La documentación en el proceso de evaluación de Sistemas de Clasificación Automática

Rodrigo SÁNCHEZ JIMÉNEZ

Departamento de Biblioteconomía y Documentación
Universidad Complutense de Madrid

Recibido: 31-10-2006
Aceptado: 12-01-2007

RESUMEN
En este trabajo presentamos una perspectiva general de la evaluación en clasificación automática de documentos. Analizamos tanto los métodos de evaluación como fundamentalmente las colecciones de pruebas sobre los que estos se emplean, haciendo especial énfasis en la utilización de lenguajes documentales en las mismas. Hemos detectado un conjunto de imperfecciones que pueden desvirtuar los resultados de la evaluación y proponemos alternativas para solventarlas.

Palabras-clave: clasificación automática, métodos de evaluación, colecciones de pruebas, sistemas de clasificación.

Documentation in the process of evaluating Automatic Classification Systems

ABSTRACT
This paper offers a general view of evaluation in automatic document classification. We analyse both evaluation methods and test collections in which they are used, focusing in the later aspect, and especially in the use of documentary languages inside these collections. We have detected a set of imperfections that could undermine trust in evaluation process results, and propose some ways of solving them.

Key Words: Automatic Classification, Evaluation Methods, Test Collections, Classification Schemes.

1. INTRODUCCIÓN

Se puede definir la Clasificación Automática de Documentos, también denominada *categorización de textos* o *topic spotting*, como la tarea de asignar automáticamente un conjunto de documentos a una o más categorías preexistentes a través de un conjunto de documentos clasificados por expertos sobre los que el sistema lleva a cabo un proceso de aprendizaje supervisado[1].

[1] Véanse los trabajos de Yang (1999), Sebastiani (2003), Dumais (1998) o Lewis (1994) para obtener otras definiciones clásicas. Esta definición refleja el estado de la cuestión actual y no incluye los precedentes basados en ingeniería del conocimiento, ya obsoletos.

Los objetivos de la Clasificación Automática de Documentos se identifican claramente con los de la Recuperación de Información *ad-hoc*[2], aunque en Clasificación Automática no se trabaja, evidentemente, sobre la base de las consultas. Podemos caracterizar la *Clasificación Automática de Documentos* como un campo de estudio dentro del marco general de la *Recuperación de Información*, aunque existe un importante aporte disciplinar de la Inteligencia Artificial. La investigación en Clasificación Automática de Documentos tiene un carácter híbrido que viene dado por la utilización de principios y metodologías propias de la Inteligencia Artificial para conseguir objetivos propios de la Recuperación de Información.

Los trabajos científicos del área no suelen incluir reflexiones sobre los presupuestos en los que se basa la Recuperación de Información ni la Inteligencia Artificial, aunque hacen uso de técnicas y modelos que ambas disciplinas proporcionan. Por otra parte las tareas de clasificación automática son vistas por muchos de los investigadores que provienen de la inteligencia artificial como un magnífico banco de pruebas para sus investigaciones. Es por tanto un área de investigación hasta cierto punto utilitarista con respecto de las disciplinas de que se nutre.

Se trata entonces de un campo de estudio multidisciplinar, en cuyo desarrollo deberían estar involucradas tanto la Lingüística Documental como la Documentación en sentido más amplio. Dado que durante el proceso de clasificación automática y de evaluación de los clasificadores se utilizan lenguajes documentales y se lleva a cabo una representación de los documentos en torno a criterios similares a los del análisis documental, parece lógico sugerir la necesidad de llevar a cabo estos procesos en torno a criterios documentales.

Para comprobar esto último en primer lugar estudiaremos los procedimientos llevados a cabo para la clasificación automática de documentos, con el objeto de compararlos con respecto de su referente documental, la clasificación manual.

2. CLASIFICACIÓN AUTOMÁTICA Y CLASIFICACIÓN MANUAL

Desde un punto de vista general, el proceso de clasificación se puede definir como:

> "... el acto de organizar el universo del conocimiento en algún orden sistemático. Ha sido considerada la actividad más fundamental de la mente humana. El acto de clasificar consiste en el dicotómico proceso de distinguir cosas u objetos que poseen cierta característica de aquellos que no la tienen y agrupar en una clase cosas u objetos que tienen la propiedad o característica en común[3]."

---

[2] Ambas son tareas que pretenden mejorar el acceso del usuario a la información de forma interactiva, por contraste con las tareas de filtrado. Véase: BAEZA YATES, Ricardo y RIBEIRO-NETO, Berthier, *Modern Information Retrieval*. Addison Wesley & ACM Press, Nueva York, 1999. Pág. 13.

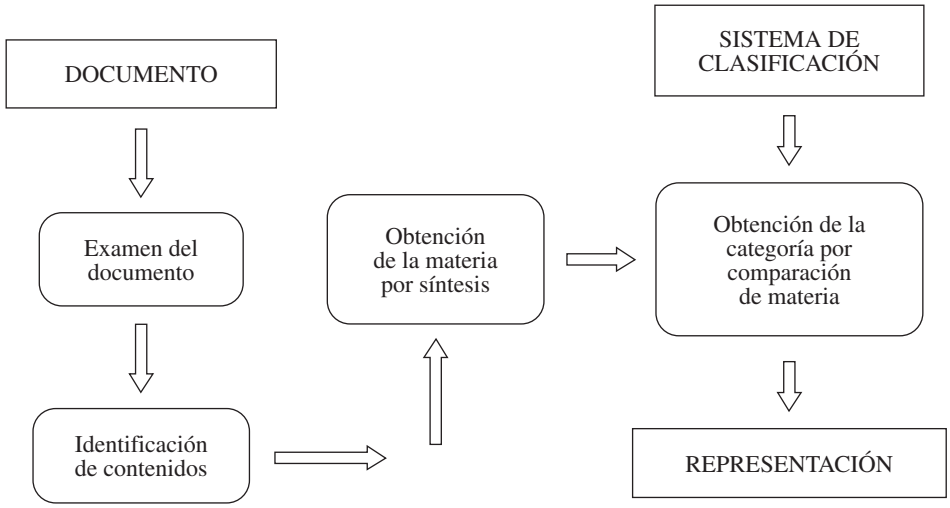
[3] CHAN, M.L. *Cataloging and classification: an introduction*. New York, McGraw-Hill, 1981, p. 209, citado en GIL URDICIAIN, Blanca, *Manual de Lenguajes Documentales*. NOESIS, Madrid, 1996.

