Generación automática de reglas de categorización de texto en un método híbrido basado en aprendizaje

Automatic generation of text categorization rules in a hybrid method based on machine learning

Sara Lana-Serrano
Universidad Politécnica de Madrid
Crta Valencia km 7 – E-28031 Madrid
slana@diatel.upm.es

Julio Villena-Román
Universidad Carlos III de Madrid
Av. de la Universidad 30 – E-28911 Leganés
jvillena@it.uc3m.es

Sonia Collada-Pérez
DAEDALUS, S.A.
Av. de la Albufera 321 – E-28031 Madrid
scollada@daedalus.es

José Carlos González-Cristóbal
Universidad Politécnica de Madrid
Ciudad Universitaria s/n – E-28040 Madrid
josecarlos.gonzalez@upm.es

Resumen: En este artículo se evalúan diferentes técnicas para la generación automática de reglas que se emplean en un método híbrido de categorización automática de texto. Este método combina un algoritmo de aprendizaje computacional con diferentes sistemas basados en reglas en cascada empleados para el filtrado y reordenación de los resultados proporcionados por dicho modelo base. Aquí se describe una implementación realizada mediante el algoritmo kNN y un lenguaje básico de reglas basado en listas de términos que aparecen en el texto a clasificar. Para la evaluación se utiliza el corpus de noticias Reuters-21578. Los resultados demuestran que los métodos de generación de reglas propuestos producen resultados muy próximos a los obtenidos con la aplicación de reglas generadas manualmente y que el sistema híbrido propuesto obtiene una precisión y cobertura comparables a la de los mejores métodos del estado del arte.
Palabras clave: Clasificación de texto, aprendizaje computacional, sistema basado en reglas, kNN, Reuters-21578, información mutua, generación automática de reglas, evaluación.

Abstract: This paper discusses several techniques for the automatic generation of rules to be used in a novel hybrid method for text categorization. This approach combines a machine learning algorithm along with a different rule-based expert systems in cascade used to filter and re-rank the output of the base model provided by the previous classifier. This paper describes an implementation based on kNN algorithm and a basic rule language that expresses lists of terms appearing in the text. The popular Reuters-21578 news corpus is used for testing. Results show that the proposed methods for automatic rule generation achieve precision values that are very similar to the ones achieved by manually defined rule sets, and that this hybrid approach achieves a precision that is comparable to other top state-of-the-art methods.
Keywords: Text categorization, machine learning, rule-based system, kNN, Reuters-21578, mutual information, automatic rule generation, evaluation.

1 Introducción

En este artículo se proponen diferentes técnicas de generación automáticas de reglas para un método híbrido de categorización automática de texto. El método híbrido utilizado consiste en un algoritmo de aprendizaje automático combinado con un sistema basado en reglas serializados. El objetivo es obtener un modelo en el que la generación de reglas se efectúe de forma automática, obteniendo unos parámetros de precisión y cobertura (*recall*) equivalentes a los proporcionados en dicho modelo utilizando un conjunto de reglas definidos manualmente.

En los siguientes apartados se describe en detalle los fundamentos y arquitectura lógica

* Esta investigación ha sido parcialmente financiada por los proyectos de I+D BUSCAMEDIA (CEN-20091026), MULTIMEDICA (TIN2010-20644-C03-01) y BRAVO (TIN2007-67407-C03-01).

del método híbrido empleado, las técnicas de generación automática de reglas que se proponen, y las diferentes pruebas que se han realizado para verificar su validez.

2 *Categorización automática de textos*

La categorización (o clasificación) automática de textos consiste en asignar automáticamente una o varias categorías (o clases) predefinidas a un determinado texto en lenguaje natural, según su similitud con respecto a otros textos etiquetados previamente, empleados como referencia.

Típicamente existen dos enfoques para la clasificación de textos (Sebastiani 2002). Por un lado está el enfoque *basado en conocimiento*, que consiste en la creación de un sistema experto con reglas de clasificación definidas de forma manual, típicamente una por categoría. Se asume comúnmente que se pueden producir reglas tan precisas como sea necesario. Estas reglas suelen ser expresiones lógicas que combinan términos (palabras y multipalabras) del texto mediante la aplicación de los operadores booleanos AND, OR y NOT.

