ANALES DE DOCUMENTACIÓN, N.º 5, 2002, PÁGS. 273-286

APLICACIONES AL ANÁLISIS AUTOMÁTICO DEL CONTENIDO PROVENIENTES DE LA TEORÍA MATEMÁTICA DE LA INFORMACIÓN

José Antonio Moreiro González[*]

Departamento de Biblioteconomía y Documentación. Universidad Carlos III de Madrid.

Resumen: Reflexión sintética para revisar las propuestas más relevantes que, siguiendo la teoría matemática de la comunicación de Shannon y Weaver, hayan afectado a los procedimientos del análisis automático del contenido documental. Partiendo del empleo de la teoría matemática en Ciencia de la Información se explican sus aplicaciones metodológicas en nuestra especialidad, en especial respecto a las técnicas de recuperación de la información. Para después describir los modelos matemáticos aplicados al análisis automático del contenido: leyes de Zipf y Goffman, antidiccionarios para índices permutados, Indización Estadística de Términos por Frecuencias, algoritmos n-grams y de stemming, así como los referidos a los métodos de agrupación y clasificación como clusters por valor de discriminación y por relevancia de los términos como son los métodos de agrupación basados en Grafos Teóricos, los basados en Centros de masas, el algoritmo K-vecinos o K-medias, el K-vecinos axial o incremental, y el algoritmo ISODATA. Para luego exponer los clasificadores cienciométricos como el método de Chen y finalmente los métodos con sistemas de aprendizaje.
Palabras clave: Análisis de contenido textual. Análisis automático. Elementos matemáticos. Métodos estadísticos. Métodos probabilísticos. Redes neuronales. Coocurrencias. Métodos basados en centroides. Clustering.

Abstract: This paper analyzes the most important proposals following the Shannon and Weaver's Mathematic Theory of Communication that have influenced in proceedings of automatic content analysis. It's explained the methodological applications of this theory in our discipline, especially about information retrieval. After this, describes the mathematical models applied to automatic content analysis: Laws of Zipf and Goffman, anti-dictionaries to permuted indexes, Statistical Indexation of terms by frequencies, n-grams and stemming algorisms. Also studies the methods of relation and classification like clusters by value of discrimination and by relevance of terms: for example, methods of relations based in Graph Theory, mass core, the K-means or incremental K-means, and the ISODATA algorism. Finally, explains the scientometrics indicators as Chen's coowording and methods with learning systems.
Keywords: Textual content analysis. Automatic analysis. Statistical methods. Probabilistic methods. Neural nets. Co-occurrences. Core methods. Clustering.

[*] jamore@bib.uc3m.es

INTRODUCCIÓN

La propuesta de analizar la información desde unidades mensurables ha sido fructífera en el campo de la ingeniería de sistemas de comunicación, pero presenta algunos problemas en lo referente al procesamiento de la información, si no es combinado con métodos lingüísticos. Si la teoría matemática ayudó a que el concepto de información y su tratamiento fuese objeto de innumerables estudios en Documentación, originados casi siempre dentro de la *American Society for Information Science*, podemos afirmar que, en general, los resultados que ha producido tienen que compatibilizarse con métodos semánticos si se quieren obtener aplicaciones válidas.

La primera teoría de la información surgió de la propuesta de Shanon y Weaver con el propósito de fijar un modelo de entropía sobre la suma de información requerida en una situación dada para eliminar la incertidumbre[1]. La información para ellos era una medida de libertad de elección al seleccionar un mensaje desde una fuente dada. Shannon y Weaver, ingenieros, buscaban un concepto de información formalizado, que pudiese expresarse en medidas. Nuestro propósito es revisar las propuestas más representativas sucesivas a la concepción de Shannon y Weaver, y que hayan tenido como destino el análisis automático del contenido documental.

Emplear una Teoría de la Comunicación, de carácter eminentemente mecánico, a una especialidad en la que tiene gran importancia la significación de los mensajes transmitidos es tarea limitada y dificultosa. Lo que no ha impedido que, más de cincuenta años después de su definición, la teoría matemática de la comunicación siga siendo aceptada o rechazada de acuerdo con aplicaciones concretas. Las medidas de la información han sido útiles para su aplicación a la recuperación documental, así como para comparar documentos, fijar nociones, hacer mediciones, y desde luego, para el análisis de contenido automático[2].

1. EL EMPLEO DE LA TEORÍA MATEMÁTICA DE LA INFORMACIÓN

En la teoría de Shannon y Weaver la cantidad de información contenida en un mensaje se define en función de la frecuencia relativa de utilización de los diferentes símbolos que lo componen:

a.- Los mensajes son transmitidos desde la fuente al usuario por una vía de comunicación,

b.- para que el mensaje pueda recorrer esa vía debe ser codificado,

c.- y luego, descodificado para que lo comprenda convenientemente el destinatario.