Desde el punto de vista más concreto del análisis documental y siguiendo los tres pasos fundamentales indicados por Blanca Gil Urdiciain (1996:38), tendríamos que clasificar básicamente consiste en:

1. Examinar un documento para obtener una idea clara de sus contenidos.
2. Sintetizar dichos contenidos en un tema principal.
3. Contrastar este tema principal o materia con un lenguaje clasificatorio con el objeto de determinar qué categoría se aproxima más a él y representar el documento mediante la notación propia de la clasificación de forma que el almacenamiento ordenado y la recuperación del documento fueran posibles.

FIGURA 1

ESQUEMA DE CLASIFICACIÓN MANUAL



En la figura 1 hemos descompuesto la tarea de clasificar en varias tareas más pequeñas con el objeto de permitir la comparación con la forma de operar de un clasificador automático. Varias de estas tareas no pueden ser afrontadas de forma directa por un clasificador artificial, sino que deben ser llevadas a cabo a partir de otros caminos que en principio deberían producir resultados similares. Analicemos paso por paso el recorrido que un clasificador automático llevaría a cabo para clasificar un documento cualquiera conforme a una clasificación preestablecida.

El primero de los problemas importantes es el que media entre el examen del documento y la identificación de los contenidos. Podemos denominar esta fase como la de representación de los temas del documento. Durante esta fase se genera una

"vista lógica"[4] del documento mediante la elección de una parte de los términos de que está compuesto y su ponderación para intentar expresar en lo posible los contenidos del documento. Esto suele llevarse a cabo mediante la utilización de métodos de indización automática, como las variaciones de la fórmula TF-IDF propuestas por Salton (1988).

Parece obvio que el proceso no tiene nada que ver con el que realiza un ser humano, ya que la lectura es secuencial y la ponderación de la importancia de los términos es estadística. Dejando a un lado consideraciones de eficacia vamos a suponer que el sistema lo lleva a cabo con cierta corrección. En este momento el sistema tendría una representación de los documentos más o menos acertada que reflejaría sus contenidos basándose en la relación que existe entre el léxico y los conceptos asociados a este. El sistema no comprende lo que estos contenidos significan, pero los representa de una forma que simula este entendimiento. Los términos de indización ya son utilizables para responder a necesidades informativas.

El siguiente paso consiste en conectar esta representación que lleva implícita los contenidos del documento con una materia que los sintetice. Dado que el sistema no ha llegado a comprender los contenidos del documento parece lógico que no pueda llevar a cabo la complicada labor de pasar del análisis a la síntesis tal y como lo haría un ser humano. De hecho el clasificador automático nunca dará este paso, no puede averiguar qué materia sintetizaría mejor el documento en conjunto. Simplemente carece del utillaje necesario para hacerlo, a no ser que encuentre en la suma de los términos de indización una representación de dicha materia, lo que en cualquier caso nosotros no hacemos al clasificar algo. Incluso en este último caso un clasificador tampoco podría decidir cuál es la materia en concreto, ya que desconoce las materias y su significado, y podría a lo sumo representarla en un nivel sub-simbólico.

El último escollo de importancia es el que media entre la materia de un documento y una categoría de la clasificación. Para un ser humano esta no es una tarea especialmente difícil, ya que los sistemas de clasificación, sean jerárquicos, mixtos o facetados, se orientan precisamente a facilitar la localización de materias determinadas. Podríamos describir este proceso como un periplo de categoría a subcategoría en el caso más frecuente de las clasificaciones jerárquicas. Un agente humano comprobaría la pertenencia de la materia sobre la que versa el documento a cada una de las clases principales de la clasificación hasta encontrar la correcta y luego iría descendiendo por sus subclases hasta encontrar la que mejor se adaptara a la materia del documento. En resumen, compara la materia del documento con la materia del sistema de clasificación hasta encontrar una correspondencia. Un clasificador artificial no podría llevar a cabo esta tarea de forma directa, ya que de nuevo desconoce las materias y su significado.

Como podemos observar el trabajo de un clasificador automático no se parece mucho al de un clasificador manual. La parte inteligente de un clasificador no inten-

---

[4] Representación del documento en términos de un modelo de recuperación de información concreto.

ta emular el comportamiento de un ser humano, sino aprovecharse de él para llegar a las mismas conclusiones mediante caminos diferentes. La forma en que se hace esto se basa en la utilización de conjuntos de documentos de entrenamiento. Estos documentos de entrenamiento han sido preclasificados por expertos, y el sistema los representa en sus propios términos mediante técnicas estadísticas de indización para poder utilizarlos.

La forma en que el sistema aprovecha las decisiones previas de los clasificadores humanos difiere de clasificador en clasificador, aunque todos llevan a cabo un proceso general inductivo que establece una serie de formas de actuar que repetidas sobre documentos nuevos pueden llevarles a clasificar con acierto un documento. Dicho de otra forma, existe una fase previa a la clasificación durante la cuál el clasificador lleva a cabo un proceso de aprendizaje del que se infieren una serie de condiciones de pertenencia que un documento d_j cualquiera debería satisfacer para poder ser considerado como clasificable bajo c_i y que son generalizables y extensibles a nuevos documentos previamente desconocidos.

Clasificar se reduce entonces a evaluar una serie de condiciones para un documento y emitir una decisión afirmativa o negativa acerca de la idoneidad de asignar un documento d_j a una categoría c_i. El clasificador permanece ajeno al significado de los documentos o al significado de las clases durante todo el proceso. Suple estas carencias mediante la utilización de las decisiones de expertos que sí poseían estos conocimientos, lo que se logra utilizando técnicas de aprendizaje automático.

