ESTUDIO DE SIMILITUDES ENTRE ÁREAS TEMÁTICAS

L. González[1], F. Velasco[1], R. M. Gasca[2] y F. de la Rosa[2]

Resumen: En este trabajo presentamos unos resultados que permiten la representación gráfica de las similitudes entre centros de interés que aparecen en las páginas Web que se encuentran distribuidas en Internet. Para ello se cuantifica y representa gráficamente por medio de un índice de similitud entre los diferentes centros de interés, de acuerdo a una interpretación intuitiva de similitud entre conjuntos. La utilización de una función procedente de la teoría del aprendizaje nos permite estudiar, a partir de la información en las páginas Web, la similitud o interrelación que existe entre diferentes líneas de investigación. El análisis visual de dichas interrelaciones se ha aplicado al área de investigación del aprendizaje referido al período mayo 2003-mayo 2004. Asimismo, se construyen varias tablas que aclaran la potencialidad futura de este índice de similitud.

Palabras clave: similitud, líneas de investigación, visualización, Internet.

Abstract: In this paper some results are presented which permit the graphic representation of the similarities between subjects which appear on web pages found on the Internet. To this end, these similarities are quantified and graphically represented by means of an index between the different subjects, in accordance with an intuitive interpretation of similarity between sets. The use of a function based on Learning Theory, enable us to study the similarity or interrelation between different lines of investigations corresponding to information on web pages. The visual analysis of the aforementioned interrelations has been applied to the area of Learning research refering to the period May 2003-May 2004. Furthermore, various tables which clarify the potential of this similarity index are constructed.

Keywords: similarity, research lines, visualization, internet.

1. Introducción

El gran volumen de información no estructurada que se encuentra dispersa en las bases de datos textuales, de las cuales Internet es un claro ejemplo, hace cada vez más difícil su tratamiento y/o análisis. Con objeto de abordar estos problemas, presentamos en este artículo unos resultados acerca de la similitud entre sucesos que permiten un análisis visual de las similitudes existentes entre centros de interés (se refieren a las áreas de investigación pertenecientes a la Teoría del Aprendizaje) que un usuario determinado desee realizar. La representación visual de las similitudes facilita, en gran medida, su comprensión y el análisis de cuestiones relativas a las relaciones cualitativas no tempo-

[1] Dep. de Economía Aplicada I. Universidad de Sevilla. Correo-e: luisgon@us.es; velasco@us.es.
[2] Dep. de Lenguajes y Sistemas Informáticos. Universidad de Sevilla. Correo-e: gasca@lsi.us.es; ffrosat@lsi.us.es.
Recibido: julio 2002; 2.ª versión: 25-2-05.

rales, tales como: ¿Está el centro de interés A muy relacionado con el B en la Web? ¿Es muy importante el peso que tiene el centro de interés C en el conjunto estudiado? y otras preguntas sobre relaciones cualitativas temporales, tales como: ¿Se han acercado los centros de interés A y B en los últimos años? ¿Cuál ha sido la evolución del centro de interés C con respecto al centro de interés D a lo largo de los años?

Las cuestiones que acabamos de plantear han sido ampliamente estudiadas y han permitido el desarrollo de una amplia gama de métodos estadísticos, denominados técnicas de análisis multivariante que, junto a técnicas de representación de gráficas, han sido adaptadas y desarrolladas en disciplinas tales como Cienciometría (1, 2), Infometría (3, 4), Bibliometría (5, 6) o Webometría (7, 8, 9), para analizar de forma visual las bases de datos documentales. Las técnicas usadas para realizar estas representaciones son conocidas fundamentalmente como técnicas de reducción de la dimensión y se caracterizan por transformar la información almacenada, normalmente en un espacio *n*-dimensional (vector de *n* atributos), en un espacio de dos o tres dimensiones, con el objeto de que la información pueda ser analizada por un observador de forma visual. Uno de los problemas principales de estas técnicas se encuentra en la pérdida de información apreciable que se produce en el proceso de reducción, y que aparece cuando la población en estudio no contiene relaciones significativas entre los distintos individuos.

Las técnicas de reducción de dimensión, utilizadas en las disciplinas relacionadas con el tratamiento de la documentación científica, pueden ser clasificadas en técnicas neuronales y estadísticas. Las técnicas neuronales utilizan la capacidad de aprendizaje de las redes neuronales, como la red de Kohonen (10, 11, 12), a fin de conseguir la reducción de la dimensión de los datos bibliométricos, habitualmente representados mediante vectores en un hiperespacio. Por otro lado, y dentro de las técnicas estadísticas, se puede realizar, a su vez, una subdivisión en técnicas de clasificación o *clustering* y técnicas de reducción de la dimensión. Las técnicas para la reducción de la dimensión de los datos bibliométricos utilizan métodos estadísticos muy variados (13). Podemos destacar, entre ellos, las técnicas de escalamiento multidimensional (MDS) (14, 15) que comenzaron a desarrollarse a finales de los años 60 en el área de la psicopsíquica, para analizar las percepciones psíquicas entre distintos individuos. Desde un punto de vista estadístico estas técnicas están incluidas dentro del análisis de técnicas multivariantes, donde las variables estadísticas se representan mediante un modelo lineal. Otra técnica estadística de reducción es el análisis de componentes principales, que busca el mejor modelo lineal con el propósito de proyectar los datos minimizando la pérdida de información.

El método de las *palabras asociadas* (2, 16, 17, 18, 19, 20) es otra de las técnicas estadísticas de clasificación jerárquica. Este método se basa en la construcción de un grafo, donde los nodos representan las palabras clave y los arcos son las frecuencias de aparición en los documentos de las palabras clave que relacionan, también conocidos como co-ocurrencia de palabras. A partir del grafo, estas técnicas son capaces de encontrar y representar los centros de interés que ocultan los documentos, es decir, zonas de la red muy enlazadas y consistentes, asimilables a «puntos calientes» o «polos de atracción» de gran intensidad informativa (16).