El problema está en la transición de los símbolos del mensaje que entró a los del mensaje que salió. Esta posibilidad de imperfección se llama ruido. Sin ruido, la canti-

[1] Shannon, C.E. y Weaver, W.- *The mathematical theory of Communication*. Urbana: University of Illinois Press, 1949.

[2] Ellis, D.-The effectiveness of information retrieval systems: the need for improved explanatory frameworks, en *Social Sciences Information Studies*, 1984, 4, nº 4: 265.

dad de información de un mensaje es la misma a la salida que a la entrada. Con ruido nacen la ambigüedad y los equívocos. Para evitarlos habrá que transmitir el mensaje con redundancia, aunque esto suponga una pérdida relativa de información. La principal objeción que desde el primer momento presentó su *Teoría matemática de la Comunicación* fue la de no considerar los aspectos relativos al significado de los mensajes, por lo que debemos considerar el cuerpo especulativo al que abrieron paso como una teoría de señales, no como una auténtica teoría de la información[3].

Aún manteniendo una postura de equilibrada duda al contemplar que las aplicaciones hechas con efectividad se habían limitado a fenómenos particulares, Jean-Bernard Marino analizó la posibilidad de nuevas aplicaciones de cada una de ellas, principalmente a través de las bases de datos accesibles. Distribuyó en tres bloques las aplicaciones de la teoría matemática[4]:

1. *Indización mediante tarjetas perforadas*: en la década de 1950 Garfield indizó documentos biomédicos mediante tarjetas perforadas. Los codificó de tal manera que el número de perforaciones coincidía con la frecuencia de uso de los descriptores en el total del glosario. Los descriptores más utilizados recibían así la codificación más breve.
2. *Evaluación de los resultados de un sistema documental*: se trata de desligar el sistema de salida del sistema de entrada, transmitiendo por una vía con ruido. Los mensajes recibidos tenían una triple codificación y su probabilidad de ser recuperados dependía de una tabla de contingencias. Fue utilizado por Meetham, Belzer, Cawkel y Guazzo.
3. *Indización por frases*: Briner aplicó los conceptos de la teoría matemática a los componentes gramaticales de un texto escrito, deduciendo una capacidad de transmisión del conocimiento por palabra análoga a la fórmula que cuantifica la capacidad de una vía. Para las palabras ambiguas Briner amplió el principio a indización de la frase entera que las contenía.

Buscando identificar las leyes que rigen los fenómenos informativos, Zunde y Gehl analizaron otras aplicaciones de carácter empírico. Justificaban así su búsqueda de explicaciones desde la línea matemática[5]:

> *"El objeto de estudio de la ciencia de la información son fenómenos empíricos asociados con procesos de información tales como la generación, transmisión, transformación, condensación, almacenamiento y recuperación. El propósito último consiste en alcanzar una comprensión mejor de la naturaleza de la información".*

[3] Fox, C.J.- *Information and misinformation: an investigation of the notions of information, misinformation, informing and misinforming*. London: Greenwood Press, 1983: 58-60.

[4] Marino, J.B.- Quelques applications de la théorie mathématique de la communication en Sciences de l'information, en *Documentaliste*, 1983, 20, nº 2: 60.

[5] Zunde, P. y Gehl, J. - Empirical foundations of Information Science, en *Annual Review on Information Science and Technology*, 1979, 14: 79.

Precisamente las dudas que existen sobre el nivel científico de nuestra especialidad tienen su origen en que las teorías son las últimas que se desarrollan dentro de los problemas y principios racionales de esta ciencia. Entre estas teorías fundamentales, las más antiguas fueron las de Zipf, Bradford y Lotka, revisadas en 1969 por Fairthorne, que originó con ellas el modelo de distribución hiperbólica de la información, cuya expresión generalizada es la *cumulative advantage*[6]. Más tarde, Price reformuló esta teoría sobre la premisa de que ciertos procesos informativos se pueden explicar a partir de que el éxito cria éxito[7]. Explicó de esta manera la *cumulative advantage* de la que son casos limitados las formulaciones matemáticas de Zipf, Lotka y Bradford, con las que se compatibilizan la mayoría de los resultados experimentales sobre análisis de frecuencia de citas. Sin embargo, algunas de estas teorías no están definitivamente demostradas, como confirma Coile sobre la ley de productividad científica, que no puede mantenerse para la bibliografía de Humanidades o de nuestra propia especialidad[8].

Los métodos matemáticos han sido el centro metodológico en nuestra especialidad a la hora de definir las técnicas de recuperación de la información. Si las más clásicas son las fijadas sobre la teoría del álgebra de Boole que establece un sistema binario {1,0} en el que se que utilizan los conectores {Y, NO, O}, tras la intervención de métodos estadísticos y probabilísticos se ha afianzado la actuación de métodos vectoriales, en los que los documentos se representan a través de vectores: en una colección de N documentos, en el que existe un total de n términos, representamos cada documento por un vector de n componentes[9]. Se actúa por el grado de similitud de la ecuación de búsqueda mediante la similitud geométrica del coseno. El más representativo es el método IDF que se sirve, como luego veremos, de la ponderación de los términos.