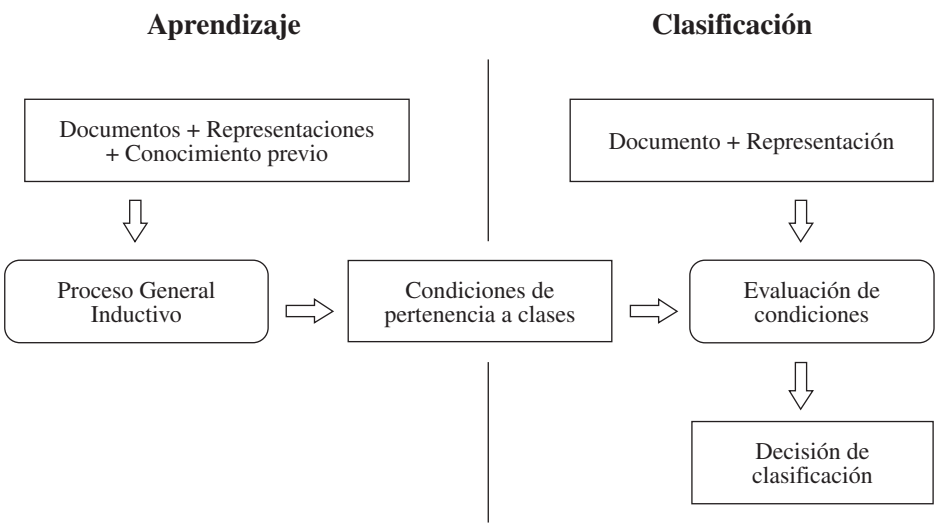
Si tomamos por buena la premisa de que la representación de un documento a partir de la selección y ponderación de las palabras que lo componen puede llegar a contener de forma implícita los temas acerca de los que trata dicho documento, es posible, si todos los documentos se representan conforme a los mismos presupuestos, que el sistema sea capaz de establecer la cercanía temática de dichos documentos y a través de comparaciones entre ellos establecer la pertenencia de un documento determinado a una clase determinada, ya que el clasificador ha sido entrenado para apreciar determinadas características y asociarlas con una categoría determinada.

Representamos en la siguiente figura el proceso de clasificación por un clasificador artificial, donde las condiciones de pertenencia se refieren a las condiciones de pertenencia a una clase determinada, y la decisión de clasificación generalmente será positiva o negativa[5].

[5] Hacemos esta aclaración porque algunos clasificadores emiten esta decisión en forma de coeficientes o porcentajes de pertenencia a una clase, como se verá más adelante.

FIGURA 2

ESQUEMA DE CLASIFICACIÓN AUTOMÁTICA



Como se puede observar, el proceso en su conjunto depende de la evaluación de unas condiciones de pertenencia generadas durante el aprendizaje. Esto puede dar una idea clara de la importancia que un algoritmo de aprendizaje puede llegar a tener en un sistema de clasificación automática. Los algoritmos de aprendizaje utilizados constituyen una materialización de los presupuestos teóricos del Aprendizaje Automático, una rama de la Inteligencia Artificial que, parafraseando a Tom Mitchel (1997: 26), podemos definir como el estudio de algoritmos de computación que mejoran automáticamente a través de la experiencia.

Por lo que se refiere la clasificación automática casi todos los enfoques que se enfrentan al problema de categorizar textos de forma automática tienen su núcleo en algoritmos que permiten el aprendizaje, digamos la parte "inteligente" del sistema.

Sin embargo, todas las inferencias llevadas a cabo por este algoritmo de aprendizaje dependen esencialmente de la calidad del conocimiento codificado de expertos. Por tanto, para mejorar tanto el proceso de aprendizaje como el establecimiento de condiciones de pertenencia a clases será necesario que este conocimiento tenga una elevada calidad. Es aquí donde un adecuado proceso de análisis documental tiene su verdadera importancia.

Es fundamental que se utilicen las prácticas aceptadas del análisis documental para llevar a cabo la clasificación previa de los documentos de entrenamiento, que son claves en los resultados del sistema. Por otra parte parece evidente que el lenguaje documental utilizado tendrá un impacto decisivo en la calidad de este conjunto de entrenamiento.

Por sorprendente que parezca no existe una preocupación real por estos aspectos en la literatura científica del área. Se da por bueno básicamente cualquier sistema de clasificación utilizado, siendo que muchos son extraordinariamente poco refinados, al mismo tiempo que se utilizan los ejemplos preclasificados más accesibles, sin preocupación alguna sobre la calidad del análisis documental.

Los ejemplos preclasificados son imprescindibles para generar las condiciones de pertenencia a clases y al mismo tiempo son utilizados para refinar (modificando diversos parámetros) la configuración del sistema de clasificación automática. Por último este conocimiento también es utilizado en el proceso de evaluación de la calidad de los diferentes sistemas de clasificación automática. La evaluación nos ofrece una medida del éxito de los diferentes algoritmos disponibles y por tanto guía la progresión de las diferentes líneas de investigación en las que están involucradas tareas de clasificación automática.

Nos centraremos en este último aspecto, teniendo siempre en cuenta que los procesos de generación de condiciones de pertenencia a clases (y el aprendizaje en su conjunto) utilizan los mismos presupuestos y los mismos mecanismos que los procesos de evaluación. De esta forma nuestras afirmaciones acerca de la necesidad de mejorar las condiciones de evaluación son extensibles al proceso de aprendizaje.

3. EVALUACIÓN

Una vez examinado el proceso de clasificación automática y obtenida una idea de conjunto sobre el mismo podemos centrarnos en la evaluación de los sistemas de clasificación automática.

La evaluación de los sistemas de clasificación automática es fundamental durante el entrenamiento de los mismos, como ya hemos visto, pero lo es además como medida de la corrección del sistema, como un conjunto de guías que van a posibilitar que decidamos si un determinado sistema es adecuado para nuestras necesidades o no. Posibilita además la comparación entre investigaciones, por lo que en conjunto las medidas de evaluación constituyen un elemento esencial para la investigación en el área.

La evaluación se lleva a cabo por lo general en torno a dos criterios distintos, los criterios de efectividad y los criterios de eficiencia. Para cada uno de estos criterios se diseñan y utilizan medidas distintas. Podemos describir las medidas de evaluación de la eficiencia como aquellas que están relacionadas con el tiempo requerido para llevar a cabo los procesos de aprendizaje y clasificación descritos, así como con los recursos empleados a lo largo de dichos procesos[6]. Por otra parte, las medidas de efectividad se refieren a nuestra capacidad para medir el comportamiento del sistema desde el punto de vista de la calidad de los resultados obtenidos por el mismo.