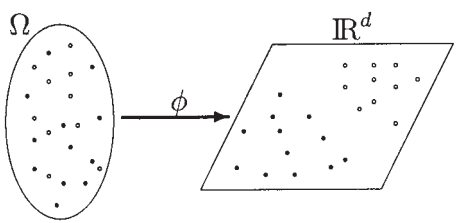
En contraste con las técnicas anteriores, la presentada en este artículo se distingue por su marcado carácter frecuencialista y se fundamenta en la Teoría del Aprendizaje Estadístico (21, 23).

Este artículo se ha estructurado como sigue: en las sección 2 presentamos una nueva técnica de análisis basada en un índice de similitud (24), que permite responder a las preguntas cualitativas planteadas al inicio de la introducción. En la sección 3 mostramos, con un ejemplo práctico, la aplicación de esta técnica para calcular nuestro índice de similitud. En particular y mediante el rastreo de recursos electrónicos con la ayuda de agentes, hemos encontrado las relaciones entre varias líneas de investigación de un área determinada, que hemos presentado en la tabla II y a continuación hemos representado gráficamente los índices entre dichas líneas. De estas representaciones gráficas, obtenidas de los índices de similitud y presentados en la tabla III, hemos construido, además, varias tablas (tabla IV, tabla V) que nos ayudan en la comprensión de este índice y que nos dan a entender su potencialidad en un futuro inmediato, como queda reflejado en la sección 4 de conclusiones y trabajos futuros.

2 Similitud entre sucesos

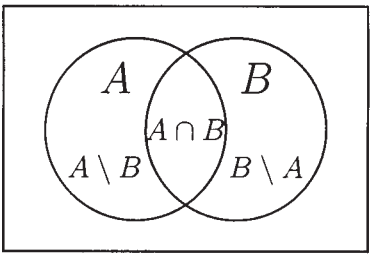
Por *similitud entre objetos* se entiende una medida de correspondencia, o parecido, entre objetos que van a ser estudiados. Esta medida debe tomar en consideración diferentes características de los objetos, según el sentido de *similar* que se quiera dar. Por ello, si se considera un conjunto de objetos Ω, al cual no se le supone ninguna estructura matemática determinada, hemos de encontrar una función tal que a cada par de objetos x, x' de Ω, le asocie un número que cuantifique cuán parecidos son. Si el espacio de trabajo es $\mathbb{R}^d$, esta similitud normalmente viene dada a través de un producto escalar (esta es una de las ideas claves en la Teoría del Aprendizaje Estadístico (21, 22)); por ello, una forma de asignar similitudes entre objetos es incrustar, a partir de una aplicación ϕ, el conjunto Ω dentro de $\mathbb{R}^d$ (ver figura 1).

Figura 1
Función ϕ de Ω en $\mathbb{R}^d$.



De esta forma, si se tiene un conjunto de objetos Ω, dados dos subconjuntos A y B (ver figura 2), intuitivamente se aprecia que éstos son tanto más similares cuanto más pequeño sea el conjunto $(A \setminus B) \cup (B \setminus A)$ (denominado *diferencia simétrica*), es decir, cuantos menos elementos tenga el conjunto A que no tenga B y recíprocamente. La diferencia simétrica es usada para estudiar la similitud entre los subconjuntos A y B y con objeto de llevar a cabo una cuantificación del grado de similitud, suponemos que nos encontramos dentro de un espacio probabilístico, ya que el utilizar una probabilidad P nos permite llevar a cabo una cuantificación de la similitud, además de posibilitar la

Figura 2
Representación de dos conjuntos *A* y *B*



asignación de pesos a los diferentes elementos de los conjuntos. A partir de estas premisas en (23, 24) se lleva a cabo un detallado estudio que conduce a la siguiente medida de similitud:

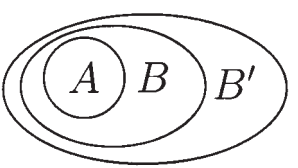
Definición 2.1. *Sea* (Ω, A, P) *un espacio probabilístico. Se define la función núcleo similitud* $k : A \times A \to \mathbb{R}$ *para todo* $A, B \in A$ *como sigue:*

$$k(A, B) = P(\mathrm{A} \cap B) - P(A) \cdot P(B) \tag{1}$$

En (25) aparece recogida por primera vez esta medida de similitud, la cual ha sido estudiada con detalle en (24). Las propiedades más importantes de esta función (para un estudio en detalle, ver (23)) son las siguientes:

1. Si $P(A) = 0$ ó $P(A) = 1$, entonces $k(A, B) = 0$ para todo $B \in \mathcal{A}$.
2. Sea $A \subseteq B \subseteq B'$, entonces: $k(A, B) = k(A, A) - P(A)P(B \setminus A)$ y $0 \leq k(A, B') \leq k(A, B) \leq \leq (A, A)$. Esta propiedad indica que hay menos similitud entre A y B cuanto mayor sea $P(B \setminus A)$. Así, si consideramos un conjunto finito e interpretamos la probabilidad en términos de número de elementos, entonces A y B con $A \subseteq B$ son más similares cuantos menos elementos tenga B que no están en A (ver figura 3). De esta propiedad se sigue que para $A \in \mathcal{A}$ fijado, se verifica $\max_{\{B: A \subseteq B\}} k(A, B) = k(A, A)$.
3. Si A y B son sucesos independientes, entonces $k(A, B) = 0$.

