

Clasificación de términos mediante el algoritmo de Kohonen.

Vicente P. Guerrero Bote - vicente@alcazaba.unex.es
Cristina López Pujalte - clopez@alcazaba.unex.es
Cristina Faba Pérez - cfabper@alcazaba.unex.es
María J. Reyes Barragán - mjreyes@alcazaba.unex.es
Felipe Zapico Alonso - fzapalo@alcazaba.unex.es
Félix de Moya Anegón – felix@goliat.ugr.es

Facultad de Biblioteconomía y Documentación (Alcazaba de Badajoz), Universidad de Extremadura, 06071 Badajoz.

Se propone un método para hallar las relaciones contextuales entre los distintos términos presentes en una base. En primer lugar se emplea el Modelo Vectorial para representar los términos como vectores en función de los documentos en que aparecen. Y en segundo lugar dichos vectores son utilizados para entrenar una red de Kohonen, que genera una organización topológica de los mismos. Dicha organización genera clusters de términos que son organizados sobre una rejilla, de modo que además de relacionar cada término con los de su mismo cluster lo relaciona con los de los clusters cercanos.

1. Introducción

Actualmente ante la gran cantidad de información no normalizada de la que se dispone, está surgiendo la necesidad de procedimientos automáticos que permitan procesar tal cantidad de información. Uno de ellos, perseguido desde hace mucho tiempo, es la clasificación automática de los términos presentes en una base. Este tipo de tesauros generados automáticamente o "*aproximaciones al espacio conceptual*" (Chen et al., 1997), muy demandadas en bases con un contenido científico muy específico, sirve de ayuda para realizar búsquedas, para el browsing, o simplemente para representar el contenido de una base documental o de unos resultados. Existen dos formas de llevarlo a cabo:

- *Con independencia del corpus*: Son relaciones dependientes del lenguaje en el que están escritos los documentos. Se han hecho estudios desde la construcción de un Tesauro hasta el proceso del lenguaje natural. También se han empleado redes de palabras (WordNet) como base (Miller et al., 1995). Sin embargo, en cuanto a rendimiento, los resultados obtenidos no han sido los que se esperaban, al igual que en

Inteligencia Artificial, se necesitan relaciones específicas para el dominio de la base.

- *Con ayuda del corpus*: En este caso se trata de desarrollar las relaciones a partir de la información contenida en la base. El resultado serán relaciones específicas del dominio del corpus. Y se puede llevar a cabo evaluando:

- La *similitud léxica* de las palabras, para lo cual se emplea técnicas de: truncajes (se consideran relacionadas las palabras que tengan la misma raíz), eliminación de sufijos (relacionan palabras tras eliminar sufijos y prefijos), n-gramas (relacionan palabras que contienen n-gramas parecidos), etc. Este tipo de técnicas dan buen resultado en los corpus en los cuales los términos tienen una fuerte relación léxica, por ejemplo los relacionados con la química y productos farmacéuticos (Oakes & Taylor, 1999).

- La *similitud semántica*, que solamente se puede llevar a cabo mediante estudios de coocurrencia. En algunos casos se han utilizado para ello técnicas del proceso del lenguaje natural (Strzalkowski, 1995) y también se están comenzando a utilizar redes neuronales (Lin, 1997; Orwig,

Chen, & Nunamaker, 1997; Chen et al., 1998; Kohonen et al., 1999; Lagus & Kaski, 1999; Lagus et al., 1999; Moya, Herrero & Guerrero, 1998; White, Lin & McCain, 1998).

En este trabajo, nosotros vamos a proponer y contrastar un procedimiento automático que utilizando un corpus documental permita hallar las relaciones semánticas contextuales entre los términos. Evidentemente estas relaciones solamente serán válidas para dicho corpus.

2. Metodología

El problema se divide en dos partes, en primer lugar el método utilizado para clasificar, que hasta ahora han sido principalmente métodos de clustering, tanto jerárquicos como heurísticos (Salton & McGill, 1983). En segundo lugar, estos métodos necesitan una representación vectorial de los objetos que se pretenden clasificar, o lo que es lo mismo, una representación como listas de características, por lo que hay que estudiar una representación documental de este tipo.