Se utilizan también métodos derivados de los conjuntos borrosos en la idea de representar, para una combinación de dos términos, el número de veces que estos coinciden en un documento. Mediante matrices sin normalizar o mediante matrices de correlación normalizada, se obtiene el valor de la pertinencia de un documento sobre un término.

2. MODELOS MATEMÁTICOS APLICADOS AL ANÁLISIS AUTOMÁTICO DEL CONTENIDO

Centrándonos en nuestro objeto, en los procesos de recuperación de información es preciso realizar análisis léxico textual de los documentos afectados. Los documentos a analizar pueden ser a texto completo, sus resúmenes, e incluso títulos o listas de térmi-

[6] Fairthorne, R. - Empirical hyperbolic distributions (Bradford, Zipf, Mandlebrot) for Bibliometric description and prediction, en *Journal of Documentation*, 1969, 25, nº 4: 319-343. Sobre la utilización de métodos estadísticos y probabilísticos véase el apartado 4.4.5.1. de este libro.

[7] Price, D. J.- A general theory of Bibliometric and other cumulative advantage processes, en *Journal of the ASIS*, 1976, 27, nº 5: 292-306: "Un trabajo que ha sido citado muchas veces es más fácil que sea citado de nuevo que uno que lo haya sido raramente. Un autor de muchos trabajos tiene más posibilidades de publicar nuevamente que uno que haya sido poco prolífico. Una publicación periódica que haya sido consultada con frecuencia, es más fácil que sea consultada de nuevo que una de uso menos frecuente".

[8] Coile, R.- Lotka's frequency distribution of scientific productivity, en *Journal of the ASIS*, 1977, 28, nº 6: 366-370.

[9] Salton, G.-*Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. New York: Addison-Wesley, 1989.

nos. A partir del texto de entrada, el proceso de análisis de texto produce un documento que lo representa en una forma que pueda ser interpretada por el ordenador para poder ser reutilizada en un proceso de recuperación de información[10].

Cuando comenzaron a aplicarse la informática y la estadística a la Documentación, en los años 50 y 60, se diseñaron unos modelos que hacían un análisis automático del texto para tratar de averiguar cuales eran los términos relevantes de un documento, partiendo de la comprobación hecha por Zipf en 1935 demostrativa de que la frecuencia de aparición de una palabra en un texto es inversamente proporcional a la posición que ocupa en el ranking de frecuencias de palabras de un texto, de forma que el producto de ambas variables es una cantidad que se puede aproximar por la constante[11]:

$$F \times r = C.$$

donde frecuencia es el número de veces que se repite una palabra en el documento, y rango el valor que corresponde a cada palabra ordenadas de mayor a menor frecuencia. El aumento del rango implica una disminución de la frecuencia. La ley de Zipf se aplica de acuerdo con esta metodología:

- Ordenación decreciente de las palabras.
- Multiplicación de la frecuencia x rango.
- Obtención de la media para términos de frecuencias iguales, que tiene como efecto disponerlos en orden alfabético.
- Se elegirían como palabras de indización aquellas que tuviesen una frecuencia de aparición superior a la constante C determinada por esa ley.

A la hora de aplicar esta ley hay que considerar que los textos científicos presentan menor diversidad de vocabulario que los textos literarios, y que un aumento en el tamaño del texto supone un comportamiento en el que la magnitud (F x r) se aleja de la constante. Es útil recordar que cuando se trabaja con raíces de palabras se reduce considerablemente el tamaño del texto a tratar, con una paralela reducción del tamaño de la estructura de los índices, ya que las raíces son más frecuentes que las palabras, lo que facilita la búsqueda.

Zipf influyó en Luhn, con sus propuestas para la constitución de índices permutados. Consistía en eliminar las palabras no significativas mediante la confrontación del documento con una lista negativa o antidiccionario construida previamente. Luhn no consideró aún en esta primera aplicación las hipótesis de Zipf, y se atuvo a un principio muy simple: cada una de las palabras que quedaban, las no vacías, se convertía en un elemento de entrada temática al índice. Lo importante de su aplicación fue que había dado comienzo a los procesos de filtrado, cuyo objetivo era eliminar previamente los términos

[10] Baeza-Yates, R. y Ribeiro-Neto, B.- *Modern Information Retrieval*. London: Addison Wesley Longman, 1999.