Figura 3
Intuitivamente *A* y *B* son más similares que *A* y *B′*



4. Sea $B \subseteq B' \subseteq \overline{A}$ ($\overline{A}$ es el conjunto complementario de A), entonces $k(A, B) = -k(A, A) + P(A) \cdot P(\overline{A} \setminus B)$ y $k(A, \overline{A}) \leq k(A, B') \leq k(A, B) \leq 0$. La interpretación, a la vista de la figura 4, en términos de conjunto es la siguiente: Como B' tiene más elementos que no están en A que B, su similitud es mayor en valor absoluto. De aquí se sigue que fijado A, se verifica $\min_{\{B: B \subseteq \overline{A}\}} k(A, B) = k(A, \overline{A})$.

Figura 4
Intuitivamente se declara que A y B son más similares que A y B'

5. Dado $A, B \in \mathcal{A}$, entonces $k(A, \overline{A}) \leq k(A, B) \leq k(A, A)$, es decir, el suceso que más similitud tiene con A es el propio A y el que menos es $\overline{A}$.
6. $k(A, \overline{B}) = -k(A, B)$, para todo $A, B \in \mathcal{A}$, es decir, si consideramos que B y $\overline{B}$ son totalmente disimilares, entonces es natural que la similitud entre A y B sea opuesta a la de A y $\overline{B}$.
7. Dado $A, B \in \mathcal{A}$, se tiene

$$|k(A, B)| \leq \min \{k(A, A), k(B, B)\} \leq \frac{1}{4} \tag{2}$$

Esta propiedad proporciona una cota superior ($\frac{1}{4}$) para la similitud y una cota inferior ($-\frac{1}{4}$) para la disimilitud (no similitud) entre sucesos.

Con objeto de tener una representación gráfica basada en la función de similitud se propone la siguiente construcción: Fijado un suceso $A \in A$, sea la función $k_A : \mathcal{A} \to \mathbb{R}$ definida como:

$$k_A(B) = k(A, B) \quad \forall\, B \in \mathcal{A}$$

Es decir, variando $B \in \mathcal{A}$, se tiene una cuantificación de todas las similitudes en relación a A. Por otro lado, y puesto que la función k_A recoge las características del suceso B a través su probabilidad, se considera la función $k_A(P(\cdot)) = k(A, \cdot)$, la cual está definida en $[0, 1]$ y su representación gráfica está dentro del rectángulo $[0, 1] \times [-\frac{1}{4}, \frac{1}{4}]$. Si denotamos por $p = P(A)$, entonces $k_A(B) = P(A \cap B) - p \cdot P(B)$ y teniendo en cuenta la descomposición de B en la forma $B = B_1 \cup B_2$, tal que $B_1 = A \cap B \subseteq A$, $B_2 = \overline{A} \cap B \subseteq \overline{A}$, que por construcción son disjuntos, se tienen las siguientes propiedades.

Si $P(B) = x \leq p$, se tiene que: $k_A(B) = (1 - p) \cdot P(B_1) - pP(B_2)$. Luego si $P(B) = x$, entonces el máximo se alcanza cuando $P(B_2) = 0$ y vale $\max_{P(B) = x} k_A(B) = (1 - p) \cdot x$, alcanzándose cuando el suceso $B \subseteq A$.

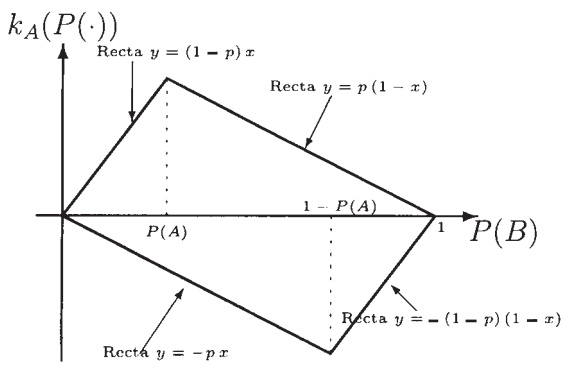
Si $P(B) = x > p$, entonces $k_A(B) = (1 - p) \cdot P(B_1) - pP(B_2)$ y el máximo se alcanza cuando $P(B_1)$ es máximo y $P(B_2)$ es mínimo y como $P(B) = x = P(B_1) + P(B_2)$; esto se consigue si $B_1 = A$, con lo que el máximo de $k_A(B)$ vale $\text{máx}_{P(B)=x} k_A(B) = (1 - p) \cdot p - p \cdot (x - p) = p \cdot (1 - x)$ y se alcanza cuando el suceso B cumple: $A \subseteq B$.

Si $P(B) = x \leq 1 - p$, se sigue que como $k_A(B) = -k_A(\overline{B})$, entonces $\text{mín}_{\{B \setminus P(B)=x\}} k_A(B) = -\text{máx}_{\{\overline{B} \setminus P(\overline{B}) = 1-x\}} k_A(\overline{B}) = -px$, ya que en este caso $P(\overline{B}) = 1 - x \geq p$ y sustituyendo para este caso en el desarrollo anterior, se tiene que el máximo se obtiene en $p(1 - (1 - x))$ y esto se cumple cuando $A \subseteq \overline{B}$. De esta forma, el mínimo se tiene cuando el suceso $B \subseteq \overline{A}$.

Por último si $P(B) = x > 1 - p$, se sigue que como $k_A(B) = -k_A(\overline{B})$, entonces $\text{mín}_{\{B \setminus P(B)=x\}} k_A(B) = -\text{máx}_{\{\overline{B} \setminus P(\overline{B}) = 1-x\}} k_A(\overline{B}) = -(1 - p)(1 - x)$, ya que en este caso $P(\overline{B}) = -1 - x \leq p$, con lo que al sustituir para este caso en el desarrollo anterior, tenemos que el máximo se alcanza en $(1 - p)(1 - x)$ y esto ocurre cuando $\overline{B} \subseteq A$. De esta forma, el mínimo se alcanza cuando el suceso $\overline{A} \subseteq B$.