2.1 Selección y representación de términos

Con el fin de probar las capacidades de este algoritmo en condiciones tan generales y cercanas a la realidad como sea posible, se ha generado una base de datos de prueba. Hemos tenido optar por una base en lengua inglesa al no disponer de un algoritmo reducción a la raíz ampliamente aceptado en castellano. Por estos motivos, hemos decidido utilizar registros de la base bibliográfica LISA para formar nuestra base de prueba. Nosotros hemos considerado cada uno de los resúmenes que contienen los registros como documentos independientes, del resto de los campos no hemos tenido en cuenta ninguno. Con objeto de que sea lo más general posible, en lugar de hacer una recuperación por temas, hemos seleccionado los últimos 954 registros de LISA (Versión Summer 96). La recuperación la hacemos seleccionando las referencias que tienen un *Accession number* mayor o igual que 9605000. De este modo, la base generada se compone de los mencionados 954 registros que contienen un total de 7758 palabras diferentes.

La cifra de términos presentes en la base nos resulta inabordable desde el punto de vista del

cálculo, de forma que se hace necesario una reducción ante lo que decidimos quedarnos con los términos más representativos de la base. Para este fin, empleamos el Valor de Discriminación de Salton (Salton & McGill, 1983).

Para ello, una vez que tenemos creada la base documental, el siguiente paso es transformar los documentos en vectores de modo que podamos calcular el mencionado valor de discriminación de los términos. Por este motivo, aplicamos el modelo del espacio vectorial que transforma cada documento en un vector.

Como esquema de pesos se escoge uno muy cercano al clásico *IDF* (denominado *tfidf*), que en el estudio de Noreault et al. (1981) es uno de los que mejores resultados obtienen. Quedando así el peso de cada componente:

$$a_{ij} = t_{ij} \log \frac{F}{f_j}$$

donde:

a_{ij} = peso asignado al término t_j en el documento D_i .

t_{ij} = número de veces que aparece el término t_j en el documento D_i .

f_j = número de veces que aparece el término t_j en toda la base.

F = número total de palabras (repetidas o no, tokens) de toda la base.

Hemos decidido utilizar este esquema en lugar del clásico *IDF*, porque éste último le asigna los pesos mayores a aquellos términos que aparecen en un solo documento. Sin embargo, dado que nosotros queremos llevar a cabo una *clasificación y organización topológica*, que muestre las relaciones semánticas existente entre todos los términos, los que aparecen en un documento solamente están relacionados con los del mismo documento, para lo que no parece necesaria ninguna clasificación ni organización. De este modo, si calculamos dicho valor de discriminación de todos los términos y los ordenamos en orden decreciente de este valor, entre los primeros clasificados hay una reducción drástica en el número de términos de frecuencia uno (entre los primeros 100 no tenemos ninguno). También existe una gran reducción de los que aparecen en un solo documento (no tan grande como la realizada para los de frecuencia uno).

Hasta ahora nos hemos preocupado de generar un conjunto de vectores documentales para poder calcular el valor de discriminación de cada término. Pero, si ponemos todos estos vectores horizontalmente y alineados veríamos que forman una matriz en la que todos los elementos de una fila corresponden a un mismo documento, mientras que todos los de una columna lo hacen a un término. Es decir, cada elemento indica el peso del término de la columna en la que se encuentra, dentro del documento de la fila correspondiente.

De esta manera si podemos coger las filas para representar a los documentos, serían los vectores documentales, también podremos tomar las columnas para representar a los términos, con lo que un término también se puede representar vectorialmente.

Si los vectores documentales de los que partimos son binarios, se generarían vectores de términos también binarios equivalentes. En la ponderación *tf-idf* cada componente consta de dos partes, una local, que indica la importancia del término en cada documento (y que variaría con la frecuencia en cada uno), y otra global, que indica la importancia del término en el total de la base (y que sería la misma para todas las coordenadas de un término). Por otro lado, siguiendo el estudio de Noreault et al. (1981), las mejores funciones de similitud son aquellas que utilizan medidas angulares, con lo que no sirve de nada la componente global del *tf-idf* (que solo modifica el módulo del vector), y estaríamos, por tanto, en una situación equivalente a la utilización de la frecuencia en el documento para cada componente.

