

Metodología para el Análisis Visual de la Evolución de Conceptos en Bases de Datos Textuales

F. de la Rosa T., R.M. Gasca y J.A. Ortega

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
{ffrosat, gasca, jortega}@lsi.us.es

Abstract. El análisis de los datos almacenados en las bases de datos textuales se hace cada vez más complejo debido al aumento del volumen y la entropía. Estos problemas obligan a buscar nuevos métodos para analizar el contenido de las bases de datos textuales. En este artículo presentamos una metodología que permite el seguimiento de núcleos de información, representados en forma de conceptos. Estos conceptos son obtenidos a partir del filtrado de los datos mediante listas de palabras claves y posteriormente son mostrados en forma de mapas mediante técnicas de Escalado Multidimensional (MDS). La metodología también permite observar la evolución de los conceptos en el tiempo mediante la producción de secuencias temporales de los mapas.

1. Introducción

Debido al crecimiento del volumen y de la entropía de los datos almacenados en las bases de datos textuales, de la cual Internet es un exponente, se hace cada vez más difícil analizar estos datos. Para tratar este problema nos planteamos en este artículo desarrollar una metodología que nos permita extraer información sobre los conceptos que estamos interesados en analizar en cualquier base de datos textual, exigiendo como requisito que permita realizar un *análisis visual* que refleje las relaciones entre los conceptos definidos. Por tanto el punto de partida de la metodología es la descripción de los conceptos y de las medidas que se extraerán de la base de datos. Aunque en este artículo aplicaremos la metodología para resolver la problemática que presenta la obtención de información cuantitativa a través de Internet [Rodríguez97], [Larson96], los resultados son fácilmente adaptable a cualquier otra base de datos, no necesariamente textual y aplicable a *cualquier área temática* que estemos interesados en investigar.

El tipo de análisis textual que proponemos en este artículo también ha sido desarrollado en otras disciplinas. Desde el punto de vista de la Cienciometría o la Informetría existen dos formas de analizar la información textual, mediante el análisis de co-citas [Larson96] que solo es aplicable a textos que tengan la característica de que se citen unos a otros y el análisis de co-palabras [López96] que realiza el análisis relacionando los textos a partir de las palabras claves que tienen en común los textos.

La metodología que proponemos utiliza el análisis de co-palabras para obtener una matriz que representan las distancias estimadas entre los distintos conceptos. Esta matriz se transformada posteriormente mediante técnicas MDS (Multidimensional Scaling [HK&JMB97] y [JRD01]) en mapas que revelan la estructura que relaciona los conceptos. En la figura 1 presentamos el esquema que describe el funcionamiento de la metodología y como esta puede ser aplicada a otras base de datos.

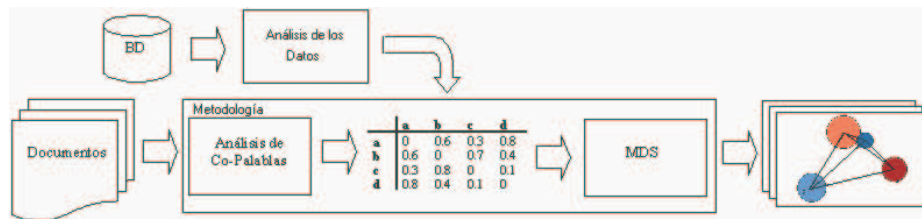


Fig. 1. Esquema de la Metodología

Las técnicas MDS empezaron a desarrollarse a finales de los sesenta en el área de la psicofísica para analizar las percepciones físicas entre distintos individuos, desde un punto de vista estadístico se interpretan como análisis multivariantes. Por ejemplo para el caso analizado en este artículo las variables estadísticas representan un hiperplano en el cual se encuentran distribuidos los conceptos y son las técnicas MDS las encargadas de buscar una proyección de las variables en un espacio de dos o tres dimensiones con la menor pérdida de información posible.

Se han desarrollado técnicas alternativas a las técnicas MDS para la producción de mapas pero para el tratamiento específico de la documentación. Básicamente se han desarrollado dos vertientes, el método de las “Palabras Asociadas” [RRB&FCC98][Luc&Claire95] que construye un grafo utilizando las palabras claves coincidentes entre documentos y a partir del cual se obtienen lo que se ha llamado Diagramas Estratégicos, que es una representación de las características estructurales de los documentos. La otra vertiente más reciente coloca de forma secuencial los documentos para trabajar con ellos en forma de vectores que pertenecen a un espacio vectorial [Lelu92], de forma que una vez clasificados los documentos, la producción de los mapas se realiza proyectando los documentos prototipos de cada clase sobre un plano. El cálculo del plano más adecuado para realizar la proyección se realiza analizando las componentes principales [Luc&Claire95] del sistema.