Por todo ello, la gráfica de la función $k_A(P(B))$ queda dentro del paralelogramo de la figura 5.

Figura 5
Dominio y recorrido de la función $k_A(P(B))$, para A fijo y $P(A) = 0,15$



3 Aprendizaje en la Red

En esta sección aplicamos la sección 2 para medir las similitudes que existen entre varias líneas de investigación pertenecientes a la Teoría del Aprendizaje, las cuales aparecen recogidas en la tabla I. El proceso de selección de éstas ha sido extraído de las distintas secciones en que se dividen las actas de varios congresos representativos dedicados al área en estudio (Inteligencia Artificial). La selección de las palabras clave utilizadas para representar las líneas de investigación han sido escogidas de las presentadas por los autores en los abstracts de los artículos admitidos en las distintas secciones de los congresos, que en nuestro caso se corresponden con las líneas de investigación del área. Así, estudiamos las similitudes existentes dentro de doce líneas de investigación

Tabla I
Relación de doce líneas de investigación relacionadas con el Aprendizaje

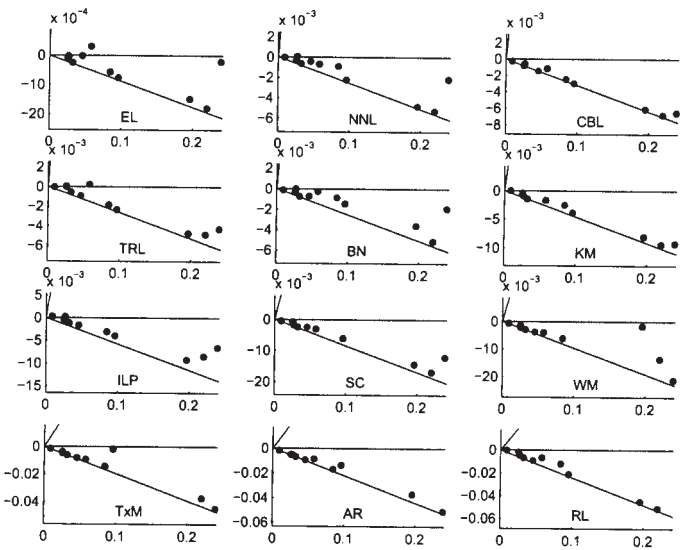
Items	*Research Subjects*
A_1	Ensemble Learning (EL)
A_2	Neural Network Learning (NNL)
A_3	Case-based Learning (CBL)
A_4	Rules-based Learning (Tree Learning) (TRL)
A_5	Bayesian Network (BN)
A_6	Kernel Methods (KM)
A_7	Inductive Logic Programming (ILP)
A_8	Soft Computing (Evolutionary Computing and Fuzzy Logic) (SC)
A_9	Web Mining (WM)
A_{10}	Text Mining (TxM)
A_{11}	Association Rules (AR)
A_{12}	Reinforcement Learning (RL)

abiertas, relacionadas con el área de estudio, utilizando para ello la información que proporciona a través de la red Internet un buscador cualquiera.

Si una página web contiene una palabra clave que sirve de descriptor de una línea de investigación se contabiliza como una página web que trata sobre dicha línea. Entendemos entonces que el número de páginas web que relacionan las distintas líneas de investigación es proporcional al número de artículos que relacionan las líneas de investigación y por tanto esta medida es utilizada como una estimación de la fuerza con que dos disciplinas están relacionadas. Para la obtención de la información utilizamos un agente software que, a partir de las palabras clave, se encarga de consultar el buscador y obtener los listados de las páginas Web, que tratan a la vez sobre alguna de las parejas de las materias investigadas. Es conocido que en la mayoría de los buscadores de páginas Web, tales como Altavista, Google, Yahoo, ... existen bases de datos de históricos, donde se recogen las direcciones de las páginas por años así como los identificadores de cada una de esas páginas. El agente software diseñado se introduce dentro de estas bases de datos, rastrea y cuantifica el número de veces que los identificadores aparecen por año. De esta forma, se ha realizado una búsqueda, en las bases de datos del buscador Altavista, del número de páginas donde aparezcan referenciados a la vez dos ítems concretos en el año 2004; y hemos denotado los números de enlaces (frecuencias absolutas) por n_{ij}, con $i, j = 1, 2, ..., 12$. Después de realizar la búsqueda, hemos encontrado un total de $N = 426.661$ páginas, donde al menos aparece un ítem, obteniendo la tabla II.

Para enlazar estos datos con la construcción de similitudes se ha tenido en cuenta que el número de datos disponibles es suficientemente grande, y hemos optado por una interpretación frecuencialista de la probabilidad, considerando de esta forma que $P(A_i \cap A_j) = \frac{n_{ij}}{N}$ para $i, j = 1, 2, ..., 12$, con lo cual estamos en condiciones de aplicar la función núcleo similitud. Hemos realizado los cálculos (ver tabla III) y representado las similitudes en la figura 6, donde podemos ver la posición (para interpretar esta gráfica, y la siguiente, es necesario saber que los diferentes ítems se encuentran ordenados en función de sus probabilidades) de los valores $k_{A_i}(A_j)$ variando A_i desde $i = 1$ hasta 12, con respecto a todos los restantes ítems.

Figura 6
Representación gráfica de todas las similitudes



Del estudio de estas gráficas una a una tenemos las siguientes conclusiones acerca de la similitud entre una de las temáticas y su relación con las demás.

Gr1.-A_1 tiene similitud positiva con A_7 (ver tabla III) y con el resto tiene disimilitud. Así, con A_6, A_4 y A_5 dicha disimilitud (similitud negativa) es prácticamente nula. Ahora bien, con las que más disimilitud tiene A_1 es con A_{11} y A_{10} con una disimilitud similar a ambas.