Por este motivo, igual que se hace en los vectores documentales para los términos, se pueden generar diferentes pesos de documentos, basados por ejemplo, en el número de términos que contienen, los que contienen muchos es lógico que pesen menos que aquellos que contienen pocos. De esta forma se puede utilizar un esquema de pesos similar al *tf-idf* para los términos que sería:

$$a_{ij} = t_{ij} \log \frac{D}{ND_i}$$

donde:

a_{ij} = peso asignado al documento D_i en el término t_j .

t_{ij} = número de veces que aparece el término t_j en el documento D_i .

ND_i = número de términos que aparecen en el documento D_i .

D = número total de términos de toda base.

Una vez que tenemos los términos representados como vectores, de igual forma que se calculan similitudes entre documentos utilizando funciones definidas para tal fin, también se pueden calcular similitudes entre términos mediante las mismas funciones.

2.2. Algoritmo de Kohonen

En la actualidad, están tomando un especial auge las redes neuronales artificiales, como mecanismos para llevar a cabo ciertas tareas intelectuales básicas. Las mencionadas redes han sido diseñadas, emulando el funcionamiento del cerebro, para aprender a asignar salidas multidimensionales a entradas multidimensionales, lo que llevan a cabo con una gran capacidad de generalización.

Su entrenamiento puede ser *supervisado* o *no-supervisado* (según el tipo de red). En el *supervisado* se emplea una serie de pares formados por una entrada y su correspondiente salida, de modo que la red se va adaptando a las salidas deseadas a partir de los errores que va cometiendo. Mientras que en el *no-supervisado*, se sirven solamente las entradas (sin sus correspondientes salidas), y la red hace un clustering de las entradas. Estos clusters tienen una serie de ventajas sobre los clásicos. Por un lado son *adaptativos*, es decir, no se le tienen que facilitar parámetros de tamaño, solapamiento, etc., sino que en gran medida se adaptan a la base en cuestión, por lo cual podemos decir que aprenden de los objetos que tiene que tratar. Una de las consecuencias de esto es que los clusters se encontrarán más juntos allá donde hay más densidad de términos. Así como, la zona que cubre un cluster puede aumentar o disminuir en función de la anteriormente mencionada densidad de términos. Para ello, se lleva a cabo una comparación global de los vectores, capaz de incluir en un cluster términos que no aparecen exactamente en los mismos documentos.

Uno de los modelos utilizados para la clasificación de términos ha sido la red de Hopfield empleada para asistir a los usuarios, memorizando o generando tesauros, listas de encabezamiento, etc., para luego sugerir términos y rehacer la pregunta (Chen et al., 1997). El modelo de Kohonen ha sido

empleado por el mismo Teuvo Kohonen para la clasificación de términos, consiguiendo clasificarlos por la función sintáctica que desempeñan en los textos (Kohonen et al., 1999; Lagus & Kaski, 1999; Lagus et al., 1999), simplemente utilizando la información derivada del término anterior y posterior, o para la clasificación de palabras clave de una base (Moya, Herrero & Guerrero, 1998).

En nuestro trabajo nos hemos centrado en el último de estos algoritmos, el modelo de Kohonen, que permite hacer tanto un clustering como una clasificación topológica (White, Lin & McCain, 1998; Guerrero, Moya & Herrero, 2002a). Nos hemos propuesto aplicarlo a la clasificación semántica de términos, para lo cual también hemos tenido que estudiar un procedimiento de selección de términos representativos de la base, así como de representación vectorial de los mismos.

La simulación hardware conlleva la creación de una capa competitiva de cierta complejidad, donde cada neurona ejerce una influencia sobre el resto de las neuronas de su capa que va a ser función de la distancia entre las mismas, es decir, cada neurona ejerce una influencia positiva sobre sí misma y sobre las neuronas topológicamente cercanas. Esta influencia va decreciendo a medida que aumenta la distancia entre las neuronas, hasta hacerse negativa, para tener finalmente una influencia positiva sobre las más alejadas. Como consecuencia de esto en la capa se da una burbuja de actividad, formada por todas aquellas unidades que están cercanas a la ganadora, las cuales participan del refuerzo correspondiente al aprendizaje.