Los mapas temáticos también han ido evolucionando a medida que la tecnología lo ha permitido, como muestra de ello se han generado mapas que muestran secuencias temporales en forma de películas, en los que se puede observar la evolución de los conceptos. También se han generado mapas dinámicos que asocian los distintos elementos de los mapas (conceptos, enlaces, etc) con operaciones, como puede ser acceder al conjunto de artículos indexados por las palabras claves o incluso profundizar en otros niveles del mapa. En estos ejemplos se puede observar como los mapas son utilizados como navegadores de distintas base de datos textuales.

Modelo Mapa	Características	Tecnología
Mapa 2D	Representan las relaciones entre varios conceptos	Gif
Mapa 2D temporales	Representan la evolución temporal de los conceptos.	Gifs Animadas, Avi, etc.
Mapa 2D dinámicos	Los mapas son interactivos, permitiendo el acceso a otros recursos (registros, mapas de segundo nivel, etc)	HTML Dinámico

Table 1. Evolución de las técnicas de mapeado

Como ejemplo de esta evolución en la página web del grupo “Centre for Science and Technology Studies (CWTS)”, en la Universidad de Leiden, de los Países Bajos podemos observar la evolución que han tenido estos mapas a través de los estudios Bibliométricos que el grupo [CWTS] ha ido publicando.

La principal novedad que planteamos en este artículo es la construcción de una *metodología* que defina las etapas necesarias para aglutinar las diferentes técnicas empleadas durante la construcción de mapas temáticos temporales y la definición de un *modelo* que abstraer las distintas operaciones permitidas en las base de datos textual y sobre el que se construyen las medidas que utiliza la metodología.

El artículo está estructurado en las siguientes secciones, en la primera sección se definen el modelo que servirá de base en apartados posteriores. En las siguientes secciones describimos la metodología y algunas características observadas al implementar la metodología. Finalizaremos el artículo presentando un ejemplo que ilustre el tipo de información que se infiere de los análisis y las conclusiones.

2. Definiciones y Notación

En este apartado definiremos un modelo de base de datos textual que nos proporcionará la terminología necesaria para construir la metodología que presentamos en el artículo. Antes de seguir debemos aclarar que uno de los objetivos marcados en el artículo es aplicar la metodología en el dominio de Internet, por tanto nos referiremos indistintamente a lo largo del artículo a los documentos o páginas Web como componentes básicos o registros de una base de datos textual. Una vez realizadas estas aclaraciones comenzamos con la definición del modelo:

Conceptos u Objetos: Son los distintos aspectos del área de investigación que queremos estudiar. Suele corresponder con temas, subtemas, dominios, técnicas, etc. Es similar al término Centro de Interés definido en [RRB&FCC98].

Buscador: Sistema que mantiene indexado los documentos o páginas Web mediante palabras claves, esto permite la posterior recuperación de los documentos utilizando palabras claves en las consultas. En el artículo utilizaremos la función, $W(i)$, como

una representación abstracta del sistema, que devuelve un conjunto con los documentos que almacena en el instante de tiempo.

Conjunto de Ítems o Palabras Clave: Son aquellas palabras que tienen la mayor probabilidad de indexar los documentos que mejor representen a los conceptos. Representaremos el conjunto de ítems como:

$$CI(c) = \{pc_1, pc_2, \dots\} \quad (1)$$

donde c representa un concepto y CI un conjunto finito de palabras claves. Los documentos recuperados para cada concepto serán utilizados para calcular los índices que representarán el concepto.

Los ejemplos que presentamos en este artículo se han aplicado al área de *Machine Learning* y en la tabla 2 se asociamos a cada concepto su conjunto de ítems.

Abrev.	Concepto	Items (pc)
AL	Automated Learning	Automated Learning
AD	Automated Discovery	Automated Discovery
SVM	Support Vector Machines	Support Vector Machines, Decision Trees
DM	Data mining	Data Mining
HS	Hybrid Systems	Hybrid Systems, Neural and Symbolic Processing
LIR	Learning for Information Retrieval	Learning for Information Retrieval
ML	Machine Learning	Machine Learning
NN	Neural Network	Neural Network
RL	Reinforcement Learnig	Reinforcement Learnig

Table 2. Palabras claves asociada a cada concepto.