---

[6] La memoria utilizada por un sistema, el espacio de almacenamiento o la capacidad de proceso requerida, por lo general. Para una revisión exhaustiva de la evaluación de la eficiencia de los sistemas de clasificación véase: YANG, Yiming, ZHANG, Jian y KIESEL, Bryan, "A scalability analysis of classifiers in text categorization". En: *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval*. ACM Press, New York, 2003, pp. 96–103.

3.1. MEDIDAS DE EVALUACIÓN DE LA EFECTIVIDAD

Evidentemente lo más relevante para el caso que nos ocupa es la evaluación de la calidad de los resultados. Para este supuesto se utilizan medidas de evaluación comunes en todo el ámbito de la recuperación de información, como la precisión, la exhaustividad, la medida f y la exhaustividad y precisión interpoladas[7]. Para calcularlas se utiliza una matriz de contingencia que sí es distinta a la habitual y que ofrecemos a continuación.

TABLA 1. MATRIZ DE CONTINGENCIA PARA EVALUACIÓN

Asignación vs. corrección	Documentos asignados	Documentos no asignados
Documentos correctos	Dac	Da¬c
Documentos incorrectos	Dac¬	Da¬c¬

En clasificación automática no existe un conjunto de documentos recuperados, sino un conjunto de documentos asignados a una categoría y tampoco un conjunto de documentos relevantes para una demanda de información, sino un conjunto de documentos cuya categorización bajo una clase sería correcta.

Tendríamos entonces cuatro subconjuntos de documentos, *Dac*, o documentos asignados a la categoría y correctos para dicha categoría, *Da¬c*, documentos no asignados a una categoría pero correctos para dicha categoría, *Dac¬*, documentos asignados a una categoría pero no correctos para dicha categoría y *Da¬c¬* documentos no asignados a una categoría y no correctos para dicha categoría. A partir de estos subconjuntos es fácil construir las medidas de evaluación habituales adaptándolas a la tarea de clasificar documentos.

Sin embargo, para construir estos subconjuntos de documentos es necesario contar con un conjunto de juicios de clasificación emitidos por una fuente de confianza. Es decir, un conjunto de documentos y de categorías puestos en relación por expertos que nos permitan decidir en cada caso cuándo una decisión de asignación a una categoría es correcta o no.

3.2. CORPUS DE PRUEBA

Un corpus de prueba para clasificación automática se puede definir como una colección de pruebas que incluye: documentos que versan sobre temas dispares, asignaciones de dichos documentos a diferentes categorías de un mismo sistema de clasificación y el propio sistema de clasificación.

---

[7] Van Rijsbergen ofrece una revisión exhaustiva de las medidas de evaluación clásicas en Recuperación de Información en: VAN RIJSBERGEN, Keith, *Information Retrieval*. Rutterworths, London, 1979. Véase especialmente el séptimo capítulo.

Para poder evaluar de forma objetiva el comportamiento de un sistema de clasificación automática es imprescindible contar con un corpus de prueba adecuado, una colección de pruebas que, a ser posible, haya sido utilizada por otros investigadores con anterioridad. En la actualidad existen múltiples corpus de prueba utilizados con asiduidad en la literatura científica del área. Sin embargo la calidad de estos corpus de entrenamiento es en nuestra opinión bastante discutible. Hemos detectado importantes deficiencias en su concepción que desvirtúan en cierta forma su capacidad para establecer la calidad de los diferentes sistemas de clasificación automática existentes. Por este motivo describiremos a continuación los bancos de pruebas más utilizados en clasificación automática con la intención de poner de relieve las deficiencias más sustanciales desde el punto de vista de su utilización práctica en tareas de evaluación[8].

3.2.1. OHSUMED

Es una colección bibliográfica de documentos que reúne 348.000 resúmenes procedentes de la base de datos bibliográfica de medicina MEDLINE. Estos resúmenes han sido clasificados mediante la lista de encabezamientos de materia de la *National Library of Medicine* (MeSH)[9].

MeSH es una lista de encabezamientos de materia de calidad, que utiliza reenvíos y notas de alcance y se ha consolidado como una referencia en su campo. Sin embargo no es probablemente la mejor elección para las tareas de clasificación automática, debido al hecho de que para cada documento existe un elevado número de asignaciones de diferentes materias, unas doce de media , y también debido a la ausencia de una estructura jerárquica que es típica de un sistema de clasificación.

Estos dos aspectos están directamente relacionados con el tipo de lenguaje documental utilizado, de estructura combinatoria y no jerárquica. Probablemente no es por tanto la mejor opción utilizar listas de encabezamientos de materias para clasificar los documentos, aunque convivan bien con las clasificaciones en las bibliotecas, dado que son lenguajes documentales diseñados para usos distintos.

En realidad podemos decir que MeSH es útil para un tipo concreto de tarea dentro del ámbito de la clasificación automática (si se entiende esta desde la perspectiva más general de la detección de temas). Esta tarea consistiría en la asignación de categorías múltiples a documentos sin la utilización de una estructura jerárquica, o en otras palabras, la asignación automática de encabezamientos de materia.

Sin embargo, dado el hecho de que la asignación de encabezamientos de materia y la asignación de materias de un sistema de clasificación a documentos son tareas

---

[8] No describiremos aquí otras colecciones de documentos utilizadas para experimentos específicos, sino aquellas que gozan de cierto grado de aceptación en el conjunto de la comunidad investigadora, por lo que son colecciones en Inglés salvo el caso de EFE-DATA.

[9] Se puede obtener una copia de estos encabezamientos en: http://www.nlm.nih.gov/mesh/filelist.html

[10] Véase: YANG, Yiming y PEDERSEN, Jan O. "A comparative study on feature selection in text categorization". En: *Proceedings of ICML-97, 14th International Conference on Machine Learning*. Morgan Kaufmann Publishers, San Francisco, 1997. Pp. 412–420.

distintas, es posible que fuera necesario replantear tanto las formas de aprendizaje como las formas de evaluación de los resultados del sistema.

Desde un punto de vista meramente estadístico, es mucho más fácil que se produzca una sola coincidencia entre la materia asignada por el sistema de clasificación automática y las doce materias asignadas por el experto humano, al mismo tiempo que es extraordinariamente difícil que se produzca una coincidencia plena con el total de las materias asignables.