Tabla II
Número de citas en las que aparecen recogidas algunas de las líneas de investigación relacionadas con el Aprendizaje en el período mayo 2003-mayo 2004

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
A_1	3.780	78	16	88	41	167	348	76	35	106	64	816
A_2	78	10.900	78	319	137	318	346	542	91	76	137	1.680
A_3	16	78	13.600	120	18	9	306	101	40	66	81	471
A_4	88	319	120	11.300	290	114	755	160	82	190	424	871
A_5	41	137	18	290	10.800	187	512	548	410	587	200	1.760
A_6	167	318	9	114	187	19.400	423	569	224	382	271	725
A_7	348	346	306	755	512	423	24.600	779	675	890	1.790	3.110
A_8	76	542	101	160	548	569	779	36.000	847	952	750	3.490
A_9	35	91	40	82	410	224	675	847	40.800	7.260	3.180	712
A_{10}	106	76	66	190	587	382	890	952	7.260	83.500	2.570	797
A_{11}	64	137	81	424	200	271	1.790	750	3.180	2.570	93.800	703
A_{12}	816	1.680	471	871	1.760	725	3.110	3.490	712	797	703	102.000

Tabla III
Similitudes entre todas las líneas de investigación

	$P(A_i)$	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
A_1	0,008859	0,00878100	-4,3520E-05	-0,000245	-2,8389E-05	-0,00012816	-1,142E-05	0,00030482	-0,00056940	-0,00076517	-0,00148541	-0,00179773	-0,000205
A_2	0,025547	-4,3520E-05	0,02489455	-0,00063151	7,1055E-05	-0,00032557	-0,00041629	-0,00066203	-0,00088525	-0,00222970	-0,00482161	-0,00529537	-0,00216
A_3	0,031875	-0,000245	-0,00063151	0,03085938	-0,00056296	-0,00076467	-0,00142826	-0,00112065	-0,00245280	-0,00295438	-0,00608351	-0,00681786	-0,006516
A_4	0,026484	-2,8389E-05	7,1055E-05	-0,00056296	0,02578329	9,2930E-06	-0,00093705	0,00024252	-0,00185967	-0,00234045	-0,00473789	-0,00482882	-0,004290
A_5	0,025312	-0,00012816	-0,00032557	-0,00076467	9,2930E-06	0,02467210	-0,00071267	-0,00025945	-0,00085141	-0,00145962	-0,00357807	-0,00509619	-0,001926
A_6	0,045469	-1,142E-05	-0,00041629	-0,00142826	-0,00093705	-0,00071267	0,04340189	-0,00163021	-0,00250292	-0,00382306	-0,00800329	-0,00936112	-0,00917
A_7	0,057657	0,00030482	-0,00066203	-0,00112065	0,00024252	-0,00163021	0,05433268	0,05433268	-0,0030391	-0,00393147	-0,00919784	-0,00848034	-0,006494
A_8	0,084376	-0,00056940	-0,00088525	-0,00245280	-0,00185967	-0,00085141	-0,0030391	-0,0030391	0,07725679	-0,00608339	-0,0142816	-0,01679197	-0,011991
A_9	0,095626	-0,00076517	-0,00222970	-0,00295438	-0,00234045	-0,00145962	-0,00382306	-0,00393147	-0,01139603	0,08648189	-0,00169876	-0,01356989	-0,021192
A_{10}	0,195705	-0,00148541	-0,00482161	-0,00608351	-0,00473789	-0,00357807	-0,00800329	-0,00919784	-0,02174047	--0,00169876	0,15740499	-0,03700173	-0,044918
A_{11}	0,21984	-0,00179773	-0,00529537	-0,00681786	-0,00482882	-0,00509619	-0,00848034	-0,00848034	-0,02383721	-0,01356989	-0,03700173	0,17151411	-0,050910
A_{12}	0,239065	-0,00020548	-0,0021699	-0,00651640	-0,00429016	-0,00192638	-0,0091709	-0,00649465	-0,02065617	-0,02119219	-0,04491853	-0,05091011	0,181913

Gr2.- A_2 tiene similitud positiva con A_4 y disimilitud con el resto, teniendo en cuenta que es prácticamente nula con A_1. Con las que más disimilitud tiene es otra vez con A_{11} y A_{10}. Hemos de hacer notar que la escala en esta gráfica es distinta a la anterior, así, en esta gráfica, la disimilitud es mayor que la anterior, sin embargo la similitud es mayor en la primera.

Gr3.- A_3 tiene la misma escala que la gráfica segunda y podemos apreciar que existe disimilitud con todas las líneas, siendo otra vez A_{11} con la que más disimilitud presenta.

Gr4.- A_4 tiene similitud positiva, aunque pequeña, con A_7, A_2 y A_5, aunque con estas dos últimas sea practicamente nula. De igual manera es A_{11} la más disímil.

Gr5.- A_5 tiene similitud positiva con A_4 como hemos dicho anteriormente en Gr4. Con el resto de las líneas tiene disimilitud y sigue siendo A_{11} la más disímil con esta línea.

Gr6.- A_6 es disímil con todas las líneas, siendo con A_1 prácticamente nula.También es aquí A_{11} la más disímil.

Gr7.- A_7 tiene similitud positiva con A_1 y A_4 como ya hemos indicado. En este caso es A_{10} la más disímil, pero seguida muy de cerca de A_{11}.

Gr8.- A_8 tiene similitud negativa con cada línea de investigación y A_{11} es la línea con más disimilitud.

Gr9.- A_9 tiene similitud negativa con cada línea de investigación y A_{12} es la línea con más disimilitud.

Gr10.- A_{10} tiene similitud negativa con cada línea de investigación y A_{12} es la línea con más disimilitud.

Gr11.- A_{11} tiene similitud negativa con cada línea de investigación y A_{12} es la línea con más disimilitud.