Sin embargo, el proceso que lleva a cabo durante el aprendizaje cada vez que se le presenta un vector, lo podemos resumir en los siguientes pasos:

- *Seleccionar como nodo ganador (cada nodo de la red está representado por un vector de pesos de la misma dimensión que la entrada) el más cercano a la entrada presentada (para ello, típicamente se emplea la distancia euclídea).*
- *Modificar el vector de pesos del nodo ganador y los nodos correspondientes a su vecindad acercándolo hacia el vector de entrada (en algunos casos el refuerzo es igual para toda la vecindad, en otros decrece al aumentar la distancia al ganador).*

Tras el entrenamiento las neuronas topológicamente cercanas resultan ganadoras

con clusters de vectores que son cercanos en el espacio de entrada. En este caso las neuronas han sido dotadas de conciencia. Dicho mecanismo reduce la probabilidad de que una neurona pueda ganar un gran número de competiciones, utilizado para evitar el problema del vector pegado (Guerrero & Moya, 2001; Guerrero, Moya & Herrero, 2002a, 2002b), permitiendo que las victorias sean compartidas por todas la neuronas de la red.

Debido a esta organización topológica a veces lo único que interesa es el clustering llevado a cabo por la capa oculta, y se selecciona todo el conjunto de vectores para el entrenamiento con el único motivo de ver la organización topológica resultante. Dentro de esta última aplicación existen dos posibilidades, por un lado si se tienen más unidades ocultas que vectores de entrenamiento lo que se consigue es una proyección óptima sobre la topología que se elija. Si el número de unidades es menor al número de vectores lo que se consigue es una capa que hace clustering y ordena cada cluster topológicamente. El número de clusters resultantes será igual al de neuronas que formen la capa oculta.

Recientemente se han utilizado redes de Kohonen para la generación de *mapas topológicos* de un conjunto de documentos, etiquetando incluso las zonas de influencia de cada palabra o término (Lin, 1997; Chen et al., 1998; Kohonen et al., 1999; Lagus & Kaski, 1999; Lagus et al., 1999; Guerrero & Moya, 2001; Guerrero, Moya & Herrero, 2002a, 2002b), o para el análisis de dominio (White, Lin & McCain, 1998).

3. Resultados

Los vectores generados han sido aplicados a una red de 20x20 neuronas en la capa oculta y, por tanto, clusters. El resultado completo resulta difícil de mostrar en una figura, de modo iremos mostrando parcialmente algunas partes de interés.

En realidad, sería más correcto hablar de raíces de palabras en lugar de términos, ya que se ha practicado la reducción a la raíz antes de ser aplicados a la red. Dentro de un mismo cluster los términos se han ordenado en función del grado de pertenencia a ese cluster que devuelve el módulo de fuzzificación (Guerrero y Moya, 2001).

Como son muchas celdas y términos, vamos a comentar aquellas partes de la clasificación

que nos parezcan más representativas Así en primer lugar nos vamos a fijar en la esquina inferior izquierda que podemos ver ampliada en la figura 1. En ella podemos ver como en el cluster de la posición (19,19), y en algunos de los más cercanos ((19,18) y (18,19)), aparecen términos (mejor dicho raíces) relacionados con la medicina como *physician* (médico), *patient* (paciente), *clinic* (clínica), *medic* (médico, estudiante de medicina), *biomed* (biomedicina), *care* (cuidado), *health* (salud). Aunque también aparecen otros como *informat* (informática), sin embargo, esto es relativamente lógico, debido al carácter instrumental de la informática puede aparecer aquí como en cualquier otra parte donde se utilice la misma.

facilit, optim, tool, nlm, evid	physician, patient, clinic	18
medic, medicin, biomed, partii, ebm	health, care, informat	19
18	19	

Figure 1: Ampliación de la esquina inferior derecha de la organización topológica, contiene términos relacionados con la medicina y el cuidado de la salud.