Por este motivo, cuando obtengamos la matriz de contingencia antes señalada deberemos encontrar una forma nueva de definir el éxito del clasificador, una forma que no se base en ninguno de los extremos mencionados. Esto es algo que no hemos encontrado en ningún trabajo de investigación hasta la fecha.

En resumen, podríamos decir que ni el lenguaje documental ni la técnica de análisis documental utilizada en esta colección es la más apropiada para una colección de pruebas en clasificación automática.

3.2.2. Reuters

Esta es la colección de pruebas que probablemente más se ha utilizado en Clasificación Automática. Es una colección de tamaño relativamente pequeño (en torno a los 22.000 documentos) distribuida en 135 categorías sin una estructura jerárquica. Se trata de noticias, documentos cortos por tanto, relacionadas fundamentalmente con el área de los negocios y la economía. Existen al menos cuatro versiones en las que se utilizan categorías ligeramente distintas[11].

Dada la importancia y representatividad de esta colección reproducimos en la tabla 2 un extracto de las categorías de Reuters. Como se puede observar no existen relaciones jerárquicas, salvo que consideremos como tales las existentes entre los encabezados del tipo *Economic Indicator Codes* y las categorías propiamente dichas. Sin embargo no existen asignaciones de documentos a lo que podríamos denominar clases principales, ni juegan ningún otro papel funcional en el proceso de evaluación, por lo que carecen de utilidad práctica.

Lo segundo que llama nuestra atención es que se mezclan al mismo tiempo materias y temas, ya que "carbón", "coco", "balanza de pagos"... pueden pasar perfectamente por descriptores de un tesauro, mientras que "Ingresos y Previsión de Ingresos" (*Earnings and Earnings Forecasts)* parece claramente una materia.

De esta forma podemos poner en duda que el lenguaje controlado utilizado en la colección de Reuters sea en realidad un sistema de clasificación. Se puede caracterizar como un híbrido entre lista de encabezamientos y lista de descriptores, ya que no posee estructura ni se decide por un criterio claro de representación de los documentos. Teniendo en cuenta que la colección de Reuters ha sido clasificada con estas categorías parece fácil deducir que el resultado distará de ser idóneo.

---

[11] La versión Reuters-21578, que nosotros hemos revisado con ocasión de este trabajo, se puede obtener en línea en: http://www.daviddlewis.com/resources/testcollections/reuters21578/

Por otra parte, durante el proceso de análisis documental se asignan múltiples categorías a cada uno de los documentos, por lo que tampoco queda claro que el empleo del lenguaje controlado sea el más idóneo. En esta situación, ¿qué validez tiene el corpus de Reuters para evaluar los experimentos sobre clasificación automática? Desde el punto de vista práctico parece que los sistemas de clasificación automática aplicados a Reuters han respondido bastante bien a las expectativas, es decir, que son capaces de asignar las categorías con corrección durante el proceso de validación y el de pruebas, pero no resulta fácil mostrarse seguros acerca de la posibilidad de extender las conclusiones obtenidas en Reuters a otros entornos de trabajo.

Tabla 2. Extracto del listado de categorías de Reuters

****Subject Codes (135)	
Money/Foreign Exchange (MONEY-FX)	COPRA-CAKE
Shipping (SHIP)	CORN-OIL
Interest Rates (INTEREST)	CORN CORNGLUTENFEED
**Economic Indicator Codes (16)	COTTON COTTON-MEAL
Balance of Payments (BOP)	COTTON-OIL
Trade (TRADE)	COTTONSEED
Consumer Price Index (CPI)	F-CATTLE
Wholesale Price Index (WPI)	FISHMEAL
Unemployment (JOBS)	FLAXSEED
Industrial Production Index (IPI)	GOLD
Capacity Utilisation (CPU)	GRAIN
Gross National/Domestic Product (GNP)	GROUNDNUT
Money Supply (MONEY-SUPPLY)	GROUNDNUT-MEAL
Reserves (RESERVES)	GROUNDNUT-OIL
Leading Economic Indicators (LEI)	IRON-STEEL
Housing Starts (HOUSING)	LEAD
Personal Income (INCOME)	LIN-MEAL
Inventories (INVENTORIES)	LIN-OIL
Instalment Debt/Consumer Credit (INSTAL-DEBT)	LINSEED LIVESTOCK
Retail Sales (RETAIL)	L-CATTLE HOG
**Currency Codes (27)	LUMBER LUPIN

TABLA 2 *(CONTINUACIÓN)*. EXTRACTO DEL LISTADO DE CATEGORÍAS DE REUTERS

****Subject Codes (135)	
U.S. Dollar (DLR)	MEAL-FEED
Sterling (STG)	NICKEL
D-Mark (DMK)	OAT
Japanese Yen (YEN)	OILSEED
Swiss Franc (SFR)	ORANGE
French Franc (FFR)	PALLADIUM
Belgian Franc (BFR)	PALM-MEAL
Netherlands Guilder/Florin (DFL)	PALM-OIL
Brazilian Cruzado (CRUZADO)	PALMKERNEL
Argentine Austral (AUSTRAL)	PLATINUM
Saudi Arabian Riyal (SAUDRIYAL)	PLYWOOD
South African Rand (RAND)	PORK-BELLY
Indonesian Rupiah (RUPIAH)	POTATO
Malaysian Ringitt (RINGGIT)	RAPE-MEAL
Portuguese Escudo (ESCUDO)	RAPE-OIL
Spanish Peseta (PESETA)	RAPESEED
Greek Drachma (DRACHMA)	RED-BEAN
…	RICE
**Corporate Codes (2)	SORGHUM
	WHEAT
Mergers/Acquisitions (ACQ)	…
Earnings and Earnings Forecasts (EARN)	
	**Energy Codes (9)
**Commodity Codes (78)	
	Crude Oil (CRUDE)
ALUM	Heating Oil/Gas Oil (HEAT)
BARLEY	Fuel Oil (FUEL)
CARCASS	Gasoline (GAS)
CITRUSPULP	Natural Gas (NAT-GAS)
COCOA	Petro-Chemicals (PET-CHEM)
COCONUT-OIL	Propane (PROPANE)
COCONUT	Jet and Kerosene (JET)
COFFEE	Naphtha (NAPHTHA)
COPPER	

Dado que los temas reflejados en este listado hacen referencia a conceptos bastante concretos, utilizar este tipo de lenguaje implica un enfoque analítico sobre los temas de los documentos. En otras palabras, la descripción de los documentos pasará muchas veces por ser una indización por descriptores en lugar de una indización por materias. Por ejemplo, para un determinado documento se puede llegar a utilizar una descripción en la que se incluyen los siguientes temas de Reuters: *grain, wheat, corn, barley, oat, sorghum*.