Todas estas definiciones se relacionan entre sí mediante las operaciones de **consultas** representadas por la función, $Q(c,t)$, que obtiene como resultado el conjunto de documentos indexados por las palabras claves que representan al concepto c en el instante t . Es posible construir operaciones más complejas transformando los operadores lógicos aplicados a los conceptos en operadores de conjunto aplicados sobre las consultas simples.

$$Q(c_1 \wedge c_2, t_1) = Q(c_1, t_1) \cap Q(c_2, t_1) = \{D\}^1 \quad (2)$$

$$Q(c_1 \vee c_2, t_2) = Q(c_1, t_2) \cup Q(c_2, t_2) = \{C, G, B, E, P, D, A\}$$

Nuestro modelo también define las **consultas periódicas**, $QP(c,p)$, como aquellas consultas que recoge información que ha existido en un período de tiempo, esto es desde el instante t_1 al instante t_2 . En la ecuación 4 mostramos como estas consultas se construyen a partir de consultas mas simples. A lo largo del artículo sobreentenderemos que las consultas temporales abarcan el periodo de un año.

¹ $\{D\}$ y $\{C, G, B, E, P, D, A\}$ representan conjuntos de documentos o páginas Web.

$$QP(c_3, p) = QP(c_3, [t_1, t_2]) = Q(c_3, t_1) \cap Q(c_3, t_2) = \{H, E, D\}^2 \quad (3)$$

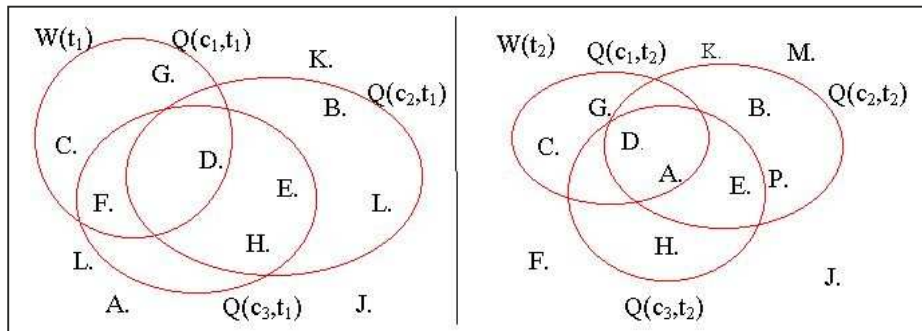


Fig. 2. En la figura presentamos los estados de un sistema de base de datos textual en dos instantes diferentes t_1 y t_2 . Si realizásemos una analogía entre el modelo e Internet interpretaríamos $W(t)$ como todas las páginas Web que existen en un buscador en el instante t y $Q(c,t)$ sería las páginas Web indexadas por el sistema las palabras claves del concepto c en una consulta.

Indicador o Índice de Impacto de un Concepto: Es la estimación de la importancia de un concepto dentro del área en estudio, de forma general mide alguna característica común al conjunto de conceptos, como puede ser la productividad, la madurez, etc. Su cálculo no puede ser realizado de forma directa sino que es estimado a partir de información medible en el conjunto de documentos o páginas. En la metodología que presentamos será representada como $I(c)$ y su estimación será realizada por el número de documentos que indexan las palabras claves asociada al concepto.

$$I(c_2, t_1) = |Q(c_2, t_1)| = 5; I(c_3, p) = I(c_3, [t_1, t_2]) = |QP(c_3, [t_1, t_2])| = 3 \quad (4)$$

Indicador o Índice de Afinidad entre dos Conceptos: Es un índice que estima la distancia que existe entre dos conceptos. Es utilizado en la metodología para el cálculo de los mapas temáticos. El índice debe cumplir una serie de propiedades dependientes del tipo de técnica MDS utilizada para calcular los puntos de los mapas, para las técnicas *métricas* deberá de cumplir las propiedades de las distancias mientras que para las técnicas *no métricas* este requisito se relaja. Nos referiremos al índice con la notación $I(c_1, c_2, t)$ y su estimación será realizada por la inversa del número de documentos que indexan las palabras claves de dos conceptos.

² La ecuación cumple $t_1 > t_2$.

$$I(c_1, c_3, t_1) = \frac{1}{|Q(c_1 \wedge c_3, t_1)|} = \frac{1}{3} = 0'333 \quad (5)$$

$$I(c_1, c_2, [t_1, t_2]) = \frac{1}{|QP(c_1 \wedge c_2, [t_1, t_2])|} = \frac{1}{1} = 1$$

Mapa Bibliométrico o Mapa Temático: Es una representación gráfica de la estructura que relaciona los distintos conceptos y se interpreta de forma semejante a un mapa topográfico de dos o tres dimensiones. En estos mapas existe una relación entre los índices de Impacto y Afinidad y los distintos elementos que componen el mapa, de forma que la afinidad entre conceptos viene representada por los disposición de los puntos en el mapa y el impacto se representa de forma implícita en los círculos o bolas que identifica cada concepto en el mapa, este efecto se consigue ajustando de forma proporcional el área o el diámetro de los círculos al índice $I(c)$. En las secuencias de mapas temporales también es posible utilizar el color de los círculos para representar la propiedad de monotonía creciente (tonos rojos) o decreciente (tonos azules) del impacto como se propone en [CTWS].