[11] Zipf, G. K.- *Human Behavior and the Principe of Least Effort: an Introduction to Human Ecology*. New York: Haffner, 1948: "Si se prepara una tabla del conjunto de palabras de un texto cualquiera, clasificadas por orden de frecuencia decreciente, se constata que el producto que resulta de multiplicar las frecuencias (f) de observación de las palabras de los textos por el valor numérico (r) del rango que ocupan estas palabras en una distribución de frecuencias de observación, es constante. F x r = C".

no preferentes por confrontación con una lista de palabras vacías, y que, de alguna manera, se han mantenido en los sistemas de indización automática hasta el presente. La palabras vacías pueden también suprimirse a posteriori, eliminándolas si consiguen eludir el proceso de filtrado. Para asegurarlo se introdujo también una lista o diccionario de palabras significativas, para comparar positivamente las palabras restantes del texto candidatas a ser entradas de indización.

Para obtener los posibles términos candidatos, los primeros procesos de filtrado, con sus correspondientes cálculos estadísticos, se propusieron eliminar previamente los términos vacíos[12]. Existe para cada idioma un conjunto de palabras vacías, comunes a todos los dominios, fácilmente identificable: artículos, preposiciones, conjunciones, etc., aunque también puede haber verbos, adverbios y adjetivos. Las palabras vacías sólo son descartadas cuando se trate de obtener descriptores simples, ya que pueden formar parte de descriptores compuestos. La eliminación de palabras vacías reduce considerablemente el tamaño de la estructura de indización. A pesar de los beneficios que supone la eliminación de palabras vacías, es posible que esta eliminación reduzca la respuesta a una consulta. En efecto, en el caso de que se pregunte por una frase que contenga una palabra vacía, la recuperación será imposible porque dicha palabra vacía no será un término de indización. Este sistema es el utilizado en los índices permutados tipo KWIC y KWOC.

Las palabras que son demasiado frecuentes en los documentos de una determinada colección no aportan información. De hecho, se considera que una palabra que aparezca en al menos el 80% de los documentos de una determinada colección carece de utilidad en tareas de recuperación de información. Estas palabras se consideran también vacías y normalmente se eliminan durante el proceso de análisis automático de texto para evitar que puedan ser consideradas como potenciales términos de indización[13].

La intervención de la estadística ha mantenido el principio más duradero dentro de las técnicas de indización automática: servirse de la eliminación en el texto de las palabras vacías para indizar las palabras o expresiones significativas[14]. Eliminadas las palabras vacías, era fácil suponer que el paso siguiente se daría con las palabras restantes, concediendo mayor peso a unas que a otras, para lo que se introducirían nuevos criterios estadísticos y probabilísticos. La utilización de los nuevos criterios se aplicó calculando la frecuencia estadística de aparición de las palabras. Partiendo de la hipótesis de Zipf, Luhn estableció un umbral superior y otro inferior para la frecuencia de ocurrencia de un término en un texto, de tal modo que los términos que queden por encima del umbral superior o por debajo del umbral inferior se consideran términos poco relevantes para el sistema. De hecho, las palabras que exceden el umbral superior se consideran muy comunes, y aquellas que quedan por debajo del umbral inferior se consideran raras y, por tanto, ninguna de ellas contribuye de forma significante a proporcionar información. De

[12] Se puede consultar una lista de palabras vacías para el idioma inglés y apreciar la categoría gramatical que suelen tener en Frakes, W., Prieto-Díaz, R, y Fox, C.- DARE: Domain Analysis and Reuse Environment, en *Annals of Software Engineering*, 1998, 5: 125-141.

[13] Gil Leiva, I. La automatización de la indización de los documentos. Gijón: Trea, 1998.

[14] Van Rijsbergen, C.J.- *Information retrieval*. 2ª ed., Londres, Butterworth, 1981: 23.

todos modos, no hay ninguna regla que se pueda establecer a priori sobre cómo establecer estos umbrales de forma general.

Así pues, de los términos significativos encontrados solo pasan la selección aquellos cuya tasa se sitúe en torno a una frecuencia de aparición media. Quedando fuera de las entradas tanto los que están en el umbral superior, como en el inferior. La única discriminación léxica que se puede aplicar en este método primitivo o método de extracción estadística, es la posibilidad de trabajar con las raíces de los términos, para poder tener en cuenta sus formas flexionadas.

Las leyes de Zipf establecen cuándo un término es representativo en un documento:

Si las palabras que aparecen en un texto se disponen en orden decreciente, en función del número de apariciones y se les asigna un número de orden, entonces el producto de la frecuencia relativa de cada palabra por el orden se puede aproximar a una constante. De hecho establece que si se ordenan las palabras de un texto por orden decreciente de ocurrencia, la frecuencia de la palabra que ocupa la posición r-ésima vendrá dada por una distribución frecuencia-rango del tipo:

$f(r) = k / (r+w)^b$, *donde* $r = 1, 2, 3, \ldots$,

con w y b parámetros variables. El valor de k sólo depende del tamaño del texto.