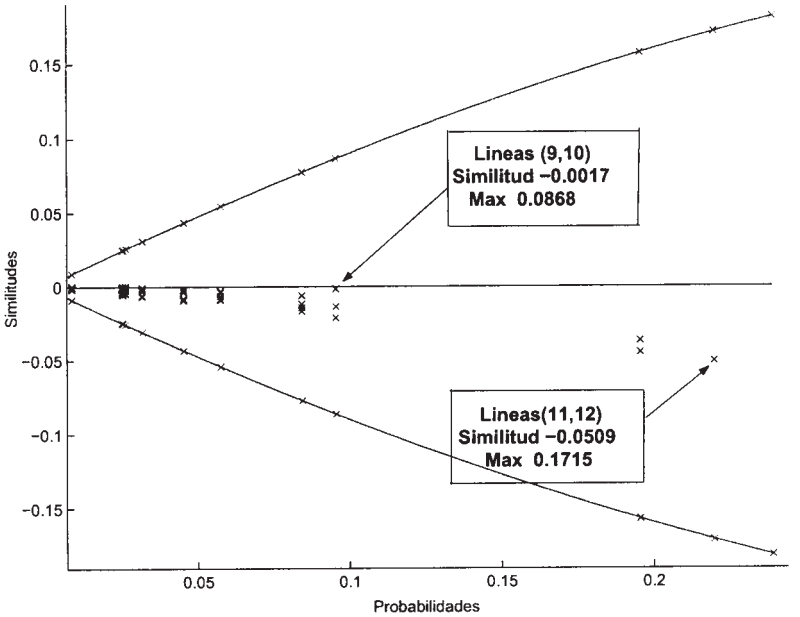
Gr12.- A_{12}, por simetría a A_{11}, tiene similitud negativa con cada línea de investigación y A_{12} es la línea con más disimilitud.

De los comentarios hechos en estos gráficos, hemos de resaltar que la línea A_{11} es la más disímil con casi todas las líneas de investigación, lo que se ampliará y aclarará con la construcción de la tabla IV. Asimismo, se puede señalar que debido a la naturaleza todavía joven de estos estudios, 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.

Hemos de hacer notar que el principal defecto que presenta la figura 6 es que necesita tantas gráficas como ítems se estudian y que las escalas no coinciden, lo que hace difícil la comparación entre gráficas. 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 el siguiente gráfico.

En primer lugar, consideramos los sucesos $\{A_1, A_2, ..., A_n\}$ ordenados a partir de sus probabilidades, es decir, $0 \leq P(A_1) \leq P(A_2) \leq \cdots \leq P(A_n)$, y se representan estas probabilidades sobre el eje de abscisas. Sobre el eje de ordenadas representamos las similitudes de la siguiente forma: Se toma el primer suceso A_1 y se representa el conjunto $\{-k(A_1, A_1), k(A_1, A_2), ..., k(A_1, A_n), k(A_1, A_1)\}$ con abscisa $P(A_1)$, a continuación se toma el suceso A_2 y se representa el conjunto $\{-k(A_2, A_2), k(A_2, A_3), ..., k(A_2, A_n), k(A_2, A_2)\}$

Figura 7
Representación gráfica de todas las similitudes en un único gráfico



con abscisa $P(A_2)$, y así sucesivamente hasta el suceso A_n. La explicación del por qué se actúa de esta forma viene motivada por la desigualdad (2). 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 $|k(A_i, A_j)| \leq k(A_i, A_i)$ si $i < j$, lo cual valida la construcción realizada. Por otro lado, si existe algún suceso con probabilidad superior a 1/2 se trabaja con su complementario y, gracias a las propiedades de la función núcleo similitud y de las conclusiones sobre el complementario, se tiene la inicial.

La representación gráfica, siguiendo este proceso, aplicada a las doce líneas de investigación relacionadas con la Teoría del Aprendizaje, puede verse en la figura 7. En esta gráfica el símbolo × representa la similitud y el gráfico está diseñado de tal forma que cuando se sitúa el ratón sobre el símbolo × aparece un cuadro indicando los ítems de referencia. De la visualización de esta gráfica, se sigue que entre los ítems con mayor probabilidad (ítems A_{10}, A_{11} y A_{12}) no existe mucha similitud, con lo que podemos concluir que estas líneas de investigación siguen caminos distintos con pocas interrelaciones. Por otro lado, nótese que, si se fija el ítem A_7, las similitudes $k(A_7, A_8)$ y $k(A_7, A_9)$ están próximas, pero sin embargo la similitud entre los ítems A_8 y A_9 es negativa y alta (ver la tabla III).

Por otro lado, surge la siguiente pregunta: ¿Cómo se ve en la figura 7 que dos ítems son muy similares o muy disimilares? Para ello, si la similitud entre ellos es positiva, se dibuja 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 muy alta. De igual manera, si la similitud entre ellos es negativa, se dibuja el trián-

gulo $\Delta_2 = \{k(B, B), k(A, B), k(A, A)\}$, y si el área encerrada en el triángulo es pequeña, entonces se tiene que son sucesos de tamaño parecido y disimilitud muy 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.

Por otro lado, para comprobar qué líneas son las más disímiles (símiles) con las demás se ha construido la tabla IV como sigue: en cada columna, en orden ascendente la similitud, así por ejemplo en la columna A_1; la más disímil con ella es A_{11}, a continuación A_{10} y así sucesivamente hasta A_7, que es con la que tiene más similitud. En la figura 6, el primer gráfico nos indica que ello es así. No obstante, mirando la tabla II con valores absolutos no concuerda la relación de orden con los obtenidos, ya que, por ejemplo, la temática A_{12} es la primera en la tabla II en relación al número n_{ij}; sin embargo, en la tabla IV, A_{12} ocupa el sexto lugar y A_4 el quinto lugar en la tabla II y el tercer lugar en la tabla IV. Hay que destacar que A_7 ocupa el segundo lugar en la tabla II y el primer lugar en la tabla IV en cuanto a similitud. Todo esto es debido a que en la tabla IV hemos tenido en cuenta valores relativos y no absolutos como en la tabla II. La tabla IV, por tanto, nos da una relación de orden más acorde con el sentido de afinidad entre temáticas.

