

Metodología para la Representación Gráfica de Relaciones Cualitativas entre Tópicos de Interés

F. de la Rosa, L. González, R. M. Gasca, F. Velasco
Universidad de Sevilla

ffrosat@lsi.us.es, luisgon@us.es, gasca@lsi.us.es, velasco@us.es

Resumen

En este trabajo presentamos una metodología que permite cuantificar y representar gráficamente el grado de similitud entre diferentes tópicos de interés, de acuerdo a una interpretación intuitiva de la similitud entre conjuntos. Para ello utilizamos una función procedente de la teoría del aprendizaje, que nos permitirá estudiar a partir de las páginas Web, la similitud que existe entre las diferentes líneas de investigación sobre aprendizaje referente al año 2000.

Palabras clave: Relaciones Cualitativas, Similitud, Aprendizaje, Espacios probabilísticos.

1. Introducción

Cuando se comienza un análisis sobre el estado del arte o las bases documentales sobre una determinada área de investigación, se necesita tener una visión global de los tópicos de interés de dicha área, la evolución en el tiempo de los mismos, los tópicos de más importancia documental y las relaciones cualitativas de similitud o no similitud que se establecen entre los tópicos de acuerdo a unos parámetros previamente definidos. Hoy en día, es importante acudir a las bases documentales como la Web para encontrar una primera información de partida para realizar el análisis inicialmente reseñado.

En este análisis nos podemos plantear cuestiones sobre relaciones cualitativas no temporales, tales como: ¿Está el tópico de interés A muy relacio-

nado con el tópico B en la Web?, ¿Es mucho el peso que tiene el tópico C en el área de investigación?, y preguntas sobre relaciones cualitativas temporales como: ¿Se han acercado los tópicos de interés A y B en los últimos años?, ¿Cuál ha sido la evolución del tópico C con respecto a la D a lo largo de los años?. Una adecuada representación gráfica de la información recogida podría facilitar la visualización de las preguntas anteriores.

En disciplinas tales como la Cienciometría, Infometría o Bibliometría se han desarrollado numerosas técnicas para analizar bases de datos documentales, que clasifican y/o representan gráficamente los documentos que estas almacenan. Las técnicas usadas para tratar estas bases de datos documentales se pueden dividir en técnicas estadísticas y neuronales. Las técnicas neuronales utilizan la capacidad de aprendizaje de las redes neuronales, como la red de Kohonen, para reducir la dimensión de los datos bibliométricos, que habitualmente se encuentran representados como vectores en un hiperespacio.

Las técnicas para reducir la dimensión de los datos bibliométricos que utilizan los métodos estadísticos son muy variadas. Podemos destacar entre ellas la técnica de escalamiento multidimensional (MDS)[3][4] que comenzaron a ser desarrollados a finales de los años 60 en el área de la psicopsiquica para analizar las percepciones psíquicas entre distintos individuos. Desde un punto de vista estadístico estas técnicas están incluidas dentro del análisis de las técnicas multivariantes, donde las variables estadísticas están representadas por medio de un modelo lineal. Otra técnica estadística es el análisis

de componentes principales que busca el mejor modelo lineal con el propósito de proyectar los datos, minimizando la pérdida de información.

Otro método estadístico de interés es el de las "palabras asociadas" [5][6]. Este método se basa en la construcción de un grafo, donde los nodos representan las palabras clave y los arcos contienen las frecuencias de aparición en los documentos de las palabras clave que relacionan. A partir de este grafo se encuentran y representan las distintas materias de interés que están implícitas en los documentos.

En este trabajo presentamos una metodología que permite extraer información sobre las similitudes entre las distintas materias de interés de un área, a partir de los datos de una Base de datos documental, como puede ser la Red Internet, para posteriormente realizar una representación gráfica de esta información que nos permita responder a las preguntas cualitativas que nos planteamos al inicio de la introducción. Para finalizar el artículo presentamos un ejemplo práctico de la aplicación de la metodología basándonos en la representación gráfica de la función núcleo de similitud presentada en el trabajo [8].