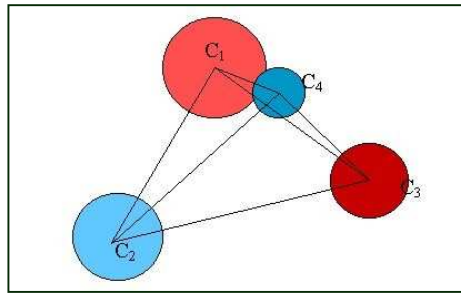


Fig. 3. En la figura exponemos un ejemplo de cómo interpretar los distintos elementos de los mapas. Si con las etiquetas C_i , R_i y D_{ij} identificamos el concepto i -ésimo, el radio del concepto i -ésimo y la distancia entre el concepto i -ésimo y el j -ésimo. Los elementos de la figura cumplirían que $R_4 < R_3 < R_2 < R_1 \Rightarrow I_4 < I_3 < I_2 < I_1$ que se interpreta como que el concepto 1 tiene más impacto que el resto de conceptos. Y $D_{13} < D_{12} \Rightarrow I_{13} < I_{12}$ que indica que los conceptos 1 y 3 son más afines que los conceptos 1 y 2.

Producción Cartográfica: Al proceso de creación de estos mapas temáticos.

Patrón: Regla que indica como obtener medidas del conjunto de documentos almacenados en la base de datos, con el objeto de utilizarlas para la estimaciones de los índices.

3. Metodología para la Construcción de Mapas Temáticos Temporales

La metodología que presentamos en este artículo es una abstracción de los procesos que son necesarios seguir para obtener una serie de mapas temporales. Para la definición de la metodología se han utilizado tres elementos las *etapas* o *módulos* que encapsulan conjuntos de operaciones que se ejecutan ordenadamente para obtener como resultado un tipo de producto, los *almacenes* donde se almacenan los productos de entrada y de salida a las etapas y los *interfaces* que son una clase especial de módulos que tiene la peculiaridad de poder interactuar con el usuario.

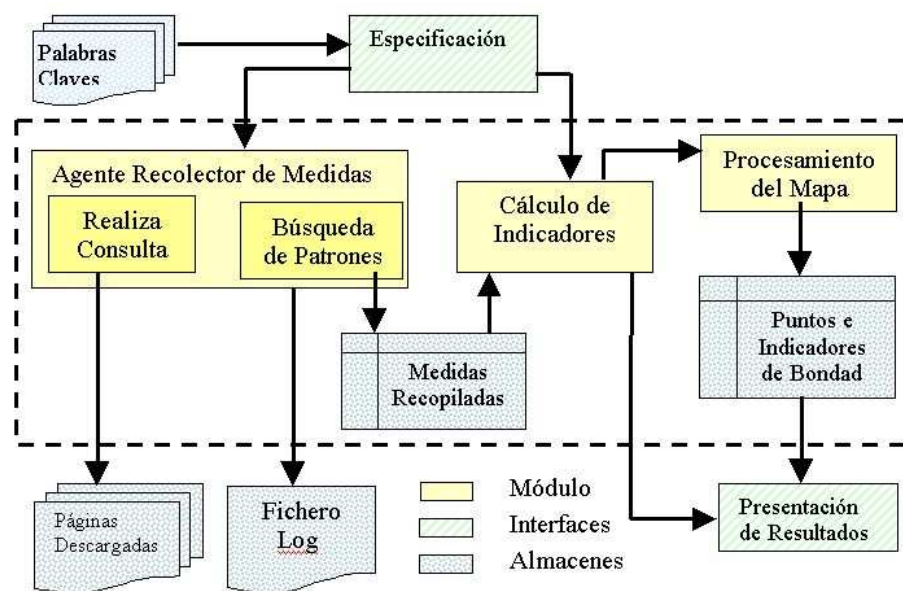


Fig. 4. Esquema de una metodología para la construcción de Mapas Temáticos

Como podemos observar en el esquema de la metodología se ha dividido en las etapas de: Especificación, Obtención de Medidas, Cálculo de Indicadores, Procesamiento de Mapas y Presentación de Resultados.

La etapa de *Especificación* es la encargada de recopilar los parámetros que permiten al usuario programar la metodología para resolver el problema, los parámetros modificables son:

- Los conceptos o centros de interés que tendrá el mapa y sus atributos como el conjunto de palabras claves, las leyendas, etc.