Por otro lado, está el enfoque de *aprendizaje automático*. En este caso, se proporciona al sistema un conjunto de textos etiquetados para cada categoría, que se usa como conjunto de entrenamiento para construir un clasificador. La ventaja es que sólo se necesita un mínimo conocimiento del dominio para asignar una categoría a cada texto existente en el conjunto de entrenamiento, lo que implica una carga de trabajo mucho menor que la escritura de las reglas. Se han propuesto numerosos algoritmos y técnicas de aprendizaje supervisado para construir los clasificadores.

Aunque se ha demostrado que este enfoque puede generar clasificadores igual de buenos que los sistemas basados en reglas, pero con un menor esfuerzo, su inconveniente fundamental es que, en la mayoría de los algoritmos de aprendizaje empleados, el modelo no es comprensible por el ser humano, con lo que es difícil diagnosticar la razón de los falsos positivos/negativos para ajustar el sistema. Así, la única manera de mejorar el clasificador es invertir más esfuerzo en la construcción del conjunto de entrenamiento y alternativas de selección de términos.

3 *Enfoque Híbrido*

Como resultado de nuestra investigación en diferentes estrategias estadísticas y/o semánticas de expansión automática de las consultas para mejorar la precisión y la cobertura en diferentes tareas de recuperación de información y clasificación automática (Villena-Román et al. 2009) llegamos a la conclusión de que la utilización de técnicas de expansión semántica del corpus (como sinónimos o términos que coaparecen con un cierto valor de confianza), aun cuando generaban mejoras en la cobertura de nuestros experimentos, por lo general siempre empeoraban los valores de precisión.

Por ello decidimos adoptar una estrategia diferente y actuar en la etapa de post-procesamiento de la salida del clasificador automático, proponiendo el método híbrido descrito en (Villena-Román et al. 2011) y mostrado en la Figura 1.

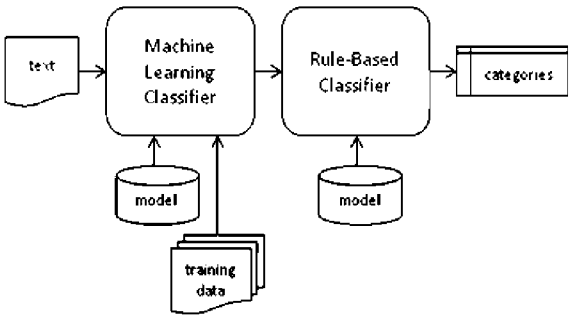

Figura 1: Arquitectura lógica del sistema

Existen trabajos que proponen métodos híbridos para diferentes tareas de clasificación, pero casi todos ellos se basan en la combinación de dos o más clasificadores de aprendizaje automático como (Kim y Myoung 2003). Es novedosa la aplicación combinada de un algoritmo de aprendizaje utilizado para la construcción de un modelo base, con un sistema basado en reglas para el filtrado de resultados.

3.1 Clasificador basado en Aprendizaje Automático

El primer paso del modelo híbrido tiene como objeto entrenar un modelo a partir de un corpus de entrenamiento. Se puede utilizar cualquier algoritmo de aprendizaje que pueda dar respuesta a los requisitos de cada escenario en concreto.

Sin embargo sería deseable que fuera capaz de proporcionar una clasificación multietiqueta manejando un elevado número de categorías no balanceadas (algunas categorías con muy pocos textos y otras con demasiados). Además, desde una perspectiva pragmática (comercial), podrían resultar interesantes otros requisitos no

funcionales, como clasificación en tiempo real o la edición del modelo *en caliente*.

Para nuestra implementación, elegimos el algoritmo de los k vecinos más cercanos (*k-Nearest Neighbour* o kNN) basado en la distancia euclídea, por su simplicidad y buen rendimiento demostrado en otros trabajos (Joachims, 1998). Nuestra implementación se basa en el motor de código abierto de recuperación de información Apache Lucene.

Independientemente del algoritmo de aprendizaje, todos los clasificadores se basan en el hecho de que cuantas más veces aparezca un término en el texto, más relevante es ese término. Según el Modelo de Espacio de Vectores (Salton, et al., 1975), cada texto del corpus se representa como un vector multidimensional $(w_{i1}\, w_{i2}\, \ldots w_{iN})$, donde cada valor w_{ij} del vector representa el grado en el que el termino está presente (o ausente) en el texto.

En nuestro sistema, utilizamos TF*IDF para construir los vectores, y reducimos la dimensionalidad por selección de características seleccionando, mediante el método de Bag-Of-Words, los N términos con más peso (en concreto se utilizan 200 términos).