2. Metodología para la Construcción de Mapas Temáticos

La metodología que presentamos en esta sección describe los procesos que son necesarios seguir para obtener una serie de mapas temáticos temporales. En la descripción de la metodología se destacan tres elementos fundamentales:

- **Etapas:** Definen una serie de procesos que se ejecutan en bloque y que transforman y/o extraen información.
- **Almacenes:** Se asemejan a contenedores de productos, en nuestro caso información de entrada y/o de salida que las etapas consumen y/o producen.
- **Interfaces:** Son una clase especial de etapas que se caracterizan por interactuar con el usuario para recibir o transmitir información.

En la figura 1 se presenta el proceso de transformación que sufre la información al utilizar la metodología. En esta figura se distinguen las siguientes etapas de la metodología:

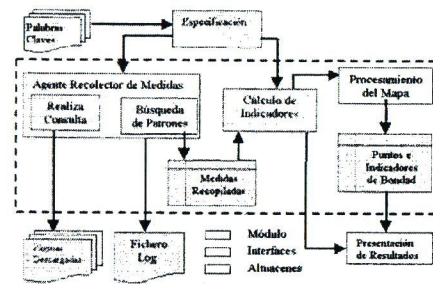


Figura 1: Esquema de una metodología para la construcción de Mapas Temáticos

- **Especificación:** Es la etapa encargada de recopilar los parámetros, que permiten al usuario definir como algunos de los procesos de la metodología van a transformar la información y donde se permite definir:
 - Los conceptos o tópicos de interés con que trabajará el mapa y sus atributos: conjunto de palabras claves, leyendas, colores, etc.
 - Como construir las consultas: temporalidad, buscador utilizado, etc.
 - Indicar como calcular los índices de Afinidad o Impacto: fórmulas, patrones a buscar, medidas a recopilar, etc.
 - Indicar los periodos, si el mapa representa una serie temporal.

Algunas de las características configurables de la metodología, como pueden ser las series temporales o el tipo de cálculo utilizados en los índices, dependerán de las posibilidades que ofrezcan los recursos utilizados.

- **Obtención de Medidas:** Es el proceso de obtención de medidas configurado en la etapa anterior y se encarga de consultar los recursos y de extraer las medidas necesarias para estimar los recursos. En el contexto de la Red Internet, esta etapa es realizada por un agente software, que se encarga de realizar las operaciones que una persona realizaría comúnmente como: consultar un buscador, descargar las páginas, comprobar en las páginas que aparece la información que buscamos y extraer de las

páginas medidas. Seguidamente exponemos el pseudocódigo del agente software:

```

T := {Consultas (Buscador, Palabras Claves,
Operación, Periodo)}
mientras T ≠ {Conjunto Vacío}
  consulta = primer_elemento(T)
  T := T - {consulta}
  pagina_html := descarga(url(consulta))
  numero_paginas := BuscaPatrones(pagina_html)
  inserta_registro(datos(consulta), numero_paginas)
  inserta_linea_log(patron_detectado, pagina, numero_paginas)
  retardo
fin mientras
  
```

Figura 2: Agente Software.

- **Cálculo de Indicadores:** Los procesos de esta etapa calculan la información previa necesaria para la construcción de las representaciones gráficas, planteando el cálculo de dos indicadores como la abstracción de esta información. El primer indicador es la Afinidad, que estima la similitud entre dos materias de interés y el segundo indicador es el Impacto que estima la importancia de la materia de interés en comparación con las demás. El método utilizado para estimar los índices es independiente de la metodología y las posibles estimaciones de los indicadores depende en gran medida de las posibilidades ofrecidas por los recursos utilizados.
- **Procesamiento de Mapas:** Esta etapa recibe los índices calculados en la etapa anterior y sobre ellos se aplica la técnica definida en la metodología para calcular la información necesaria para representar gráficamente los datos.
- **Presentación de Resultados:** Es la última etapa de la metodología y presenta al usuario toda la información recopilada, utilizando un paradigma gráfico. Esta etapa permitirá representar dos tipos de paradigmas según presente o no la evolución en el tiempo.