- Indicar el buscador que queremos utilizar, algunas capacidades como realizar series temporales o el tipo de calculo utilizado para estimar los índices dependerán de las posibilidades del buscador.
- Indicar los periodos si el mapa representa una serie temporal.

El proceso de *Obtención de Medidas* debe ser capaz de imitar las operaciones que realizaría una persona recuperar las paginas, contar palabras (obtención de medidas), etc. Dada la naturaleza de estas tareas proponemos que sean implementadas en un agente software, que se representa en el siguiente pseudocódigo:

```

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

```

El *módulo de cálculo de índices* utiliza las medidas recogidas por el agente software para calcular los índices de Impacto y Afinidad. El método utilizado para estimar los índices es independiente de la metodología y las posibles estimaciones depende en gran medida de las capacidades del buscador sobre el que se realizan las consultas. El método que propuesto en este artículo para estimar los índices es muy simple pero tiene la ventaja de que se puede aplicar a la gran mayoría de buscadores. Se podrían mejorar la precisión de estos índices discriminando con pesos a los elementos que tiene más relevancia, por ejemplo en el caso de que el índice contase el numero de *enlaces* que hacen referencia a una palabra clave, se podría dar más importancia a los *enlaces* que pertenecen a dominios de tipo “.edu” que a los “.com”.

De los índices calculados el de afinidad es utilizado en la etapa de *Procesamiento de Mapas* para obtener la disposición de los conceptos en el plano y la bondad de la solución calculada. La metodología no propone ningún modelo específico para realizar estos cálculos, sino que define una interfaz que permite utilizar cualquier técnica que pueda realizar estos cálculos a partir de la *matriz de distancia*. Esta matriz es simétrica y se calcula a partir del índice de afinidad. Definiremos la matriz como:

$$M(i, j, p) = \{I(c_i, c_j, p) \quad \forall i < j\} \quad (6)$$

La fila i y la columna j de la matriz hace referencia a los conceptos c_i y c_j .

Finalmente la etapa de *Presentación de Resultados* muestra al usuario toda la información recopilada en la metodología a través de los mapas temáticos y su evolución en el tiempo, esta etapa debe calcular también la bondad de la solución.

5. Implementación

En este apartado comentaremos como han sido implementadas algunas de las tareas de la metodología. Para implementar la etapa de *Obtención de Medidas* se ha implementado el agente utilizando el lenguaje *Perl* debido a las facilidades que ofrece tanto para la lectura y escritura en ficheros como para analizar léxicamente los documentos, funciones que han sido utilizadas para recuperar las cadenas que contenían información sobre las medidas. También se observó que era muy importante que el agente generase un *log* de las operaciones que iba realizando ya que en ocasiones desconocemos el comportamiento que se ha programado en el buscador y es fácil que se descubra la existencia de algún patrón que no se tuvo en cuenta.

La etapa de *Procesamiento del Mapa* se ha implementado utilizando el modelo MDS desarrollado en la herramienta *Kyst* [KYST] que se encuentra disponible en la red y que permite su ejecución en modo batch. El modelo implementado por *Kyst* se basa en los trabajos de [Kruskal64a y 64b] y [Young72] y básicamente calcula la posición de cada concepto minimizando una función objetivo también llamada función de Stress. El método utilizado para minimizar la función es el *método del gradiente* y la función de Stress que utiliza es la que propuso Kruskal:

$$Stress = \sqrt{\frac{\sum_{i < j} (M(i, j, p) - d(i, j, p))^2}{\sum_{i < j} M(i, j, p)^2}} \quad (7)$$

Donde $d(i, j, p)$ es la distancia entre los conceptos i y j en el mapa temático y p es el periodo. Al minimizar esta función en sucesivas iteraciones se consigue que las distancias entre las posiciones de los conceptos $d(i, j, p)$ tiendan a mantener las distancias originales del modelo real. Esta función también ha sido utilizada como indicador de bondad ya que su valor se puede interpretar cualitativamente, según [AJ&WW98] como:

Stress	Bondad
>20%	No aceptable
[20-10)%	Pobre
[10-5)%	Aceptable
[5-2,5)%	Buena
[2,5-0)%	Excelente
[0)%	Perfecta

Table 3. Tabla con la interpretación cualitativa de los resultados de la función de cruskal.

Por último destacar que el módulo de *Presentación de Resultados* se ha utilizado los mapas de burbujas que facilita el asistente de Excel para representar los mapas.

4. Características Propias Observables

En este apartado describimos dos características particulares observadas durante nuestras investigaciones en las base de datos que hemos utilizados y que lo diferencia respecto otras base de datos textuales.

