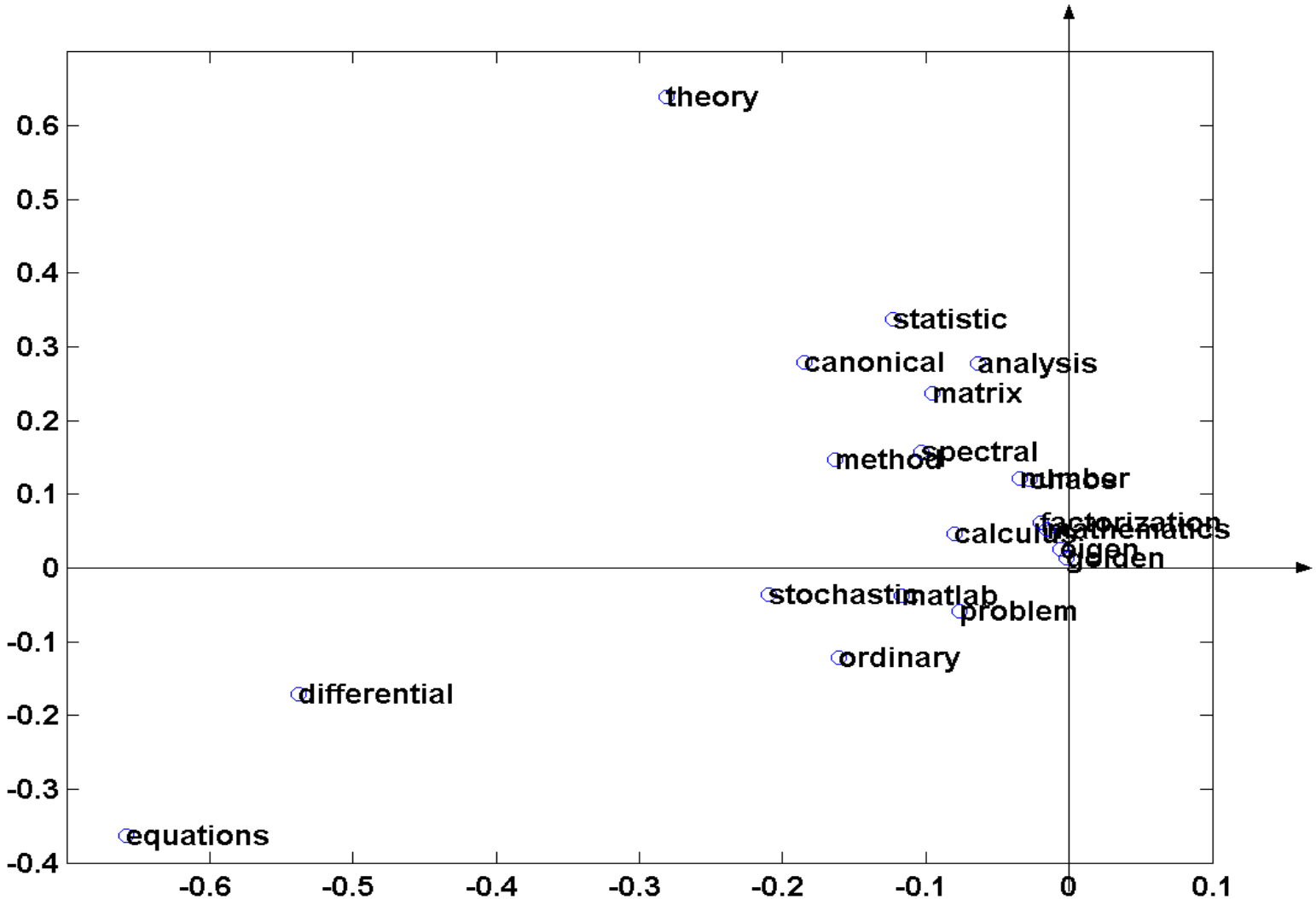
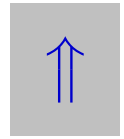


Figura 13: Descripción de los términos.



4. Elaboración automática de tesauros

Un tesauro es un diccionario de sinónimos. En general, estos diccionarios están hechos por humanos. Pero la construcción automatizada de tesauros ha empezado. Una aproximación podría ser consecuencia de la aproximación de la matriz $\boldsymbol{A}$ de *términos* $\times$ *documentos* mediante la matriz de rango inferior $\boldsymbol{A}_k$. Denotemos por $\boldsymbol{f}_1^{(k)}, \ldots, \boldsymbol{f}_m^{(k)}$ las filas de $\boldsymbol{A}_k$. De este modo, si T_i es un término que figura en la base de datos a la que está asociada $\boldsymbol{A}$, veamos cómo podríamos buscar sus sinónimos en el universo semántico dado así. Tomando un número umbral de tolerancia $0 < \tau < 1$ llamaríamos $(\tau; k)$-sinónimos de T_i a todos los términos T_ℓ tales que

$$\cos(\boldsymbol{f}_i^{(k)}, \boldsymbol{f}_\ell^{(k)}) > \tau.$$

Por ser $\cos(\boldsymbol{f}_i^{(k)}, \boldsymbol{f}_i^{(k)}) = 1$, se tiene que T_i es $(\tau; k)$-sinónimo de sí mismo, para todo τ.