Aunque esta metodología se presenta en [7], en la siguiente sección resaltamos como aspecto metodológico de interés el proceso de abstracción sufrido por el índice de Afinidad, modelado mediante una matriz cuadrada donde el elemento de la fila i y la columna j referencia la afinidad entre las materias de interés i -ésimo y j -ésimo, ya que dicha matriz es compartida por las distintas etapas para transformar la información.

3. Aplicación de la Metodología a un Caso Práctico

En esta sección presentamos una aplicación de la metodología para realizar un estudio sobre las relaciones que existen entre varias áreas que perteneciente a la Teoría del Aprendizaje.

3.1. Especificación de la Metodología

En esta etapa definimos las materias de interés del Área a estudiar. Para conseguir nuestros objetivos definimos las materias en que estamos interesados, mediante un conjunto de palabras clave. La elección de estas palabras clave se realiza a partir de las inicialmente propuestas en los artículos publicados en varias actas de Congresos dedicados a la Inteligencia Artificial. En concreto, se estudia la evolución de nueve líneas de investigación donde cada una de estas líneas se corresponde con nuestras materias de interés que son denotados por A_i :

Items	Líneas de Investigación
A_1	Reinforcement Learning
A_2	Learning for information Retrieval
A_3	Automated Learning
A_4	Automated Discovery
A_5	Hyb. Syst., Neural and Symb. Process.
A_6	SVM, Dec. Trees and Dec. Trees Gr.
A_7	Machine Learning
A_8	Neural Network
A_9	Data Mining

En esta etapa se indican los recursos utilizados para estimar los índices de Afinidad e Impacto y como van a ser calculados. Hemos de hacer notar que en este estudio hemos utilizado para estos cálculos la información obtenida a través del buscador ALTA-VISTA.

3.2. Proceso de Obtención de Medidas

Una vez definidos las materias de interés, utilizamos un agente software que, a partir de las palabras clave, se encarga de consultar el buscador ALTA-VISTA, a fin de obtener los listados de las páginas Web que tratan a la vez sobre una de las parejas de las materias investigadas. Una vez recuperada las consultas, el agente extrae el número de páginas en las que aparecen los dos items, que denotamos por n_{ij} . De un total de 250 000 páginas rastreadas, se obtuvieron las siguientes secuencias:

	A ₁	A ₂	A ₃	A ₄	A ₅
A ₁	1	0	0	0	0
A ₂	0	75	1	5	2
A ₃	0	1	679	16	6
A ₄	0	5	16	1434	24
A ₅	0	2	6	24	11417
A ₆	0	13	93	141	161
A ₇	0	60	185	457	490
A ₈	0	26	90	210	829
A ₉	0	24	268	442	490

	A ₆	A ₇	A ₈	A ₉
A ₁	0	0	0	0
A ₂	13	60	26	24
A ₃	95	185	90	268
A ₄	144	457	210	442
A ₅	162	490	829	490
A ₆	16984	6271	4474	5372
A ₇	6271	41001	7581	9604
A ₈	4474	7581	85706	6566
A ₉	5372	9604	6566	122970

Por su propia naturaleza, la tabla ha de ser simétrica, pero en caso de no serlo, hemos tomado como valor de referencia el mayor de los dos items.

3.3. Cálculo de los Índice de Afinidad

Una vez extraídas las medidas, número de páginas que tratan sobre una pareja de temas, utilizamos la función núcleo de similitud:

$$k(A, B) = P(A \cap B) - P(A) \cdot P(B)$$

propuesta en [1] y [8], para calcular el índice de Afinidad. Esta función tiene sus orígenes en la Teoría del Aprendizaje Estadístico (ver [2]) y mediante una interpretación frecuencialista de la probabilidad, permite suponer que $\frac{n_{ij}}{N} = P(A_i \cap A_j)$. Utilizamos entonces la función núcleo similitud, para poder transformar las secuencias anteriores en la siguiente tabla:

	A ₁	A ₂	A ₃	A ₄	A ₅
A ₁	0.0000	-0.0000	-0.0000	-0.0000	-0.0000
A ₂	-0.0000	0.0003	0.0000	0.0000	-0.0000
A ₃	-0.0000	0.0000	0.0027	0.0000	-0.0001
A ₄	-0.0000	0.0000	0.0000	0.0057	-0.0002
A ₅	-0.0000	-0.0000	-0.0001	-0.0002	0.0436
A ₆	-0.0000	0.0000	0.0002	0.0002	-0.0025
A ₇	-0.0000	0.0002	0.0003	0.0009	-0.0055
A ₈	-0.0000	0.0000	-0.0006	-0.0011	-0.0123
A ₉	-0.0000	-0.0001	-0.0003	-0.0011	-0.0205

	A ₆	A ₇	A ₈	A ₉
A ₁	-0.0000	-0.0000	-0.0000	-0.0000
A ₂	0.0000	0.0002	0.0000	-0.0001
A ₃	0.0002	0.0003	-0.0006	-0.0003
A ₄	0.0002	0.0009	-0.0011	-0.0011
A ₅	-0.0025	-0.0055	-0.0123	-0.0205
A ₆	0.0633	0.0139	-0.0054	-0.0119
A ₇	0.0139	0.1371	-0.0259	-0.0423
A ₈	-0.0054	-0.0259	0.2253	-0.1424
A ₉	-0.0119	-0.0423	-0.1424	0.2499

Aunque la medida utilizada, número de páginas que trata sobre una pareja de items obtenida de ALTAVISTA, es una medida débil de la similitud, tiene la ventaja de que es difícil encontrar un buscador que no la acepte. Aún así, se puede mejorar la medida discriminando con pesos las distintas categorías de páginas Web recuperadas en las consultas, así por ejemplo, se podría dar más importancia a los enlaces que pertenezcan a dominios de tipo ".edu" que a los ".com".

3.4. Procesamiento del Mapa Temático

Si fijamos un suceso A, y consideramos la función definida en [0, 1]: $k_A(P(B)) = k(A, B)$, entonces la representación gráfica del dominio y recorrido de la función $k_A(P(B))$, es el área encerrada por el paralelogramo que aparece representado en la figura 3, donde, además, se ha indicado en que situaciones se alcanzan los lados de la figura.

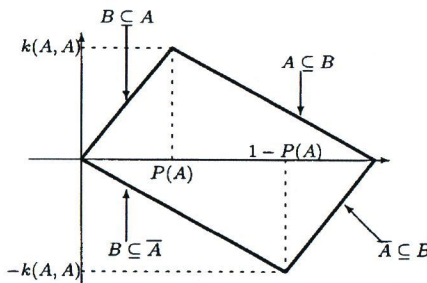


Figura 3: La región encerrada representan el dominio y recorrido de la función $k_A(P(B))$, con A fijado.

Ahora bien, ¿conociendo las similitudes $k(A, B)$ y $k(A, B')$ se puede realizar alguna afirmación sobre $k(B, B')$? La respuesta es negativa ya que hay dos

situaciones que nos lo aclara: i) Si $A = \emptyset$ y $B = B'$ con $P(B) = 1/2$ entonces $k(A, B) = k(A, B') = 0$ y sin embargo $k(B, B') = 1/4$; y ii) Si tomamos $A = \emptyset$ y $B = \bar{B}'$ con $P(B) = 1/2$, entonces se sigue que $k(A, B) = k(A, B') = 0$ y sin embargo $k(B, B') = -1/4$ (ver [1] para una mayor aclaración).

Con las propiedades de la función núcleo similitud, se ve en [1] como se comporta esta función respecto a distintas formas de los conjuntos A, B y B'. Existen otras posibilidades de combinar los conjuntos, como por ejemplo las dadas en la figura 4. Estos casos son igualmente interesantes, sin

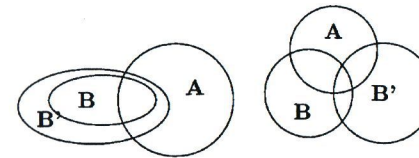


Figura 4: Dos ejemplos de posibilidades que se pueden presentar entre los conjuntos A, B y B'

embargo, no es posible realizar ninguna afirmación sobre sus similitudes (ver [1]).