Otra parte bastante opuesta la podemos ver en la figura 2. Donde podemos ver como aparece un gran número de términos relacionados con el World-Wide Web, como son por ejemplo: *java*, *hotjava*, *sun*, *browser*, *client*, *mosaic*, *html*, *markup*, *hypertext*, *map*, *hypermedia*, *navig*, *interact*, *sgml*, *webmap*, *server*, etc., aunque también vemos que aparecen otras palabras con una relación secundaria y muy circunstancial.

	7	8	9	10	11	12	13	14
0	sun, java, hotjava	devic, pda, browser	client, distribut	chicago, illinoi, mosaic, second		html, macro, word, markup	hypertext	schema, brows, exploit, conceptu
1	realiti, assist	germani, carri, uniform	usabl, week, converg	server, collabor	webmap, immedi, window, frog, file	graphic, automat, sgml	interac, expres, manipul, abil	map, hyperme dia, semant, navig, illustr

Figure 2: Ampliación de otra zona de la organización topológica, relacionada con el World-Wide Web.

Muy cercana a esta zona, como se muestra en la figura 3, existen algunas celdas que contienen términos político-religiosos relacionados con Asia. No podemos decir que esta zona sea en exclusividad la que contiene términos de este tipo, existe una zona cercana aparentemente relacionada con la economía capitalista y los países democráticos, donde aparecen términos relacionados con China como podemos ver en la figura 4.

	2	3
0		hong, kong, proceed, asian
1	moslem, ac, com	southeast, republ, asia, singapor

Figure 3: Ampliación de una zona de la organización topológica que contiene términos político-religiosos relacionados con Asia.

Por último vamos a mostrar un área donde aparecen términos relacionados con la Biblioteconomía y la documentación en la figura 5.

	0	1	2
3	western, economy, low	china, internetwork, chinese, pose, market	financi, country, connect, company
4	profit	cee, eastern, overview, cash	
5	payment, europa	foundat, opportun, aslib, secur	copyright, ronchei
6	retrospect, conver	tackl, gateway, theme	
7	ensur	societi, democrat, easi, deliveri	workplac, highlight
8	danish, realis, ministri	australian, nordic, uncov, readi, erin	polit, administr, feder
9	self, housekeep, fast	initi, competit	hour
10	darmstadt, hess, allegro	easili, expand, member	superhighwa i

Figure 4: Ampliación de otra zona de la organización topológica relacionada con la economía capitalista y los países democráticos.

	16	17	18	19
13	referenc, mark, rapid, analys		edition, revision, japanes	cite
14	rule, inqueri, iso, failur, congress	canadian, descript, archivist, entri	format, appear, classif, danbib	aacr, marc, quarter, festschrift, centuri
15	minim, mlc, less, whole, risk		index, bibliograph, descrip	

Figure 5: Ampliación de una zona de la organización topológica relacionada con la Documentación.

4. Conclusiones

Como encabezábamos el trabajo, la extracción de relaciones entre los términos es algo buscado desde hace mucho tiempo. Cuando esto se ha hecho de manera general para toda una lengua no se han obtenido los resultados deseados, lo que hace que siempre se hayan demandado para bases en particular.

El procedimiento diseñado ha demostrado ser bastante útil, ya que hemos podido observar que descubre bastantes relaciones semánticas entre términos. Esto lo hace a partir de la información presente en la misma base, de forma que los resultados están especialmente indicados para el corpus correspondiente a esa base. Al mismo tiempo, dado que la red hace una especie de análisis de coocurrencia (aunque diferente), nos encontramos con algunas relaciones anómalas debidas principalmente al pequeño tamaño de nuestra base que disminuirá a medida que crezca (igual que ocurre con otros métodos que utilizan la coocurrencia).

A ello tenemos que sumar, la gran cantidad de información que ofrece el algoritmo de Kohonen, que no solamente es capaz de realizar el agrupamiento de términos cercanos, sino que también ofrece una organización topológica de los grupos creados. Por este motivo son muchas las posibilidades que ofrece en Documentación, puede permitir ampliar una búsqueda incluyendo los términos del cluster

correspondiente, o incluso los de toda una zona. Y también tiene una gran utilidad para el browsing, ya que representa de alguna forma el contenido de la base.