La primera ley de Zipf funciona bien para palabras de alta frecuencia, pero la constante deja de ser tal cuando consideramos palabras de baja frecuencia. Para estas últimas existe otra ley enunciada por Goffman que estableció un procedimiento para calcular esta zona de transición utilizando la segunda ley de Zipf, estimando un número óptimo de frecuencia de aparición *n* alrededor del cuál construye un intervalo de aceptación[15]:

Sea I_1 *el número de palabras con frecuencia absoluta 1 y* I_n *el número de palabras con frecuencia absoluta n, entonces se verifica de modo aproximado que:*

$I_1/I_n = n(n+1)/2$.

Los términos con mayor contenido semántico de un documento se encuentran en la zona de transición entre las palabras de frecuencia muy alta (artículos, conjunciones, preposiciones, etc.) y las de muy baja (las que denotan el estilo de vocabulario del autor). Es muy interesante realizar filtrados sobre los posibles términos representativos de un dominio, ya que a la hora de buscar relaciones entre los términos es necesario que el número de estos sea reducido, debido a que los métodos estadísticos y de redes neuronales que proporcionan estas relaciones trabajan con un conjunto limitado de elementos[16].

La ley de Zifp conoce en la actualidad nuevas formulaciones, como son[17]:

- La aproximación polinomial de la Constante de Zipf.

[15] Chaumier, J. y Dejean, M.- L'indexation documentaire: de l'analyse conceptuel humaine à l'analyse automatique morphosyntaxique, en *Documentaliste – Sciences de l'information*, 1990, 27, nº 6: 277.

[16] Blair, D. C.- *Language and representation in information retrieval*. Amsterdam: Elsevier Sciencie Publishers, 1990.

[17] Kowalski, G.- *Information Retrieval Systems: Theory and Implementation*. Amsterdam: Kluwer Academic Publishers, 1998.

- La aproximación lineal (tamaños de textos fuera del intervalo de estudio).

Otras técnicas han sido pensadas también para discriminar entre los términos que se consideran representativos de un texto y los que se consideran sin importancia. Cada una de estas técnicas presenta una problemática específica, a la par que unas ventajas incuestionables, como sucede con estos procesos algorítmicos:

1. *IDF* o Indización Estadística de Términos por Frecuencias: sistema de filtrado basado en la ley de Zipf y que persigue identificar las palabras que aparecen en la zona media de la función de distribución de frecuencias, las que mejor representan al documento. La técnica IDF establece un sistema de pesos en función de la frecuencia relativa de cada término en cada documento. En el caso de que un término tenga una frecuencia en un documento mayor que la media fijada en el resto de documentos, se tomará como descriptor. En el momento que se tome como descriptor para un documento será considerado como tal en el resto de documentos, es decir, no es necesario que un término aparezca en todos los documentos a filtrar para que sea descriptor. Se aplica primero la ley de Zipf para el cálculo de la zona de transición y después el método IDF para ponderar por documentos.
2. *Método N-grams*: que modifica la ley de Zipf para aprovechar la información que nos proporciona el tratamiento de palabras compuestas, pues la ley de Zipf no estaba pensada para filtrar términos compuestos. El algoritmo *n-grams* es utilizado para reconocer palabras compuestas: trabaja con cadenas de longitud fija cuya frecuencia en un documento es comparada con la frecuencia de esa misma cadena de caracteres en otro documento denominado *background* contra el que se compara. El resultado de este método es dependiente en gran medida del documento de comparación que se elija[18]. El algoritmo *N-grams* filtra de modo parecido a los anteriores, de tal forma que la frecuencia se calcula no sobre cada término o palabra compuesta si no sobre cadenas de caracteres de longitud predeterminada y fija. El número *n*, la longitud de la cadena, toma valores entre 3 y 6. En cada aplicación se toma un valor para poder tener un carácter central en el *n-gram*. La construcción del background necesario para realizar la comparación de frecuencias con los documentos del corpus del dominio no es un paso en absoluto trivial. El filtrado variará en función de la información que componga el background. Para comprobar que el background responde a características generales del lenguaje se han utilizado estudios estadísticos propios sobre cómo aparecen las cadenas en cada idioma.

[18] Meadow, C. T.- *Text Information Retrieval Systems*. San Diego: Academic Press. Meadow, C. T.- *Text Information Retrieval Systems*. San Diego: Academic Press, 1992. Meadow, C. T., Boyce B.R. y Kraft D.H.- *Text Information Retrieval Systems*. 2ª ed. San Diego: Academic Press, 1999.