Por todo ello, la única afirmación que se puede realizar es la siguiente: Si $P(A) \simeq P(B_1) \simeq P(B_2)$ y $k(A, B_1) \simeq k(A, B_2)$ entonces $k(A, B_1) \simeq k(A, B_2) \simeq k(B_1, B_2)$. Entonces si trabajamos con un conjunto de sucesos, $\{A_i\}_{i=1}^n$, elegimos como suceso A, de referencia en la función k_A , aquel que cumpla la condición

$$d(A) = \min_i \sum_{j=1}^n (P(A_i) - P(A_j))^2,$$

con objeto que cuando sea posible se pueda aplicar la afirmación anterior (ver [1] u [8]).

Consideramos entonces la función $k_A(P(B))$ con A el óptimo definido anteriormente, siendo éste $A = A_7$, y representamos gráficamente esta función en la figura 5.

3.5. Interpretación de los Resultados

Para interpretar esta gráfica, hacemos notar que los items se encuentran ordenados en función de sus respectivas probabilidades. En esta figura, salvo A_6 que presenta una similitud pequeña y positiva, las restantes líneas de investigación presentan una similitud negativa, en especial los items A_5, A_8 y A_9 .

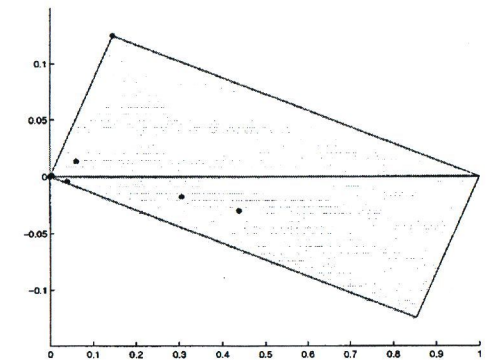


Figura 5: Similitudes $k_{A_7}(P(A_i))$ con $i = 1, \dots, 9$.

Todo ello nos permite declarar que no existe una alta similitud entre el suceso A_7 y las restantes líneas, es decir, la línea "Machine Learning" presentaba en el año 2000 un conjunto de páginas Web que no tenían muchos nexos con las restantes páginas.

Hemos de notar que en este gráfico solo es posible observar como se comportan los restantes items con respecto al que sirve de referencia. Por ello, si queremos realizar un análisis conjunto hemos de realizar tantas gráficas como items estemos estudiando.

Así, en la figura 6 podemos ver la posición de los

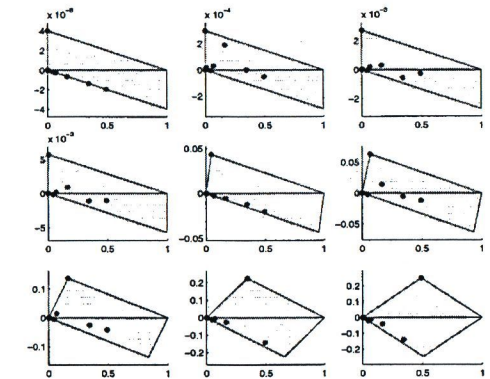


Figura 6: Representación gráfica de cada línea frente a las demás.

valores $k_{A_i}(A_j)$ variando A_i desde $i = 1$ hasta 9, con respecto a todos los restantes items. Del estudio de estas gráficas una a una se observa lo siguiente:

- Gr1.- Todas las líneas presentan la máxima disimilitud con "Reinforcement Learning", lo cual resulta evidente desde el primer momento, puesto que el número de referencias de este ítem es casi nulo. Se tiene entonces que es una línea muy poco desarrollada y apartada totalmente de los desarrollos de las restantes.
- Gr2.- Representa las similitudes con la línea "Learning for Information Retrieval". Lo más destacable es la ausencia de similitud con las redes neuronales (A_8) y su alta similitud con "Machine Learning" (A_7).
- Gr3.- Lo más destacable es la alta disimilitud entre "Automated Learning" y las redes neuronales.
- Gr4.- Similar al anterior.
- Gr5.- Todas las líneas presentan casi la máxima disimilitud con "Hybrid Systems, Neural and Symbolic Processing", por ello, se indicaría que es una línea apartada de los desarrollos de las restantes.
- Gr6.- La interpretación, referente a "Support Vector Machine, Decision Trees and Decision Trees Graphics", es similar a la dada en el gráfico 3.
- Gr7.- Línea "Machine Learning". Lo más resaltante es que presenta un grado intermedio de disimilitud con las otras dos grandes líneas de investigación (grandes en el sentido dado por sus probabilidades).
- Gr8.- Se observa que "Neural Network" presenta una alta disimilitud con "Data Mining".
- Gr9.- Igual comentario que el anterior gráfico.