3.2.3. Corpus 20NG

20 News Groups[12] es una colección de documentos provenientes de los grupos de noticias UseNet, recopiladas durante 1993. Los documentos están agrupados en torno a los grupos de discusión de UseNet, por lo que cada grupo de discusión se asimila a una categoría. Reproducimos a continuación las categorías de 20NG:

Tabla 3. Categorías 20NG

alt.atheism	rec.autos	sci.space
comp.graphics	rec.motorcycles	soc.religion.christian
comp.os.ms-windows.misc	rec.sport.baseball	talk.politics.guns
comp.sys.ibm.pc.hardware	rec.sport.hockey	talk.politics.mideast
comp.sys.mac.hardware	sci.crypt	talk.politics.misc
comp.windows.x	sci.electronics	talk.religion.misc
misc.forsale	sci.med	

Existen dos niveles jerárquicos, de los cuales el superior contiene cinco clases principales, charla, entretenimiento, ciencia, computadores y miscelánea, que se dividen a su vez en varias subcategorías, hasta un total de 20, y que incluyen temas heterogéneos como la medicina, la religión o el software de encriptación de documentos. En total 20.000 documentos distribuidos de forma bastante regular por las veinte categorías. Sin embargo, como en el caso de Reuters esta jerarquía no tiene un papel funcional en la colección de pruebas dado que las categorías más genéricas no reciben la asignación de ningún documento en concreto, lo que hace difícil poner a prueba la capacidad de un sistema para decidir sobre la generalidad o especificidad de los temas.

[12] Podemos obtener el conjunto de los documentos en: http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.tar.gz

En este caso no existe un problema de asignación múltiple, ya que cada documento pertenece únicamente a una categoría. Sin embargo la asimilación del esquema de las categorías de UseNet a un sistema de clasificación propiamente dicho no parece adecuada.

3.2.4. Colección EFE-DATA

La colección EFE-DATA ha sido utilizada como elemento de evaluación tanto para tareas de recuperación de información monolingüe como para tareas de recuperación de información multilingüe en el TREC. No conocemos hasta la fecha su utilización como marco de evaluación para tareas de clasificación automática, aunque sí la utilización de otras colecciones de prensa en similares circunstancias. Introducimos aquí nuestras reflexiones sobre EFE-DATA como un reflejo de las colecciones de prensa que es extensible a otros ejemplos concretos en castellano, como la colección de documentos de El Mundo utilizada por Zazo, Figuerola y Berrocal[13].

En conjunto las colecciones de prensa permiten utilizar las secciones de tipo periodístico como categorías y la distribución de los documentos sobre las mismas como ejemplos de asignación a categorías por parte de expertos. Sin embargo la utilidad de estas categorías es dudosa desde un punto de vista estricto. Reproducimos a continuación las categorías principales de la colección EFE-DATA 1994 para ilustrar este aspecto.

Tabla 4. Categorías EFE-DATA

AUTONÓMICAS	CASA REAL	CEE	CIENCIA
CORTES	CULTURA	DEFENSA	DEPORTES
DOCUMENTACIÓN	ECONOMÍA	EDUCACIÓN	EXTERIORES
GOBIERNO	LABORAL	MUNICIPALES	ORDEN PÚBLICO
OTAN	PARTIDOS	POLÍTICA	PRENSA
PRUEBAS	PUERTO RICO	REGIONAL	RELIGIÓN
SANIDAD	SOCIEDAD	SUCESOS	TIEMPO
TOROS	TRIBUNALES	VARIOS	

Como se puede observar las categorías son de amplio espectro y reducida definición, lo que tiende a producir fenómenos de solapamiento de varias categorías con

[13] Figuerola, Zazo y Berrocal utilizan por ejemplo la colección de documentos de El Mundo, sobre la que no reflexionaremos aquí por no disponer de una copia original de la misma. Para más detalles véase: FIGUEROLA, Carlos G., ZAZO, Ángel F. y BERROCAL, José L. "Categorización automática de documentos en español: algunos resultados experimentales". En: actas de las Jornadas sobre Bibliotecas Digitales, JBIDI (2000).

respecto de un solo documento. Esto hace que a la hora de evaluar los resultados un documento que objetivamente parece bien indicado para una categoría no reciba una evaluación positiva basándose en la asignación original, dado que cada documento sólo puede estar en una categoría y estas no se han definido con la suficiente precisión. Se echa en falta de nuevo la existencia de una estructura jerárquica que permita incluir materias más específicas en las que incluir a los documentos.

4. OBSERVACIONES SOBRE LA EVALUACIÓN EN CLASIFICACIÓN AUTOMÁTICA

Como se ha podido observar con anterioridad los lenguajes controlados utilizados para los corpus de pruebas distan de ser perfectos. Sintetizaremos nuestra observación de estas imperfecciones en torno a tres puntos fundamentales: tipo de conceptos, estructura de los lenguajes y utilización de los mismos en la representación de documentos.

En primer lugar, podemos observar cierta confusión a la hora de establecer la tipología de conceptos para cada una de las entradas del lenguaje controlado utilizado. Esto se puede apreciar examinando la capacidad de representación de las categorías utilizadas en los mismos. Estas no tienen en muchos casos un carácter consistente, de forma que se alternan términos que incluyen una noción de conceptos de carácter analítico y términos que incluyen una noción de concepto sintético, o materias propiamente dichas. Siguiendo a Maniez (1992: 208) no podemos describir como sistemas de clasificación aquellos lenguajes controlados que no se basen de forma exclusiva en conceptos de carácter sintético.

Reuters mezcla conceptos analíticos y conceptos sintéticos, como también lo hace EFE-DATA. El lenguaje de EFE-DATA parece también más un listado de descriptores que de materias. En el otro extremo, MESH es evidentemente el lenguaje de más calidad de entre los examinados, dado que en su construcción sí se utilizan únicamente materias. Para 20NG se utilizan materias, aunque la expresión de estas es bastante deficiente.