3.2 Clasificador basado en Reglas

Tras el paso de clasificación automática, se aplica en cascada un sistema experto encargado de post-procesar la salida del clasificador. Este sistema experto utiliza reglas basadas en expresiones lógicas definidas sobre términos en lenguaje natural. En general, cada categoría puede tener ninguna, una o varias reglas asociadas.

Cada regla se evalúa sobre el texto de entrada q para aceptar (validar), rechazar (invalidar) o proponer (incluir) dicha categoría en la lista de resultados, según si el texto satisface o no las condiciones expresadas en la regla. El rechazo de una categoría elimina los falsos positivos devueltos por el clasificador basado en aprendizaje automático, por lo que mejora la precisión. La inclusión de una categoría adicional resuelve los falsos negativos, con lo que mejora la cobertura. Las reglas también son utilizadas para reordenar la lista de resultados. La aplicación de una regla puede incrementar (reforzar) el grado de pertenencia de un documento a una categoría, por ejemplo, dependiendo del número de términos que satisfagan la expresión lógica, contribuyendo a mejorar la precisión global.

En nuestra implementación, diseñamos un lenguaje de reglas sencillo. Para cada i-ésima categoría, la regla tiene cuatro componentes:

- Términos *positivos* $P_i = \{p_{i1},\ p_{i2} \ldots p_{ip}\}$: al menos uno de estos p términos debe aparecer obligatoriamente en el texto, es decir:

  if $(p_{i1}\ OR\ p_{i2}\ OR \ldots OR\ p_{ip})$ **then**
  (categoría aceptada) (1)
  else
  (categoría rechazada)

- Términos *negativos* $N_i = \{n_{i1},\ n_{i2} \ldots n_{in}\}$: ninguno de estos n términos debe aparecer en el texto, es decir:

  if $(n_{i1}\ OR\ n_{i2}\ OR \ldots OR\ n_{in})$ **then**
  (categoría rechazada) (2)
  else
  (categoría aceptada)

- Términos *relevantes* $R_i = \{r_{i1},\ r_{i2} \ldots r_{ir}\}$: se usan para incrementar la relevancia de la categoría, como se describe más adelante.

- Términos *irrelevantes* $I_i = \{i_{i1},\ i_{i2} \ldots i_{ir}\}$: similar al caso anterior, pero reduciéndola.

El factor de *boosting* de una determinada categoría se muestra en la Ecuación 3. Los términos negativos se utilizan para rechazar la categoría, así que su relevancia final es cero.

El algoritmo de aprendizaje proporciona una lista de categorías $(c_i \in C)$, ordenadas según su *categorization status value* (CSV)[1] con respecto al texto de entrada q:

$$D_q = \begin{pmatrix} csv_{q,1} \\ \cdots \\ csv_{q,K} \end{pmatrix} \quad (3)$$

El resultado de este segundo bloque, y por tanto, del sistema global, es una lista de categorías reordenada con su nuevo CSV:

$$D'_q = D_q * B_q = \begin{pmatrix} csv'_{q,1} \\ \cdots \\ csv'_{q,K} \end{pmatrix} \quad (4)$$

$$B'_{q,i} = \begin{cases} 0 & si \exists t_i \in N \\ 1 + cuenta(t_i \in P) + cuenta(t_i \in R) & \text{si no} \\ \quad - cuenta(t_i \in I) & \end{cases}$$

$$B_{q,i} = \begin{cases} B'_{q,i} & if\ B'_{q,i} \geq 0 \\ 1/B'_{q,i} & \text{si no} \end{cases} \quad (5)$$

Los términos en las reglas pueden ser palabras individuales (por ejemplo *gasolina*) o uni-

[1] Cuanto mayor es el CSV de una categoría para un documento, mayor es el grado de pertenencia de ese documento a dicha categoría (Sebastiani 2002).

dades multipalabra (como *producto interior bruto*). En este caso, la condición booleana es *true* cuando todas las palabras están en el texto:

$$t_i \equiv t_{i1}\ AND\ t_{i2}\ AND\ ...\ AND\ t_{ik} \quad (6)$$

Por último, se definen dos reglas adicionales para el caso en que no se haya definido ninguna regla para una categoría dada. La regla ACCEPT valida la categoría, sin importar los términos del texto ($B_{q,i} = 1$), y es la regla por defecto del sistema. La regla REJECT invalida la categoría ($B_{q,i} = 0$).