Si en vez de utilizar el coseno ordinario **cos**, usamos el coseno ponderado por filas **cospf** en la definición anterior, a los elementos así definidos los llamaremos $(\tau; k)$-sinónimos ponderados de T_i.

Siguiendo con el Ejemplo 3.7 vemos que en la Figura 14 están represen-

tados los seis (0,70;2)-sinónimos ponderados de la palabra *equations*.

4.1. Polisemia de una palabra

Sabemos que la polisemia de una palabra es la variedad de distintos significados que tiene. Por ejemplo, **tarifa** es una palabra polisémica: Tarifa ciudad, tarifa de precios. La polisemia de una palabra es tanto mayor cuanto más grande sea el número de significados diferentes que encierra. Convendremos en llamar *homónimos* a los diferentes significados de *una misma palabra*. ¿Puede automatizarse la búsqueda automática de los diversos conceptos dados por una palabra? Sea T_i un término (o palabra) dado, y sea $\boldsymbol{d}_j$ un documento que precisa uno de los significados concretos de T_i; podemos decir que $\boldsymbol{d}_j$ es uno de los conceptos referidos por T_i. Los *otros* significados de T_i vienen dados por los otros documentos $\boldsymbol{d}_k$ que solo tienen la palabra T_i en común con $\boldsymbol{d}_j$. Esto puede hacerse mediante el producto $\boldsymbol{d}_j .* \boldsymbol{d}_k$, componente a componente (notación de MATLAB). Retenemos $\boldsymbol{d}_k$ si $\boldsymbol{d}_j .* \boldsymbol{d}_k$ tiene un uno en el lugar i y ceros en las demás componentes.

Un "significado" de T_i se corresponde biunívocamente con **uno** de los "grupos de documentos"(campos semánticos) que contienen a T_i.

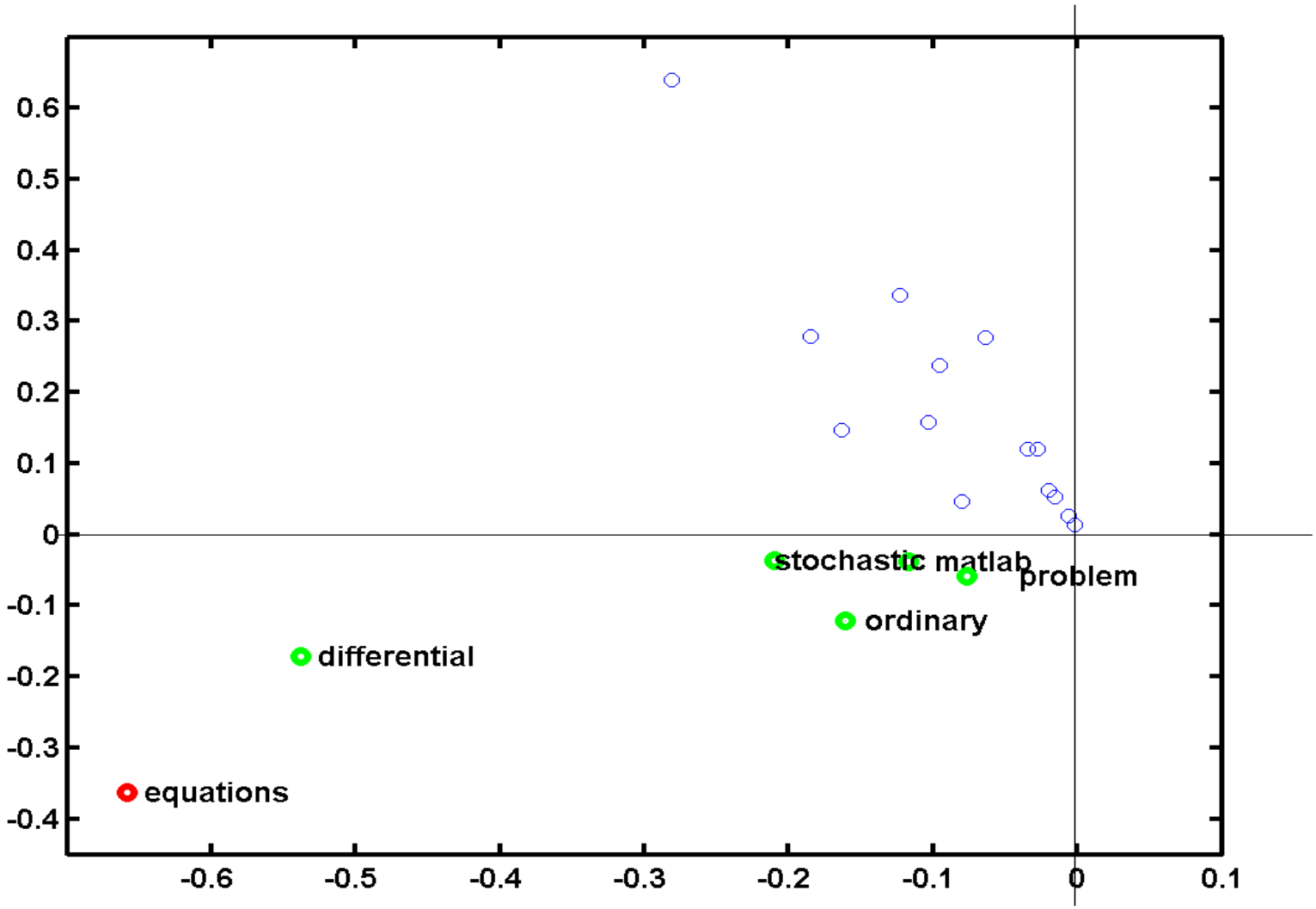
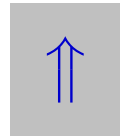