Respecto a la jerarquía de conceptos, no existe ninguna estructura jerárquica. Incluso en los casos de 20NG y Reuters esta jerarquía es meramente nominal, ya que no se utiliza de ninguna forma en la descripción de los contenidos incluida en la colección de pruebas ni se le otorga ningún papel práctico.

De esta forma sólo podemos distinguir con claridad la tipología documental de MESH, que claramente responde a la de un listado de encabezamientos de materia. El lenguaje controlado utilizado en Reuters se podría asimilar más a un listado de descriptores, aunque como hemos apuntado se utilizan materias además de descriptores. 20NG utiliza las categorías temáticas de UseNet, y no un sistema de clasificación propiamente dicho y el lenguaje utilizado en EFE-DATA se puede definir como un listado de descriptores sin estructura jerarquica.

El tercer aspecto a considerar es el empleo de los lenguajes controlados en la colección de pruebas. Durante el proceso de análisis documental se alternan los estilos de asignación única y de asignación multicategoría, de representación por síntesis y por análisis. En el caso de MESH y dada la estructura combinatoria de las lis-

tas de encabezamientos de materia, la representación de los documentos en torno a múltiples encabezamientos de materia entra dentro de la práctica normal.

Parece menos aconsejable utilizar listados de descriptores para terminar por asignar un único descriptor a cada documento, como en el caso de EFE-DATA. Como alternativa cabe pensar que las entradas del lenguaje controlado utilizado pretenden ser materias, salvo que los temas que recogen son claramente analíticos en muchos de los casos. Muchas de las noticias pueden ser descritas por varios de los elementos de que consta el listado de descriptores de EFE-DATA, lo que lleva a confusión y a una evaluación no siempre adecuada de los resultados.

El caso de Reuters es bastante similar en cuanto a la poca adecuación del lenguaje diseñado y su empleo final. Salvo que en este caso es frecuente encontrar varias materias asignadas a un mismo documento, al mismo tiempo que otro documento puede ser representado entorno a varios descriptores. El último caso es el más frecuente, de forma que la representación de los documentos resulta en una indización por descriptores, en lugar de una indización por materias.

En cualquier caso la representación final de los documentos dista mucho de ser ideal, dado que muchos de ellos podrían pertenecer en realidad a más de una de las categorías propuestas. De esta forma podemos observar cómo el diseño del lenguaje controlado y su uso posterior para el análisis documental no son realmente coherentes.

Si acordamos que un sistema de clasificación debe tener estructura jerárquica, ser precoordinado y ofrecer una representación de los documentos en torno a conceptos sintéticos, ninguno de los lenguajes controlados utilizados en las colecciones de pruebas más frecuentes cumple con estas características. Con esto podemos concluir que los lenguajes documentales utilizados son inapropiados desde un punto de vista formal.

Por último, salvo excepciones previamente mencionadas, podemos decir que las colecciones de pruebas analizadas utilizan lenguajes documentales mal definidos, que alcanzan una profundidad insuficiente y no tienen estructura jerárquica funcional. Con esto podemos concluir que la calidad de los lenguajes documentales utilizados es bastante deficiente.

4.1. EFECTOS DE LA BAJA CALIDAD DE LOS LENGUAJES DOCUMENTALES SOBRE LA EVALUACIÓN

Los tres aspectos mencionados introducen un factor de incertidumbre. Para empezar, la indefinición de los lenguajes controlados propuestos no nos permite asegurar que estemos clasificando la colección. Dado que no estamos utilizando un sistema de clasificación propiamente dicho no podemos afirmar que el análisis documental realizado sobre su base resulta en una indización por materias.

Desde un punto de vista objetivo la indización por descriptores y la indización por materias son tareas distintas. Si esta distinción no se tiene en cuenta a la hora de establecer corpus por los que evaluar los resultados de los sistemas de clasificación automática, no podemos concluir que dichos corpus sirvan para evaluar la calidad de los resultados de clasificación. Sí podremos decir que sirven para evaluar la capaci-

dad de un determinado sistema para reconocer temas comunes en documentos distintos, pero una cosa y la otra no son en puridad lo mismo.

Esto no anula nuestra capacidad para decidir cuándo un determinado sistema de clasificación automática es bueno o no. Sin embargo se introduce un factor de imprecisión notable a la hora de cuantificar esta calidad, y de ofrecer comparaciones entre los resultados obtenidos en marcos de evaluación que utilizan diferentes corpus de pruebas.

En realidad estos corpus de pruebas han sido utilizados sin mayores problemas para la evaluación de sistemas de clasificación automática y los resultados de diferentes investigadores corroboran la calidad de algunos de estos sistemas frente al bajo rendimiento de otros con bastante consenso[14]. Esto se explica porque los elementos con que se evalúa el sistema han sido "categorizados" con las mismas herramientas que los elementos sobre los que el clasificador lleva a cabo su proceso inductivo de establecimiento de condiciones de pertenencia. Por tanto, la correspondencia, si el aprendizaje es bueno, ha de ser necesariamente correcta.

De esta forma podemos deducir que el sistema aprende bien, pero no necesariamente que clasifica bien. Los medios de evaluación utilizados, basados en lenguajes documentales y colecciones defectuosas, son útiles desde el punto de vista del aprendizaje automático, área de investigación de la que proceden muchos de los autores que publican sobre clasificación automática, pero no desde el punto de vista de los criterios documentales.

Desde un punto de vista estrictamente documental persiste todavía el problema de conocer cómo reaccionarían los sistemas de clasificación automática si tuvieran que elaborar condiciones de pertenencia a verdaderas materias, en lugar de a temas de carácter analítico o una mezcla de ambos, qué impacto tendría la utilización de una estructura jerárquica, o cuáles serían las diferencias para las tareas de asignación unicategoría y multicategoría.

Detectado este problema sería necesario evaluar la capacidad de trabajar con corrección de los diferentes modelos de clasificación para diferentes tareas. Podemos resumir estas en tres, clasificación automática, asignación de encabezamientos de materias e indización mediante descriptores. La primera de ellas se basa en la utilización de sistemas de clasificación, con una estructura jerárquica de materias de entre las cuales se elige una para cada documento. La segunda de ellas utiliza listas de encabezamientos de materia, sin estructura jerárquica y de asignación múltiple basada en materias. Por último la indización por descriptores se basaría en la asignación de múltiples descriptores de entre los existentes en un lenguaje controlado.