4 *Generación automática de reglas*

El principal problema que presenta el modelo híbrido así planteado es que el proceso de generación de reglas en corpus con gran volumen de documentos y/o clases no es evidente y es necesario invertir gran cantidad de esfuerzo para su definición, por lo que se hace necesario disponer de un sistema de generación o recomendación de reglas que sirva de soporte para la definición de las mismas.

En concreto, nosotros nos hemos centrado únicamente en la definición de reglas que puedan contribuir a incrementar la relevancia (términos positivos y términos relevantes) de un término *t* para una clase *c*.

Existen trabajos interesantes basados en diferentes métodos, sobre todo algoritmos genéticos (Hirsch et al., 2007), pero nuestro trabajo consiste en el desarrollo de diferentes modelos de recomendación de reglas basadas en el concepto de Información Mutua (*Mutual Information* o MI). MI proporciona una medida de cuánto contribuye la presencia de un término *t* de un documento en el grado de pertenencia de dicho documento en una clase *c*.

$$\mathrm{MI(t,c)}=\frac{N_{11}}{N}\log_2\frac{N\,N_{11}}{N_{1.}\,N_{.1}}+\frac{N_{01}}{N}\log_2\frac{N\,N_{01}}{N_{0.}\,N_{.1}} \quad (7)$$
$$+\frac{N_{10}}{N}\log_2\frac{N\,N_{10}}{N_{1.}\,N_{.0}}+\frac{N_{00}}{N}\log_2\frac{N\,N_{00}}{N_{0.}\,N_{.0}}$$

N: documentos de la colección.
N_{11}: documentos que contienen el término *t* y pertenecen a la clase *c*.
N_{01}: documentos que no contienen el término *t* y pertenecen a la clase *c*.
N_{10}: documentos que contienen el término *t* y no pertenecen a la clase *c*.
N_{00}: documentos que no contienen el término *t* y no pertenecen a la clase *c*.
$N_{1.} = N_{11} + N_{10}$; documentos que contienen el término *t*.
$N_{0.} = N_{01} + N_{00}$; documentos que no contienen el término *t*.
$N_{.1} = N_{11} + N_{01}$; documentos que pertenecen a la clase *c*.
$N_{.0} = N_{10} + N_{00}$; documentos que no pertenecen a la clase *c*.

A partir de la definición de MI se han definido los siguientes estadísticos a evaluar:

- *MI normalizado (t,c)*: MI normalizado con respecto al valor máximo que puede tomar MI para la clase c.

$$\mathrm{MI\_N(t,c)}=\frac{MI(t,c)}{MI(c)_{|N11=N_{1.};N01=N10=0;N00=N-N11}} \quad (8)$$

- *MI ponderado (t,c):* MI ponderado por el número de ocurrencias del término en documentos de la clase respecto al número de ocurrencias en documentos que no pertenecen a la clase.

$$\mathrm{MIP(t,c)}=\mathrm{MI(t,c)}\ \frac{TF_1/N_{11}}{(TF_0+TF_1)/N_{1.}} \quad (9)$$

TF_1: ocurrencias del término *t* en los documentos de la clase *c*.
TF_0: ocurrencias del término *t* en los documentos que no pertenecen a la clase *c*.

- *MI ponderado normalizado (t,c):* normalización del anterior.

$$\mathrm{MIP\_N(t,c)}=\mathrm{MI\_N(t,c)}\ \frac{TF_1/N_{11}}{(TF_0+TF_1)/N_{1.}} \quad (10)$$

Los modelos desarrollados recomiendan, para cada una de las clases del corpus de entrenamiento, tanto términos *positivos* como términos *relevantes* atendiendo al valor que toma el estadístico evaluado.

5 *Evaluación*

Para la comparación de los diferentes modelos desarrollados se ha utilizado el corpus de noticias de prensa en inglés Reuters-21578 (en su versión R90), utilizando como métrica para la evaluación de los resultados el valor del punto de equilibrio micro-promediado (*Microaveraged Breakeven Point*, el punto donde la precisión y la cobertura se igualan).

La Tabla 1 muestra las características que diferencian a cada uno de los modelos de aprendizaje utilizados. Estos modelos únicamente se diferencian en el tipo de procesamiento aplicado al corpus para la obtención de los términos del modelo de espacio de vectores.