3. Para el tratamiento de los términos flexionados se suelen utilizar algoritmos de *stemming*. Hay varios algoritmos de este tipo, su uso depende del modo en que se traten los afijos[19]:
   - *Método diccionario* o de creación de un diccionario de raíces. Los algoritmos de eliminación de afijos extraen los sufijos y/o prefijos de términos conservando sólo la raíz.
   - *Método N-Gram*: los *stemmers* utilizan la frecuencia de secuencias de letras en el cuerpo de un texto, así el método *n-grams* que une términos basándose en el número de *n-grams* (grupos de letras) que comparten dichos términos. Divide los términos y comprueba su parecido gramatical con otros términos mediante el *Coeficiente de Dice*[20]*:*
     - Donde A se corresponde con la cantidad de términos que se comparan.
     - Donde B es el número de diagramas que contiene cada término.
     - Donde C es el número de diagramas que coinciden con el término relacionado.
   - *Método de variedad de sucesores*: el algoritmo procesa palabra por palabra, busca los sucesores en toda la colección de palabras, corta por el primer pico y halla la raíz.

3. MÉTODOS DE AGRUPACIÓN Y CLASIFICACIÓN

Aunque los métodos estadísticos fueron los primeros utilizados para automatizar la indización, se siguen manteniendo en las aplicaciones actuales[21]. Es el caso del *clustering* que agrupa, mediante el análisis de las palabras que contienen, aquellos documentos entre los que existe una asociación notable y que son aproximadamente relevantes para las mismas consultas[22]. Se pueden identificar dos utilizaciones distintas a este método[23]: la destinada a crear listas de palabras desde las que identificar los conceptos relevantes, y aquella cuya misión es elaborar extractores automáticos que identifiquen las materias. De esta forma se pueden representar tanto las preguntas de los usuarios como los textos que responden a ellas. Su elaboración exige la presencia de grandes bases de conocimientos terminológicos de un dominio dado. Operan de dos maneras fundamentales:

[19] Frakes, W y Baeza-Yates, R.- *Information Retrieval: Data structures and algorithms*. Upper Saddle River: Prentice-Hall: 1992.
[20] Pao, M.L.- *Concepts of information retrieval.* Englewood: Libraries Unlimited, Inc., 1989.
[21] Así, por ejemplo, Chi-Hong Leung and Wing-Kay Kan.- A statistical learning approach to automatic indexing of controlled index terms, en *Journal of the American Society for Information Science*, 1997, 48, nº1: 55-66 que lo aplican a bases de datos como INSPEC y MEDLINE. También se basan en los modelos estadísticos: Cohen, J.- Highlights: Language and Domain Independent Automatic Indexing Terms Abstracting, en *Journal of the American Society for Information Science*, 1995, 46, 3:162-174.
[22] Sparck Jones, K. - Some thoughts on classification for retrieval, en *Journal of Documentation*, (1970), 26: 89-101.
[23] Jardine, N. y Sibson, R.- *Mathematical Taxonomy*. London and New York: Wiley, 1971.

- *Por valor de discriminación*: partiendo del concepto de recuperación de la información propuesto por Salton[24], dentro de un espacio vectorial estadístico para la indización y la recuperación de información, se trata de conceder el valor más alto a las palabras que causan la mayor diferenciación entre los documentos de una colección que se pretende indizar.
- *Por relevancia de los términos*. También sucesivo a la frecuencia de aparición de las palabras y a las aportaciones teóricas de Salton, la relevancia de las palabras se obtiene a partir de sus valores de utilidad y de precisión, calculados desde algoritmos probabilísticos[25].

La conceptualización de estos métodos es la siguiente:

3.1 Métodos estadísticos de agrupación en Clases

Cuando cualquier sistema ha finalizado el proceso de adquisición de conocimiento, los términos obtenidos deben ser ordenados de acuerdo con las relaciones semánticas que se dan entre los términos lingüísticos y entre los elementos de los lenguajes documentales. Si se busca poder reutilizar información de manera óptima e inteligente es necesario primero clasificarla, de tal modo que se establezcan relaciones entre los componentes que la definen y describen.

La agrupación en clases puede definirse como el proceso de clasificación no supervisada de objetos. Se dispone de un conjunto de vectores {x1, …, xp}, que representan a los objetos y a partir de él se desea obtener el conjunto de clases {(1, …, (n} que los engloban[26]. El problema es que a priori no se sabe cómo se distribuyen los vectores en las clases, ni siquiera cuántas clases habrá. A partir del conjunto de vectores de características dado se trata de conseguir realizar agrupaciones de estos vectores en clases, de acuerdo con las similaridades encontradas.

3.1.1. Métodos de agrupación basados en Grafos Teóricos

Este método define clusters a partir de un grafo derivado de una medida de similitud[27]. La definición del cluster se hace simplemente en función de la representación gráfica. Para cada par de objetos se computa el valor numérico que indica su similitud

[24] Reinterpretado en Salton, G.-*Automatic text processing: the transformation, analysis and retrieval of information by computer*. Boston: Addison – Wesley, 1989.

[25] Salton, G. y McGill, M.J.- *Introduction to Modern Information Retrieval*. New York, McGraw-Hill, Inc., 1983.