Ahora bien, en la tabla IV hay algunas apreciaciones que hemos recogido en la tabla V. Así, la tabla V la hemos construido a partir de la tabla IV de la siguiente manera: Al fijarnos en la tabla IV, en la primera fila aparece repetida la temática A_{11} ocho veces, A_{12} tres veces y A_{10} una vez. Esos números son los que aparecen en la segunda columna de la tabla V por ese orden, indicando que están en el primer lugar de disimilitud y por ello la primera columna queda con el orden correspondiente de las temáticas. Análogamente en la segunda fila de la tabla IV, A_{11} aparece tres veces, A_{12} dos y A_{10} siete veces, que colocamos en la tercera columna de la tabla V. Es decir, la tabla V recoge el orden de disimilitud de cada una de las temáticas en relación con las demás, sin tener en cuenta el valor de las similitudes, sino el orden. Para poder comparar unas con otras hacemos una

Tabla IV
Orden de cada línea de investigación con respecto a las demás atendiendo a la similitud

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
1	A_{11}	A_{11}	A_{11}	A_{11}	A_{11}	A_{11}	A_{10}	A_{11}	A_{12}	A_{112}	A_{12}	A_{11}
2	A_{10}	A_{10}	A_{12}	A_{10}	A_{10}	A_{12}	A_{11}	A_{10}	A_{11}	A_{11}	A_{10}	A_{10}
3	A_9	A_9	A_{10}	A_{12}	A_{12}	A_{10}	A_{12}	A_{12}	A_8	A_8	A_8	A_9
4	A_8	A_{12}	A_9	A_9	A_9	A_9	A_9	A_9	A_7	A_7	A_9	A_8
5	A_3	A_8	A_8	A_8	A_8	A_8	A_8	A_7	A_6	A_6	A_6	A_6
6	A_{12}	A_7	A_6	A_6	A_3	A_7	A_6	A_6	A_3	A_3	A_7	A_3
7	A_5	A_3	A_7	A_3	A_6	A_3	A_3	A_3	A_4	A_2	A_3	A_7
8	A_2	A_6	A_5	A_1	A_2	A_4	A_2	A_4	A_2	A_4	A_2	A_4
9	A_4	A_5	A_2	A_5	A_7	A_5	A_5	A_2	A_{10}	A_5	A_5	A_2
10	A_6	A_1	A_4	A_2	A_1	A_2	A_4	A_5	A_5	A_9	A_4	A_5
11	A_7	A_4	A_1	A_7	A_4	A_1	A_1	A_1	A_1	A_1	A_1	A_1
12	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}

suma ponderada, tomando como peso los ordenes que están en la primera fila y los resultados los ponemos en la penúltima columna. Así, para A_{11}, tenemos $(8 \times 1) + (3 \times 2) = 14$, que hemos puesto en la penúltima columna de la tabla V. Al dividir por 11, obtenemos el orden ponderado en el que están las temáticas en relación a su disimilitud con las demás. Así, por ejemplo A7 tiene una suma ponderada valorada en 76. Reordenado esta tabla según la última columna, obtenemos la tabla V, en la que tenemos las temáticas ordenadas de menor a mayor por similitud. De esta tabla podemos ver que la temática con más relación con las demás es A_1, le sigue A_4 y finalizan A_{12} y A_{11} respectivamente, aunque podemos ver que A_{10} y A_{12} son muy parecidas en la suma ponderada. También podemos apreciar que existe un empate entre A_8 y A_9. Este desempate lo hemos deshecho reordenado otra vez las dos por la primera columna en la que existe alguna diferencia, que en este caso es la quinta. Esta tabla nos indica que las temáticas más fuertes A_{12}, A_{11}, A_{10}, en el sentido de que son las que más publican, están las primeras líneas de la tabla V, con lo cual son las que tienen más disimilitud con las demás. Este resultado nos indica que éstas se apoyan poco o comparten poco con las otras temáticas, es decir, parecen ser autosuficientes por sí mismas.

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 técnica aplicada, nos ha permitido hacer una representación gráfica de las líneas de investigación de una determinada área temática y de las relaciones cualitativas que existen entre ellas.

También podemos hacer una ordenación por disimilitudes (similitudes) de las distintas líneas de investigación, así como indicar cuáles son las más símiles con las demás.

En un futuro inmediato vamos a desarrollar varias líneas de trabajo. Una de ellas es utilizar, como bases documentales, las bases de datos científicas y realizar el estudio descrito en este artículo para distintas áreas temáticas de interés. Otra línea de trabajo va

Tabla V
Potencia de cada línea de investigación respecto a las demás

	1	2	3	4	5	6	7	8	9	10	11	*Suma*	*Rango*
A_{11}	8	3	–	–	–	–	–	–	–	–	–	14	1,27
A_{12}	3	2	4	1	–	1	–	–	–	–	–	29	2,63
A_{10}	1	7	2	–	–	–	–	–	1	–	–	30	2,72
A_9	–	–	3	7	–	–	–	–	–	1	–	47	4,27
A_8	-	–	3	2	6	–	–	–	–	–	–	47	4,27
A_6	–	–	–	–	4	4	1	1	–	1	–	69	6,27
A_3	–	–	–	–	1	4	6	–	–	–	–	71	6,45
A_7	–	–	–	2	1	3	2	–	1	–	2	76	6,91
A_2	–	–	–	–	–	–	1	5	3	2	–	94	8,55
A_5	–	–	–	–	–	–	1	1	6	3	0	99	9
A_4	–	–	–	–	–	–	1	4	1	3	2	100	9,09
A_1	–	–	–	–	–	–	–	1	–	2	8	116	10,5

a ser realizar un estudio dinámico acerca del crecimiento relativo de cada subtemática y de las similitudes temporales que existen entre las subtemáticas y entre las temáticas. Pretendemos también encontrar las similitudes entre investigadores o grupos de investigación relevantes en determinadas áreas, subáreas o líneas de investigación.