Aunque los tiempos de respuestas que nosotros hemos obtenido son grandes, una de las características de estos algoritmos es la posibilidad de ejecutarlos en paralelo, pudiendo en ese caso mejorar los tiempos varios miles de veces. El coste de las máquinas que permiten esto no es excesivamente alto y es de suponer que bajará.

5. Reconocimientos

Este trabajo ha sido financiado por la Junta de Extremadura-Consejería de Educación Ciencia y Tecnología y el Fondo Social Europeo, como parte del proyecto de investigación IPR99A047.

6. Bibliografía

- Chen, H., Houston, A.L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49, 582-603.
- Chen, H.; Ng, T.D.; Martinez, J., & Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. *Journal of the American Society for Information Science*, 48, 17-31.
- Chen, H.; Ng, T.D.; Martinez, J., & Schatz, B.R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. *Journal of the American Society for Information Science*, 48, 17-31.
- Guerrero, V.P., & Moya-Anegón, F. (2001). Reduction of the Dimension of a Document Space using the Fuzzified Output of a Kohonen Network. *Journal of the American Society for Information Science*, 52, 1234-1241.
- Guerrero, V.P.; Moya-Anegón, F., & Herrero-Solana, V. (2002a). Document organization using Kohonen's algorithm. *Information Processing & Management*, 38, 79-89.
- Guerrero, V.P.; Moya-Anegón, F., & Herrero-Solana, V. (2002b). Automatic extraction of relationships between terms by means of Kohonen's algorithm. *Libri* (en prensa).
- Kohonen, T.; Kaski, S.; Lagus, K.; Salojärvi, J.; Honkela, J.; Paatero, V., & Saarela, A. (1999). Self organization of a massive text document collection. In Oja, E. and Kaski, S. (Eds.) *Kohonen Maps* (pp. 171-182). Amsterdam, Holland: Elsevier.
- Lagus, K. & Kaski, S. (1999). Keyword selection method for characterizing text document maps. *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)* (pp. 371-376). London : Institution of Electrical Engineers.
- Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*, 13, 345-364.
- Lin, X. (1997). Maps Displays for Information Retrieval. *Journal of the American Society for Information Science*, 48, 40-54.
- Miller, G.A.; Beckwith, R.; Fellbaum, C.; Gross, D. & Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 235-244.
- Moya-Anegón, F.; Herrero-Solana, V., & Guerrero, V.P. (1998). Virtual reality interface for accessing electronic information. *Library and Information Research News*, 22, 34-39.
- Noreault, T.; McGill, M. & Koll, M.B. (1981). A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In Oddy, R. N.; Robertson, S. E.; Van Rijsbergen, C. J.; Williams P. W., (Eds.), *Information Retrieval Research: Papers given at the 1st Joint British Computer Society (BCS) and Association for Computing Machinery (ACM) Symposium: Research and Development in Information Retrieval* (pp. 57-76). London, UK: Butterworths.
- Oakes, M.P., & Taylor, M.J. (1999) Clustering of Thesaurus Terms Using Adaptive Resonance Theory, Fuzzy Cognitive Maps and Approximate String-Matching Techniques. Liverpool, UK: The University of Liverpool.
- Orwig, R.E.; Chen, H., & Nunamaker, J.F. (1997). A Graphical, Self-Organizing Approach to Classifying Electronic Meeting

- Output. *Journal of the American Society for Information Science*, 48, 157-170.
- Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management*, 31, 397-417.
- White, H., Lin, X., & McCain, K.. (1998). Two modes of automated domain analysis: multidimensional scaling vs. Kohonen feature mapping of information science authors. In Mustafa el Hadi, W., Maniez, J., & PollitErgon Verlag, S. (Eds.), *Structures and relations in knowledge organization: Proceedings of the Fifth International ISKO Conference, Lille, France* Proceedings of the 5th International ISKO Conference. Würzburg, Germany: Ergon Verlag.