A partir de estas gráficas podemos concluir que debido a la naturaleza, todavía joven, de estos estudios sobre Teoría del Aprendizaje, no existen dos líneas de investigación que presenten un alto grado de similitud en el sentido de llegar a tener un conjunto de identificadores comunes.

3.6. Representación gráfica global

El principal defecto que presenta la representación gráfica dada en la figura 6 es que necesita tantas gráficas como ítems se estudia. Evidentemente, sería mucho más práctico disponer de una única representación gráfica donde aparezca recogida toda

la información proporcionada por los gráficos anteriores. Para conseguir este fin, proponemos la siguiente solución.

En primer lugar consideramos los sucesos $\{A_1, A_2, \dots, A_n\}$ ordenados a partir de sus probabilidades, es decir, $0 \leq P(A_1) \leq \dots \leq P(A_n)$, y se representan estas probabilidades sobre el eje de abscisas¹. Sobre el eje de ordenadas representamos las similitudes de la siguiente forma: Se toma el primer suceso A_1 y se representan el conjunto de similitudes $\{-k(A_1, A_1), k(A_1, A_2), \dots, k(A_1, A_n), k(A_1, A_1)\}$ con abscisa $P(A_1)$, a continuación se toma el suceso A_2 y se representan el conjunto de similitudes $\{-k(A_2, A_2), k(A_2, A_3), \dots, k(A_2, A_n), k(A_2, A_2)\}$ con abscisa $P(A_2)$, y así sucesivamente hasta el suceso A_n . La explicación del por qué se actúa de esta forma viene motivado por los desarrollados dados en [1]. Si se tiene que $P(A_n) \leq 1/2$, de la ordenación de las probabilidades y del crecimiento de la función $f(x) = x(1-x)$ en $(0, 1/2)$, se deduce que

$$|k(A_i, A_j)| \leq k(A_i, A_i) \quad \text{si } i < j$$

lo cual valida la construcción realizada. Por otro lado, si existe algún suceso con probabilidad superior a $1/2$ se actúa sobre su complementario, gracias a las propiedades de la función núcleo similitud y de las conclusiones sobre el complementario desarrolladas en [1]. La representación gráfica, siguiendo este proceso aplicada a las anteriores líneas de investigación, puede verse en la figura 7. En esta gráfica el símbolo \times y el número al lado representa la similitud entre el ítem que aparece en la abscisa y el ítem que se referencia con el número. Se observa que entre los ítems con mayor probabilidad (ítems 6, 7, 8 y 9) no existe mucha similitud, con lo que podemos concluir que estas líneas de investigación siguen caminos distintos con pocas interrelaciones.

También, en la figura 7, y debido a que los ítems 1, 2, 3, 4 y 5 tienen una probabilidad pequeña, no se observan con nitidez, por ello si queremos verlo con más claridad hemos de realizar una ampliación dentro de la figura 7 como la que se tiene la figura 8. En este gráfico, al igual que ocurría con el anterior, la pequeñez de las probabilidades de los primeros ítems hace no visible su representación, pero se puede ver que la similitud entre ellos es despreciable; y en consecuencia declarar que todas las líneas de investigación estudiadas dentro de la teoría del

¹De esta forma se está indicando el peso que cada una de las materias tiene con respecto a las demás.

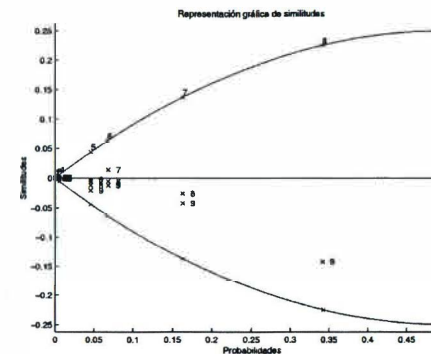


Figura 7: Gráfica de todas las similitudes.

aprendizaje seguían caminos diferentes con pocos nexos de unión en el año 2000.