Cobertura de los datos: Como es sabido los buscadores trabajan solo con una porción de las páginas web que existen en Internet, si asociamos la función $W_I(t)$ a una representación abstracta de Internet estas característica se puede expresar como:

$$|W_I(t)| > \left| \bigcup_{\substack{c \in \text{Conceptos} \\ B \in \text{Buscador}}} Q_B(c, t) \right| \quad (8)$$

A pesar de esta característica se considera que la cobertura que ofrecen los buscadores es suficientemente amplia como para representar el contenido de Internet.

Inconsistencia Temporal de las Consultas: El hecho de realizar consultas a través de Internet imposibilita realizarlas de forma instantánea tardando un periodo considerable en recoger los datos. Esta tarea puede durar días en terminarse y es la causa de la aparición de inconsistencias en los datos recopilados. Por ejemplo puede suceder que $|QP_{t_1}(a \wedge b, p)| \neq |QP_{t_2}(b \wedge a, p)|$ donde t_1 y t_2 indican que las consultas se han realizado instantes de tiempos diferentes, a primera vista parece que se ha producido un error en la recopilación de los datos ya que las dos consultas deberían haber devuelto el mismo conjunto de documentos, pero no tiene porqué, ya que estamos trabajando sobre un sistema dinámico, donde aparecen y desaparecen páginas a lo largo del tiempo. Para tratar este problema se debe tener en cuenta que el agente ha ido recopilando datos consistentes para varios instantes de tiempos incluidos t_1 y t_2 , por eso consideramos el valor medio como la mejor estimación.

$$|QP(a \wedge b, p)| = |QP(b \wedge a, p)| = \frac{|QP_{t_1}(a \wedge b, p)| + |QP_{t_2}(b \wedge a, p)|}{2} \quad (9)$$

Otra solución es realizar las dos consultas de forma consecutiva pero el problema de la inconsistencia temporal de las consultas persistiría.

6. Ejemplo de Aplicación

Para ilustrar los resultados que se obtienen al aplicar la metodología nos hemos centrado en una temática concreta como puede ser *Machine Learning* y hemos utilizado sus áreas como conceptos para observar como han ido evolucionando a la largo del tiempo. A cada área o concepto le hemos asignamos un conjunto de palabras claves que se pueden consultar en la tabla 2 y los resultado obtenidos se muestran en las figuras 5 y 6.

Para interpretar los resultados de la figura 5 hay que tener en cuenta que la herramienta Kyst no sincroniza las posiciones de los conceptos entre mapas y que el tamaño del impacto representados por Excel en los mapas de burbuja sólo es proporcional entre los conceptos de un mismo mapa. Por eso para observar la evolución temporal del impacto entre los distintos mapas utilizamos la figura 6.

La herramienta Kyst utiliza la función de Kruskal para calcular la bondad de las soluciones y en la mayoría de los mapas que mostramos este indicador varía entre 0.1 y 0.2 que están clasificados como resultados pobres pero dentro de lo aceptable, este hecho resta credibilidad a las posibles interpretaciones. Para mejorar estos resultados [Spence72] y [Sherman72] proponen utilizar conjuntamente los modelos MDS y el algoritmo de Monte Carlo.

Aunque la interpretación de los mapas no es el objeto del artículo, seguidamente exponemos algunas observaciones sobre la evolución de los conceptos que podemos obtener al analizar los mapas y que ilustran los resultados que se pueden obtener con esta metodología:

- En la figura 6 podemos observar como la evolución del Impacto es creciente a lo largo del tiempo para todos los conceptos, pero hay que destacar el grupo formado por las disciplinas etiquetadas como NN, DM y ML han tenido una evolución exponencial en cuanto a la producción de páginas Web en comparación con el resto de disciplinas cuya evolución ha sido más lineal.
- En la figura 5 podemos observar como las disciplinas etiquetadas como AL, AD y LIR forman parte de un cluster muy compacto que se conserva a lo largo del tiempo, aunque la producción de estas tres disciplinas no es muy elevada en comparación con de otras disciplinas, este hecho podría interpretarse como que estas tres disciplinas realmente se pudiesen agrupar en una disciplina mayor.
- En la figura 5 podemos observar como la disciplina DM ha evolucionado a partir de la NN y como ha ido adquiriendo mayor entidad a lo largo del tiempo, llegando a tener tanta importancia como su precursora. Esto último punto se confirma con los datos de la figura 6 donde se observa como la disciplina DM va adquiriendo más entidad respecto de la disciplina NN, hasta el punto de que esta última cede el primer puesto en el ranking a DM en el periodo 1999-2000.
- En la figura 5 también se observa que la disciplina SVM se ha mantenido aislada a lo largo del tiempo respecto al resto de disciplinas.
- En la figura 6 podemos observar una explosión en el número de páginas que se publicaron entre 1999 y 2000 en todas las disciplinas. Probablemente este fenómeno sea debido al aumento de la facilidad de acceso a Internet.