[26] Carpenter, A. y Grossberg, S.- ART-2: Stable Self-organization of Pattern Recognition Code for Analog Input Patterns, en *Applied Optics*, 1987, 26: 4919-4930.

[27] Estos métodos han sido aplicados a la clasificación mediante palabras clave por Sparck Jones, K. y Jackson, D.M.- The use of automatically-obtained key-word classifications for information retrieval, en *Information Storage and Retrieval*, 1970, 5: 175-201. Augutson, J. G. y Minker, J.- An analysis of some graph-theoretic cluster techniques, en *Journal of the ACM*, 1970, 17: 97. Vaswani, P.K.T. y Cameron, J.B.- *The National Physical Laboratory Experiments in Statistical Word Associations and their use in Document Indexing and Retrieval*. Publication 42. National Physical Laboratory, Division of Computer Science, 1970.

de modo que se genera un grafo por el que dado un umbral, dos objetos se consideran similares si su coeficiente de similitud supera dicho umbral.

3.1.2. Métodos de agrupación basados en Centros de masas

El centroide es un objeto o término que representa los objetos o términos de un determinado cluster. La similitud de los objetos del cluster respecto a su centroide se mide por una función de comparación. Este tipo de algoritmos utiliza una serie de parámetros determinados de forma empírica:

- El número de clusters deseados.
- Un tamaño mínimo y máximo para cada cluster.
- Un umbral para la función de comparación por debajo del cual un objeto no pertenece al cluster..
- El control de solapamiento entre clusters.
- Una función objetivo que debe ser optimizada.

Estos algoritmos funcionan así:

- Las descripciones de los objetos se procesan en serie.
- El primer objeto se toma como el centroide del primer cluster.
- Cada uno de los siguientes objetos se confrontan con cada uno de los centroides de cada cluster.
- Cada uno de los objetos es asignado a un cluster de acuerdo con la función de comparación que se haya definido.
- Cuando un objeto se asigna a un cluster, el centroide debe ser recalculado.
- Si un objeto no verifica la función de comparación para ningún cluster, el propio objeto se convierte en el centroide de un nuevo cluster.

La clasificación final depende de ciertos parámetros determinados de forma empírica y explicitados a priori.

3.1.2.1. Algoritmo K-vecinos o K-medias: es uno de los algoritmos de agrupación aplicado más comúnmente. Hay muchas variantes del algoritmo, y una de las más eficientes es el algoritmo convergente de las k-medias de Anderberg[28]. Es un algoritmo rápido y eficaz que busca minimizar un índice de rendimiento, basado en la suma de distancias euclídeas de todos los miembros de un grupo a su centroide. Es rápido y eficaz si la distancia que utiliza es adecuada para el problema considerado. Busca minimizar un índice de rendimiento, basado en la suma de distancias euclídeas cuadráticas de todos los miembros de un cluster a su centroide. Exige conocer el número de clusters k en los que se desea clasificar la muestra de vectores de la población. Si el número de clases no se

[28] Anderberg, M. R.- *Cluster Analysis for Applications*. New York: Academic Press, 1973.

conoce por adelantado, se puede dejar que el algoritmo determine el número de clusters utilizando parámetros definidos por el usuario. El modo de funcionamiento del algoritmo consiste en mover cada vector al cluster cuyo centroide esté más cercano al mismo, y actualizar después los centroides de los clusters. Su convergencia depende mucho del número de clases.

3.1.2.2. Algoritmo K-vecinos axial o incremental, variante del algoritmo anterior que permite que el número de grupos sea desconocido a priori, determinando éste de forma adaptativa. Este algoritmo, como su nombre indica, calcula los clusters de forma incremental. Pertenece a la familia de algoritmos de clasificación por centros móviles. Es una variante del algoritmo k-vecinos en su versión adaptativa, y del algoritmo de Forgy, en el caso iterativo. Dado un patrón de entrada, el algoritmo debe actualizar la representación de los clusters y devolver el índice del cluster actual al cual pertenece el patrón, sin necesitar tener presentes los demás patrones. De este modo puede tratarse una sucesión arbitrariamente grande de patrones en tiempo real. Los algoritmos de cluster incremental son muy atractivos para el tratamiento de patrones documentales, dado el gran espacio de almacenamiento que requieren dichos patrones. El algoritmo de cluster euclídeo no converge necesariamente en un conjunto fijo de prototipos: los prototipos pueden variar infinitamente, sin converger en el tiempo. El número de clusters creados tampoco es necesariamente finito, y depende de las funciones utilizadas en el algoritmo.

3.1.2.3. Algoritmo ISODATA: el *Interactive self organizing data analysis techniques* (ISODATA) es un algoritmo interactivo basado en el algoritmo k-vecinos, al que se introduce una consideración heurística que le permite funcionar bien cuando el conocimiento sobre el número de clases no es bueno. Tiene una serie de parámetros que pueden modificar en gran medida la agrupación.