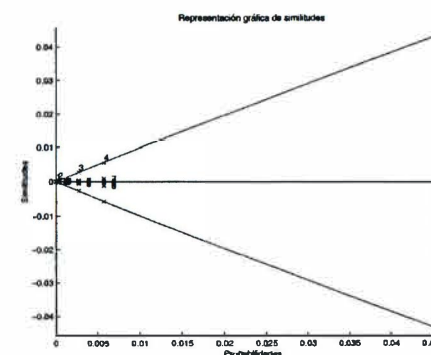


Figura 8: Gráfica con las cinco primeras líneas.

Sin embargo, ¿cómo se vería en los dos últimos gráficos, que dos ítems son muy similares o muy disimilares? Para ello, si la similitud entre ellos es positiva, basta con dibujar el triángulo formado por $\Delta_1 = \{k(B, B), k(A, B), k(A, A)\}$, y si el área encerrada en el triángulo es pequeña entonces se tiene que son sucesos de tamaño parecido y similitud alta. De igual manera, si la similitud entre ellos es negativa, basta con dibujar el triángulo formado por $\Delta_2 = \{-k(B, B), k(A, B), -k(A, A)\}$, y si el área encerrada en el triángulo es pequeña, entonces se tiene que son sucesos de tamaño parecido y disimilitud alta. En general, a medida que el área encerrada por el triángulo Δ_1 (si la similitud conjunta es

positiva) o el triángulo Δ_2 (si la similitud conjunta es negativa) es mayor, entonces los sucesos son más diferentes en tamaño y presentan menor similitud.

4. Conclusiones y trabajos futuros

En este artículo se ha puesto de manifiesto que la utilidad de la similitud entre sucesos puede ser usada tanto en problemas teóricos como prácticos. La elección de dicha función de similitud junto a la metodología aplicada nos ha permitido hacer una representación gráfica de las líneas de investigación de un determinada área y de las relaciones cualitativas que existen entre ellas, sin pérdidas de información, como sucedía en otras técnicas propuestas anteriormente. Como trabajo futuro pensamos realizar un estudio dinámico del comportamiento de las líneas de investigación. De esta manera podremos comprobar cual es el comportamiento que las líneas van teniendo entre ellas a través del tiempo.

Agradecimientos

Este trabajo ha sido soportado en parte por la ayuda ACC-265-TIC-2001 concedida por la Junta de Andalucía.

Referencias

- [1] L. González, *Análisis Discriminante utilizando Máquinas Núcleos de Vectores Soporte. Función Núcleo Similitud*, Tesis Doctoral, Universidad de Sevilla.
- [2] V. Vapnik, *Statistical Learning Theory*, Editorial John Wiley & sons, 1995
- [3] José Eulogio Deus, *Escalamiento Multidimensional*, Cuadernos de Estadísticas, Editorial La Muralla, 2001
- [4] Hansjörg Klock, Joachim M. Buhman, *Data visualization by Multidimensional Scaling: A Deterministic Annealing Approach*, 1997
- [5] Rosario Ruíz-Baños, Francisco Contreas-Cortes, *Como Consultar Eficazmente una Base de Datos Bibliográfica. El Método de las Palabras Asociadas.*, <http://www.ugr.es/fccortes/curriculum/toledo.html>, 1998

- [6] Luc Grivel, Claire François, *A workstation to classikfy, chart and ana Une station de travail pour classer, cartographie et analyser l'information bibliographique dans une perspective de veille scientifique et technique*, Solarion, 1995
- [7] F. de la Rosa T., R.M. Gasca, J.A.Ortega, *Metodología para el Análisis Visual de la Evolución de Conceptos en Base de Datos Textuales*, II Jornadas de Bibliotecas Digitales, 2001
- [8] L. González, F. Velasco, J. M. Alba, R. Gasca, F. de la Rosa, *Aprendizaje en Internet mediante la función Núcleo Similitud*, CCIA02.