7. Conclusiones y Trabajos Futuros

La principal novedad que planteamos con esta metodología consiste en *partir de unos conceptos que estamos interesados en estudiar* en vez de obtener los conceptos a partir de la clasificación de los documentos. Esto permite que la metodología se adapte mejor al estudio que planteamos. También se observa la necesidad de que la

metodología sea aplicada por un experto en el tema que se va a investigar, capaz de asignar de forma adecuada los ítems a cada concepto e interpretar los mapas de forma correcta. Estas operaciones realizadas por una persona que carezca del conocimiento suficiente sobre el tema investigado obtendrá resultados poco fiables.

Otra ventaja de la metodología es que en el caso de que la bondad de los mapas obtenidos sea menor de la esperada la metodología da la opción de modificar el conjunto de palabras claves de forma interactiva e ir mejorando la precisión de los mapas. En este sentido podemos indicar como desventaja que el proceso de asignación de estas palabras claves a los conceptos puede llegar a ser controvertido ya que la elección maliciosa de las palabras claves permitiría la obtención de resultados no fiables.

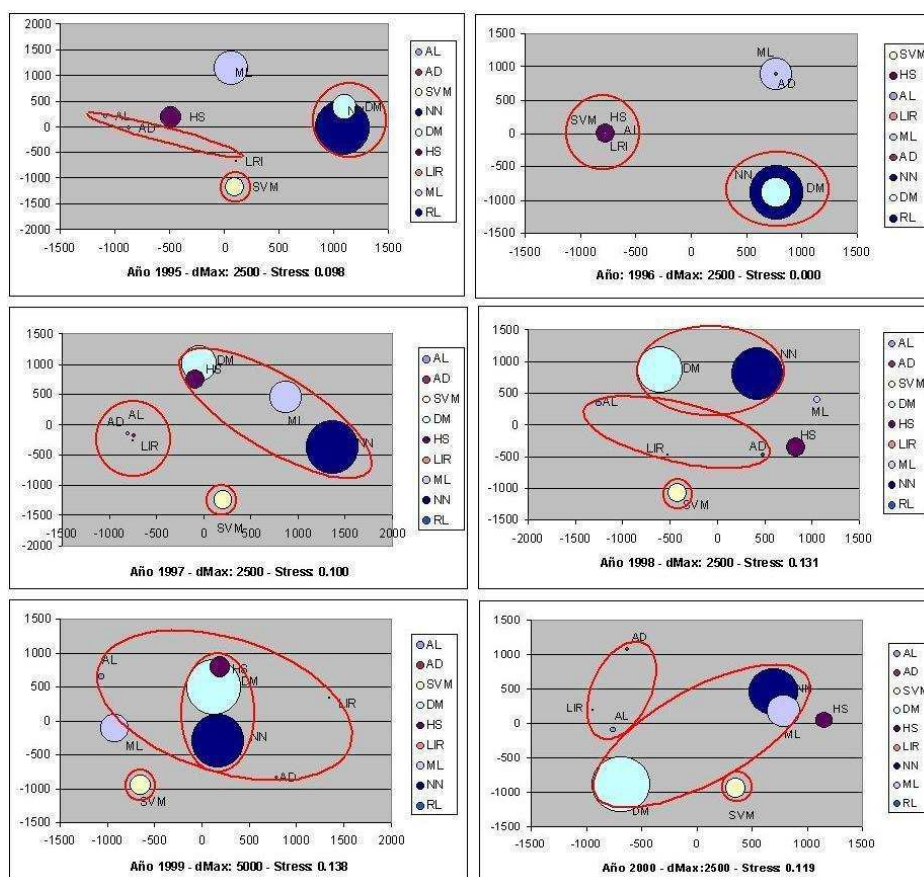


Fig. 5. Serie Temporal (1995-2000) Machine Learning

También se ha observado una característica de inconsistencia temporal de los datos peculiar de los buscadores en comparación con otras base de datos textuales, debido a

que a medida que pasa el tiempo pueden aparecer nuevos núcleos de documentos relacionados.

Como trabajos futuros nos planteamos buscar índices de impacto y afinidad más precisos y obtener técnicas que obtenga series temporales de mapas con los conceptos sincronizados.