3.1.3. Clasificadores cienciométricos: Método de Chen

El análisis de coocurrencia de palabras estudia el uso de grupos de palabras que aparecen simultáneamente en varios documentos. Las palabras pueden pertenecer a un lenguaje controlado o a texto libre. El método de coocurrencias capaz de evaluar la relación entre dos descriptores se considera, por tanto, un método de clasificación. Su propósito es establecer un peso a la relación que existe entre dos descriptores. Para aplicar tal método se deben haber identificado los descriptores, y posteriormente se debe proceder a realizar el análisis de coocurrencias para todos los documentos del corpus documental. Se calcula un peso para cada término basado en el modelo de espacio vectorial y en una función de semejanza asimétrica[29]. Mediante esta técnica se consiguen relaciones de equivalencia (sinonimia) y asociaciones.

3.2 Métodos con sistemas de aprendizaje

Atendemos finalmente a los algoritmos de clasificación basados en Redes neuronales. Las redes neuronales se utilizan como herramientas o métodos para resolver problemas,

[29] Chen, H.; Lynch, K. J.- Automatic Construction of Networks of Concepts Characterizing Document Databases, en *IEEE Transactions on Systems, Man and Cybernetics*, 1992, 22: 885-902.

en especial los relacionados con el conocimiento humano: reconocimiento de patrones, reconocimiento del lenguaje hablado, reconocimiento de imágenes, procesos de control adaptativo y estudio del comportamiento de ciertos problemas para los que no están muy bien dotados los ordenadores tradicionales. El aprendizaje de una red neuronal está relacionado con los pesos de las conexiones entre sus nodos[30]. Cuando se presenta un patrón a la red, ésta produce una respuesta. Si la respuesta o salida de la red no es la supuesta, habrán de hacerse modificaciones para acercar la respuesta obtenida a la esperada. La señal que se recibe en la capa de neuronas de entrada cuando se le presenta el patrón se mueve a través de los enlaces o conexiones entre capas, hacia las neuronas de la capa de salida. Estos enlaces modulan la señal a su paso con los pesos que los caracterizan. Por lo tanto, si se quiere modificar la señal que llega al final a la capa de salida, habrá que actuar sobre dichos pesos.

Las reglas de aprendizaje especifican cómo se irán modificando los pesos de las conexiones a medida que se entrena la red para mejorar el rendimiento de la misma, es decir, que la salida se vaya aproximando cada vez más a la esperada. Existen dos tipos fundamentales de aprendizaje: Supervisado y No supervisado. Un clasificador es un sistema que va a permitir determinar cuál de las M clases es la más representativa para un patrón de entrada no estático que contiene N elementos. El clasificador neuronal actúa en dos etapas, contabilizándose en la primera el número de elementos que pertenecen a cada clase y en la segunda se selecciona el máximo. La primera etapa se alimenta con los N elementos del patrón de entrada en paralelo, produciéndose aquí la comparación del patrón de entrada con los prototipos de las distintas clases y pasando los resultados intermedios a la siguiente etapa en paralelo. En la segunda etapa se selecciona el máximo. Habrá salida para todas las clases, pero al acabar la clasificación sólo será apreciable la salida para la clase con mayor probabilidad, y el resto serán valores muy bajos o inapreciables. Se pueden utilizar las salidas como realimentación de la primera etapa adaptando los pesos iniciales según un determinado algoritmo de aprendizaje (principio de realimentación negativa)[31].

CONCLUSIÓN

Cualquier desarrollo que involucre un proceso de organización de información debe proporcionar un mecanismo adecuado que permita clasificar dicha información. Más aún en el contexto en que nos movemos, pues por una parte pretendemos recuperar información, para lo cual necesitamos un esquema de representación, y por otra, hacer la representación de esos dominios de forma automática. En cualquiera de los dos casos es necesario establecer algún método de clasificación que permita organizar la información que se maneja.

Ajenos a la intervención del procesamiento semántico, hemos descrito los métodos basados en la expresión textual como objeto, representada mediante elementos matemá-

[30] Hebb, D.- *Organization of Behavior.* New York: Wiley & Sons, 1949.

[31] Frants, V., Shapiro, J., y Voiskunskii, V.- *Automated Information Retrieval: Theory and Methods.* San Diego: Academic Press, 1997. Harter, S.- *Online information retrieval. Concepts, Principles and Techniques.* London: Academic Press, 1986.

ticos, y que constituyen una de las dos metodologías de trabajo en cualquier proyecto de investigación de nuestra especialidad.

No nos propusimos cuestionar aquí la eficacia de simular experimentalmente la recuperación de la información, si no hacer una relación de utilidades imprescindibles para el análisis de contenido automático, cuyo origen está en el método estadístico.