Figura 14: (0,70;2)-sinónimos ponderados de *equations*.

Otra manera de confección automática de tesauros ha sido considerada en la literatura. Consiste en una particularización de la idea del"Pagerank" a un determinado tema. Si t es un tema o término en el que ponemos nuestra atención, por ejemplo, *coche*, podemos dar un valor $R_t(u)$ a cada página web u en términos de los enlaces de páginas que contienen la palabra t (coche) y que apuntan hacia ella. En las páginas mejor R_t-valoradas estarían los sinónimos de t.



5. Matemáticas suscitadas por Internet

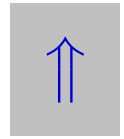
Es claro que Internet está siendo una fuente de problemas matemáticos, y de subsiguientes desarrollos de las Matemáticas. A modo de ejemplo, además de las cuestiones planteadas en este artículo, citemos:

- **Reorganización** más eficiente de un conjunto de páginas web, que arrancan de una `HomePage.html` o de una `index.html`. La referencia [1] trata de entropías, cladogramas, árboles, estructuras de datos, etc.

- **Comunicación de las matemáticas** a través de Internet. El lenguaje `OpenMath`. Software de álgebra simbólica en la Red. Búsqueda en documentos con extracción automática del significado semántico de signos y fórmulas matemáticos. Razonamiento automático con los correctores de demostraciones ("proof checkers"). Véase [5].

- **TCP**: Ecuaciones diferenciales, funciones de Lyapunov, etc. Véase [7].

Un problema de porcentajes relacionado con el protocolo `FTP`.



6. Acreditaciones

Las Figuras 1 y 2 han sido tomadas del artículo [12]. Los cálculos han sido realizados con MATLAB.

6.1. Agradecimientos

El autor agradece a José María González de Durana la lectura de este documento y por avisarle de algunas erratas.

Referencias

[1] D. J. Aldous. Reorganizing large web sites. *Amer. Math. Monthly*, 108 (1):16–27, Jan. 2001. 78

[2] P. Arciniega and J.-M. Gracia. On an iteration relative to root and vector of Perron. Unpublised Report, The University of the Basque Country, P.O. Box 450, 01080 Vitoria-Gasteiz, Spain, 1988. 5

[3] M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM J. Matrix Anal. Appl.*, 1994. 35

[4] M. W. Berry, Z. Drmač, and E. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362, 1999. 35

[5] A. M. Cohen. Communicating mathematics across the web. In B. Engquist and W. Schmid, editors, *Mathematics Unlimited– 2001 and Beyond*, pages 283–300. Springer, Berlin, 2001. 78

[6] G. Golub and C. van Loan. *Matrix Computations*. The Johns Hopkins University Press, second edition, 1989. 41

[7] F. Kelly. Mathematical modelling of the internet. In B. Engquist and W. Schmid, editors, *Mathematics Unlimited– 2001 and Beyond*, pages 685–702. Springer, Berlin, 2001. 78

[8] P. Lancaster and M. Tismenetsky. *The Theory of Matrices with Applications*. Academic Press, second edition, 1985. 11

[9] J. Martín. Googlandia. *Ciberp@ís Mensual*, (17):25–29, diciembre 2001. Madrid. 5

[10] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2000. 35, 41, 46, 48, 49

[11] K. K. Nambiar. Theory of search engines. URL http://www.rci.rutgers.edu/~kannan/computer_technology.html. 2001. 27

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, 1998. 8, 79

[13] D. Rafiei and A. O. Mendelzon. What is this page known for? computing web page reputations. Technical report, Department of Computer

Science, University of Toronto, 2000. URL http://www.cs.toronto.edu/~mendel/papers.html. 18

Sobre este documento

Este artículo ha sido escrito en LaTeX con ayuda del paquete web, escrito por D.P. Story. Véase http://www.math.uakron.edu/~dpstory/acrotex.html. Después el fichero fuente busca.tex ha sido compilado con pdflatex.