	AL	AD	SVM	DM	HS	LIR	ML	NN	RL
1995	13	14	247	380	282	2	793	1951	0
1996	40	22	586	1463	553	5	1533	4151	0
1997	76	50	1102	3963	1076	15	3045	8091	0
1998	161	106	1974	9901	1804	17	374	12550	0
1999	267	158	3679	24386	3707	15	6932	22819	2
2000	1434	679	16984	122970	11417	70	41001	85706	1

Table 4. Evolución del impacto en el periodo 1995-2000

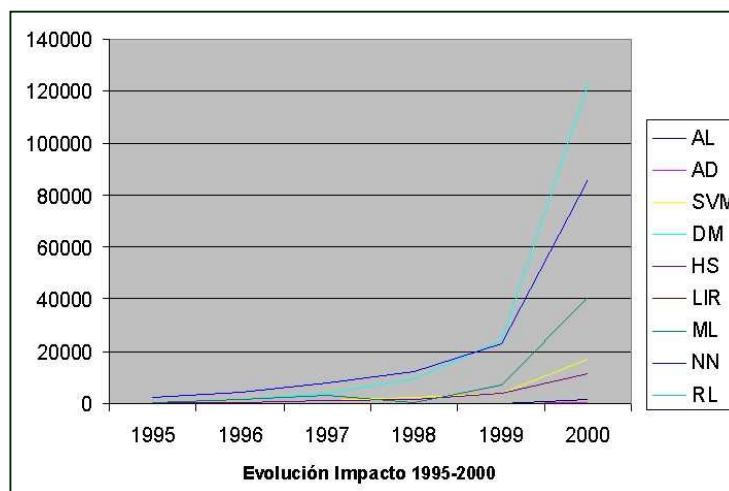


Fig. 6. Gráfica con la evolución del Impacto entre 1995-2000

Referencias

[JRD01] “Escalamiento Multidimensional”; José Eulogio Real Deus; 2001; Cuadernos de Estadísticas; La Muralla; 84-7133-700-X

- [**HK&JMB97**] "Data Visualization by Multidimensional Scaling: A Deterministic Annealing Approach." *Hansjörg Klock; Joachim M. Buhmann*; 1997
- [**AJ&WW98**] "Applied Multivariate Statistical Analysis"; *Richard A. Johnson; Dean W. Wichern*; 1998; 0-13-834194-X
- [**Lelu92**] "Hypertext paradigm in the field of information retrieval: a neural approach."; *Lelu Alain*; 1992
- [**Luc&Claire95**] "A workstation to classify, chart and analyze. Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique"; *Luc Grivel; Claire François*; 1995; Solarion; <http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2grivel.html>
- [**RRB&FCC98**] "Como Consultar Eficazmente una Base de Datos Bibliográfica. El Método de las Palabras Asociadas"; *Rosario Ruiz-Baños; Francisco Contreras-Cortés*; 1998; <http://www.ugr.es/~fccortes/curriculum/toledo.html>
- [**Rodriguez97**] "Valorando el impacto de la información en Internet: Altavista, el "Citation Index" de la Red"; *Josep Manuel Rodríguez i Gairín*; 1997
- [**Kruskal64a**] "Multidimensional Scaling: A numerical method"; *Kruskal, J.B.*; 1964; *Psychometrika* 29, 1-27
- [**Kruskal64b**] "Nonmetric multidimensional scaling: A numerical method"; *Kruskal, J.B.*; 1964; *Psychometrika* 29, 115-129
- [**Young72**] "A model for polynomial conjoint analysis algorithms"; *Young, F.W.*; 1972; *Multidimensional scaling: Theory and applications in behavioral sciences*. New York Seminar Press, Vol 1, pages 9-102.
- [**López96**] "Introducción a la Bibliometría"; *Pedro López López*; 1996; Promolibro; 84-7986-146-0
- [**Larson96**] "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure"; *Larson, Ray A.*; 1996; <http://sherlock.berkeley.edu/asis96/asis96.html>
- [**KYST**] <http://netlib.bell-labs.com/netlib/mds/kytsa.dos/kyts2a.exe.gz>
- [**Sherman72**] "Nonmetric multidimensional scaling: A Monte Carlo study of the basic parameters" *Sherman, C. R.* *Psychometrika*, 1972, 51, in press.
- [**Spence72**] "A Monte Carlo evaluation of three nonmetric multidimensional scaling algorithms" *Spence, I. A.* *Psychometrika*, 1972, 31, in press.
- [**CWTS**] Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands. "Mapping Scientometrics, Informetrics and Bibliometrics," *E.C.M. Noyns; A.F.J. van Raan* <http://sahara.fsw.leidenuniv.nl/cwts/cwtshome.html>