Id	Descripción
B	Modelo base: eliminación de palabras de parada y conversión a minúsculas.
S	Modelo base + lematización.
T	Modelo base + lematización + generación de multitérminos.

Tabla 1: Modelos de aprendizaje automático

Por otra parte, los experimentos nombrados como I y II son experimentos base realizados con el objeto de tomar como modelo de aprendizaje automático aquel que mejores resultados ofrece. Los experimentos orientados a evaluar el efecto que la generación de reglas automática produce en los resultados, únicamente se han evaluado sobre el modelo de aprendizaje T por ser el que mejor resultados proporciona, su principal diferencia con respecto a los otros modelos es la utilización de bigramas de términos para la generación del modelo de espacio de vectores.

Los experimentos I emplean únicamente el modelo kNN sin ninguna regla (es decir, todas las categorías usan la regla por defecto ACCEPT). Los experimentos II utilizan reglas generadas manualmente para todas las categorías. La Tabla 2 muestra el conjunto de reglas escritas de forma manual para las 10 categorías principales (con más documentos de entrenamiento y test).

Categoría	Términos positivos (P) y relevantes(R)
acq (*fusiones*)	R:mergers|merge|acquisition| acquisition|share|shares| company|companies
corn (*maíz*)	R: corn|maize
crude (*crudo*)	R: crude|oil|barrel|barrels| petroleum
earn (*beneficios*)	R: earnings|dividend|dividends| benefit|benefits|loss|losses|growth| income|incomes|net|company| companies|deficit|deficits|debt| debts|reduce|increase
grain (*cereal*)	R: grain|grains|crop
interest (*intereses*)	R: interest|interests|rate|rates|prime| discount
money-fx (*tipos de cambio*)	R: money_exchange|exchange| exchanges|change|changes| money|value|monetary|currency| currencies|money_market
ship (*transporte marítimo*)	R: shipping|shippings|ship| waterway
trade (*comercio*)	R: trade|commerce|deficit|import| imports|export|exports|trade_deficit
wheat (*trigo*)	P: wheat

Tabla 2: Reglas manuales

Los experimentos III, IV, V y VI usan reglas generadas automáticamente con los estadísticos MI, MIP, MI_N y M, respectivamente.

Finalmente, la Tabla 3 muestra los resultados obtenidos para los diferentes modelos de generación de reglas aquí descritos. Como se observa, en todos ellos la aplicación del módulo de clasificación basada en reglas mejora sensiblemente los resultados.

Id.	Descripción	BEP
B_I	Sólo kNN (sin reglas)	0,745
B_II	Reglas manuales.	0.778
S_I	Sólo kNN (sin reglas)	0.777
S_II	Reglas manuales.	0.854
T_I	Sólo kNN (sin reglas)	0,785
T_II	Reglas manuales.	**0,863**
T_III	Reglas automáticas aplicando estadístico MI.	0,687
T_IV	Reglas automáticas aplicando estadístico MIP.	0.696
T_V	Reglas automáticas aplicando estadístico MI_N.	0,826
T_VI	Reglas automáticas aplicando estadístico MIP_N.	**0,831**

Tabla 3: Resultados obtenidos

De los resultados obtenidos se puede observar que la utilización de reglas generadas automáticamente utilizando como estadístico la información mutua normalizada y ponderada, (MIP_N) aun cuando empeora en un 3,7% los resultados obtenidos mediante la generación manual de reglas, puede utilizarse como alternativa o complemento de la anterior.

En la Tabla 4 se muestran las reglas generadas automáticamente, para las 10 categorías principales, utilizando el estadístico MIP_N. Se puede observar que, en general, los términos recomendados contienen a todos los términos definidos manualmente, sin embargo éstos tienden a calificarse como términos positivos frente a relevantes. Se ha observado que este comportamiento únicamente se produce para aquellas categorías en las que el número de documentos de entrenamiento es alto. Por el contrario, en aquellas categorías en las que el número de documento es pequeño el conjunto de términos recomendados suele estar asignado a términos relevantes. Este comportamiento viene determinado por el modelo estadístico utilizado, y refleja que a mejor número de documentos de test, menor es la confianza en la regla inferida dado que el número de evidencias es muy pequeño.