La medida utilizada, número de páginas que trata sobre una pareja de ítems obtenida en el buscador, es una medida débil de la similitud, aunque tiene la ventaja de que es difícil encontrar un buscador que no la acepte. Aun así, se puede mejorar la medida discriminando con pesos las distintas categorías de páginas Web recuperadas en las consultas, así por ejemplo, podemos dar más importancia a los enlaces que pertenecen a dominios de tipo «.edu» que a los «.com».

Hemos querido destacar alguna de las aplicaciones que se pueden desarrollar con la técnica matemática descrita en la sección 2 y desarrollada en (23, 24). Esto nos indica la potencialidad de este índice de similitud que está siendo ampliada a intervalos reales y a series temporales, con lo que podremos estudiar las similitudes de ejemplos tales como series económicas, niveles de audiencia radiofónicos, televisivos, etc.

Agradecimientos

Este trabajo ha sido soportado en parte por la ayuda ACPAI-2003/14 concedida por la Junta de Andalucía y por el Ministerio de Ciencias y Tecnología bajo el proyecto DPI2003-07146-C02-01. Fondos Feder.

Referencias

1. CALLON, M.; LAW J., y RIP, A. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*, Macmillan.
2. COULTER, N.; MONARCH, I.; KONDA, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *American Society for Information Science,* 49 (13):1206-1223.
3. EGGHE, L.; ROUSSEAU, R. (1990). *Introduction to Informetrics.Quantitative methods in library, documentation and information science.* Elsevier.
4. WOLFRAM, D. (2000). *Applications of Informetrics to information retrieval research*. Informing Science 3 (2): 77-82.
5. NOYONS, E. C. M.; BUTER, R. K.; VAN RAAN, A. (2002). Bibliometric mapping as a science policy tool. *Information Visualisation,* Proceedings Sixth International Conference on: 679-684.
6. BUTER, R. K.; NOYONS, E. C. M. (2002). Using bibliometric maps to visualise term distribution in scientific papers. *Information Visualisation,* Proceedings Sixth International Conference on: 697-702.
7. ALMIND, T. C.; INGWERSEN, P. (1997). Informetric analyses on the world wide web: methodological approaches to webometrics. *J. of documentation,* 53 (4): 404-426.
8. ROUSSEAU, R. (1997). *Sitations: an exploratory study.* PhD thesis, Faculty of Industrial Sciences and Technology. Zeedijk.
9. LARSON, R. (1996). Bibliometrics of the world wide web: an exploratory analysis of the intellectual structure of cyberspace. Disponible en: http://sherlock.berkeley.edu/asis96/asis96.html.

10. KOHONEN, T. (1998). Self-organization of very large document collections: State of the art. Proceeding of ICANN98.
11. KOHONEN, T. (1998). Self-organization of a massive document collection. IEEE.
12. LIN, X.; MARCHIONINI, G. (1991). A self-organizaing semantic map for information retrieval. En Proc of 14 ACM/SIGIR Conf. Research and development in information retrieval.
13. KINNUCAN, M.; NELSON, M.; ALLEN, B. (1987). Statistical methods in information science research. *Annual Review of Information Science and Tecnology,* (22): 147-178.
14. DEUS, J. E. (2001). *Escalamiento Multidimensional.* Cuadernos de Estadísticas. La Muralla.
15. KLOCK, H.; BUHMAN, J. M. (1997). *Data visualization by Multidimensional Scaling: A Deterministic Annealing Approach*.
16. RUIZ-BAÑOS, R.; CONTRERAS, F. (1998). Cómo consultar eficazmente una base de datos bibliográfica. El método de las palabras asociadas. http://www.ugr.es/fccortes/curriculum/toledo.html.
17. LUC GRIVEL, C. F. (1995). *Une station de travail pour classer, cartographie et analyser l'information bibliographique dans une perspective de veille scientifique et technique.* Solarion.
18. CALLON, M.; COURTIAL, J. P.; LAVILLE F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics,* 22(1):155{205, 1991.
19. COURTIAL, J. P. (1994). A coword analysis of scientometrics. *Scientometrics,* 3 (31): 251-260.
20. BRAAM, R.; MOED, H. F.; VAN RAAN, A. (1991). Mapping of science by combined cocitation and word analysis.ii: dynamical aspect. *J. American Society for Information Science,* 42 (4): 252-266.
21. VAPNIK, V. (1998). *Statistical Learning Theory.* John Wiley & Sons, Inc.
22. SCHÖOLKOPF, S.; SMOLA, A. J. (2002). *Learning with Kernels.* The MIT Press, Cambridge, MA.
23. GONZÁLEZ, L. (2002). Análisis discriminante utilizando máquinas núcleos de vectores soporte. Función núcleo similitud. Tesis de doctorado. Dpto. Economía Aplicada I. Universidad de Sevilla, junio.
24. GONZÁLEZ, L.; VELASCO, F.; GASCA, R. M. (2005). A study of the similarities between topics. *Computational Statistics,* 20 (3). En prensa.
25. SKHÖLKOPF, B. (2000). Statistical learning and kernel methods. Technical Report MSR-TR-2000-23, Microsoft Research Limited, febrero.