Categoría	Términos positivos (P) y relevantes (R)
acq	P: acquir|acquisit|compani| share|stake R: merger|corp
corn	P: corn|tonn|maiz|usda R: agricultur|grain|tonn_corn
crude	P: crude|oil|barrel|crude_oil|barrel_dai| bpd|petroleum|opec|energi|mln_barrel| dlr_barrel|oil_compani|oil_price R: explor
earn	P: earn|ct|shr|net|qtr|rev|ct_net|loss| 4th|4th_qtr|mln_note|div|profit| dividend|shr_loss|net_loss R: note|avg_shr|year_shr|qtly|ct_prior
grain	P: grain|wheat|agricultur|ton R: usda|corn
interest	P: interest|rate|bank|monei_market| interest_rate|pct|prime_rate|lend_rate| prime R: bank_england|lend|market|point
money-fx	P: monei|bank|dollar|monei_market| currenc|central_bank|market|rate| central|exchang_rate R: bank_england|uk_monei|england| treasuri|dealer|yen|exchang
ship	P: ship|vessel|port|strike R: tanker|cargo
trade	P: trade|tariff|import|deficit|surplu| trade_surplu|trade_deficit|countri| japan|unit_state R: japanes|semiconductor|retali| billion_dlr|gatt|billion
wheat	P: wheat|agricultur|tonn R: agricultur_depart

Tabla 4: Reglas automáticas

6 *Conclusiones y Trabajos Futuros*

En este artículo se han propuesto diferentes técnicas de generación de reglas para categorización automática a partir de indicadores estadísticos basados en el concepto de Información Mutua (MI). Además, se han presentado los resultados de la evaluación realizada con el objeto de determinar si la calidad de los resultados obtenidos a partir de un modelo automático de generación de reglas, es comparable con los obtenidos mediante un modelo manual de definición de reglas.

Las reglas generadas se utilizan en un método híbrido, propuesto en otros trabajos, que combina un algoritmo de aprendizaje que proporciona un modelo de clasificación base relativamente poco costoso de entrenar, con un sistema experto basado en reglas, que post-procesa los resultados del primer clasificador, mejorando su precisión y cobertura filtrando los falsos positivos y resolviendo los falsos negativos. Para las pruebas se ha descrito una implementación basada en kNN y un lenguaje básico de reglas que permite expresar listas de términos positivos, negativos, relevantes e irrelevantes.

La conclusión principal que se puede obtener de la evaluación usando el corpus de noticias Reuters-21578 es que la utilización de reglas generadas automáticamente, utilizando el estadístico de Información Mutua ponderado normalizado (MIP_N), cuya finalidad es reforzar términos positivos y relevantes, proporciona resultados comparables con la generación de reglas manuales que refuerzan dichos términos, por lo que puede combinarse con el modelo híbrido ya sea como único generador de reglas o como sistema de recomendación.

Actualmente estamos trabajando en el análisis de modelos automáticos que permitan generar y/o recomendar reglas automáticas que permitan definir los términos negativos y no relevantes. Además, estudiamos la aplicación de este sistema híbrido a otros escenarios como la moderación automática de foros, el análisis de opinión y la creación de marcadores sociales y bibliotecas digitales (Heymann 2010).

Bibliografía

Heymann, P., Paepcke, A., Garcia-Molina, H. 2010. Tagging human knowledge. In 3rd ACM International Conference on Web Search and Data Mining (WSDM), 51–60.

Hirsch, L., Hirsch, R., y Saeedi, M. 2007. Evolving Lucene search queries for text classification. In *Proceedings of 9th annual conference on Genetic and Evolutionary Computation* (GECCO '07), pp 1604–1611.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, Germany, 1998), 137–142.

Kim, I.C., y Myoung, S. 2003. Text Categorization Using Hybrid Multiple Model Schemes. *Advances in Intelligent Data Analysis V.* Lecture Notes in Computer Science, 2003, Volume 2811/2003, 88–99.

Salton, G., Wong, A., y Yang, C.S. 1975. A vector space model for automatic indexing. *Communications of the ACM*, volume 18 num 11, pp 613–620.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), pp 1–47.

Villena-Román, J., Lana-Serrano, S., y González-Cristóbal, J.C. 2009. MIRACLE-GSI at ImageCLEFphoto 2008: Different Strategies for Automatic Topic Expansion. Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Carol Peters et al (Eds.). *Lecture Notes in Computer Science*, Vol. 5706.

Villena-Román, J., Lana-Serrano, S., Collada-Pérez, S., y González-Cristóbal, J.C. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *In Proceedings of FLAIRS-24, 24th International Florida Artificial Intelli-gence Research Society Conference*, Palm Beach, Florida, USA, May.