---

[14] Véanse por ejemplo los trabajos: CARDOSO-CACHOPO, Ana y OLIVEIRA, Arlindo L. "An Empirical Comparison of Text Categorization Methods". En: *Proceedings of SPIRE-03, 10th International Symposium on String Processing and Information Retrieval*. Heidelberg, Springer Verlag, 2003. Pp. 183-196. YANG, Yiming, ZHANG, Jian y KIESEL, Bryan, "A scalability analysis of classifiers in text categorization". En: *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval*. ACM Press, New York, 2003. Pp. 96–03. y por último el trabajo clásico de Lewis: LEWIS, David D., "Evaluating Text Categorization." En: *Proceedings of Speech and Natural Language Workshop*. Morgan Kaufmann, 1991, 312–318.

Dado que existen muchos tipos de algoritmos de aprendizaje, la base de los sistemas de clasificación automática, es posible que los diferentes modelos sobre los que estos se basan (probabilísticos, basados en reglas lógicas, de espacio vectorial… etc) tengan diferentes capacidades de trabajo para los casos planteados. Es decir, que quedan por demostrar aspectos tales como que un sistema de clasificación automática determinado, por ejemplo uno basado en K-vecinos, se comporte con el mismo nivel de correción para cada una de las tareas propuestas o que este sistema sea más efectivo para una tarea en concreto que otros sistemas en principio mejor valorados en la literatura científica.

Podemos concretar estas ideas con un ejemplo práctico. Existen dos tipos de clasificadores desde el punto de vista de las decisiones de clasificación que producen. La mayor parte de los modelos de clasificación existentes producen una función del tipo *C : DxC -> {V, F}*, (Sebastiani 2003) donde V (verdadero) y F (falso) son las dos únicas decisiones posibles para una categoría y un documento dados[15]. El resto de los clasificadores producen funciones de decisión del tipo VC : DxC -> [0, 1], donde [0, 1] indica un valor real entre 0 y 1. Es decir, que con este segundo tipo de clasificadores podemos evaluar el grado de pertenencia de un documento a una categoría.

Este factor es de gran importancia a la hora de llevar a cabo los dos tipos posibles de asignación de documentos a materias, la asignación a múltiples materias y la asignación a una única materia. Si nos encontramos en la necesidad de representar los documentos a través de una única materia, evidentemente el segundo tipo es con mucho el más deseable. Los clasificadores que utilizan funciones de decisión cuyo resultado es un número real son mucho mejores a la hora de decidirse sobre una única categoría que el resto de los clasificadores, dado que los clasificadores binarios no cuentan con ninguna condición de partida para evaluar la prevalencia de una sola materia en concreto.

De esta forma, podemos concluir que para la tarea de clasificar un conjunto de documentos sobre una única materia, los sistemas del estilo de K-vecinos, WORD o LLSF y las últimas versiones de SVM (todos ellos clasificadores real-valuados y no binarios) son a priori mucho más adecuados que el resto.

Este es sólo un ejemplo de un aspecto que merece comprobación en el futuro, pero de la reflexión que surge al diferenciar claramente los lenguajes documentales pueden surgir muchos otros aspectos que debieran ser comprobados, y que no obstante quedan fuera de los objetivos de este artículo, que sólo se propone poner de manifiesto la necesidad de investigar en este sentido.

5. CONCLUSIONES

En primer lugar creemos que es palpable la necesidad de elaborar colecciones de pruebas en mejores condiciones que las actuales. Estas mejoras deberían ir en la

[15] Los clasificadores basados en reglas DNF, Naive Bayes, Redes Neurales, Clasificadores Lineales basados en el Algoritmo de Rocchio, Árboles de Decisiones y Máquinas de Soporte Vectorial, entre otros, producen funciones de clasificación binarias.

línea de utilizar sistemas de clasificación de calidad, con materias bien perfiladas y una estructura jerárquica clara.

En segundo lugar es imprescindible hacer un uso adecuado de los lenguajes documentales durante el proceso de análisis documental llevado a cabo por los expertos, de manera que los sistemas de clasificación no se utilicen como listados de descriptores, ni a la inversa, o se produzca cualquiera de los malos usos identificados a lo largo del artículo.

En tercer lugar cabe instar a la creación de formas de evaluación específicas para cada una de las tareas propuestas, asignación automática de materias desde un listado de encabezamientos de materias, asignación automática de materias desde un sistema de clasificación y asignación automática de descriptores desde un listado de descriptores o desde un tesauro.

Por último es muy deseable la utilización de técnicas de análisis documental adecuadas para el establecimiento de los documentos de entrenamiento que están en la base del buen funcionamiento de cualquier sistema de clasificación sobre una colección de documentos concreta.

6. BIBLIOGRAFÍA

BAEZA YATES, Ricardo y RIBEIRO-NETO (1999), Berthier, *Modern Information Retrieval*. Addison Wesley & ACM Press, Nueva York, p. 13.

CARDOSO-CACHOPO, Ana y OLIVEIRA, Arlindo L. (2003) "An Empirical Comparison of Text Categorization Methods". En: *Proceedings of SPIRE-03, 10th International Symposium on String Processing and Information Retrieval*. Heidelberg: Springer Verlag, pp. 183-196.

CHAN, M.L. (1981) *Cataloging and classification: an introduction*. New York: McGraw-Hill.

DUMAIS, Susan T, PLATT, John, HECKERMAN, David y SAHAMI, Mehran Sahami (1998). "Inductive learning algorithms and representations for text categorization". En: *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Georges Gardarin, James C. French, Niki Pissinou, Kia Makki y Luc Bouganim, eds.). Nueva York: ACM Press, pp. 148-155.

FIGUEROLA, Carlos G., ZAZO, Ángel F. y BERROCAL, José L. (2000). "Categorización automática de documentos en español: algunos resultados experimentales". En: actas de las Jornadas sobre Bibliotecas Digitales, JBIDI.

GIL URDICIAIN, Blanca (1996). *Manual de Lenguajes Documentales*. Madrid: NOESIS.

LEWIS, David D. (1991). "Evaluating Text Categorization." En: *Proceedings of Speech and Natural Language Workshop*. Morgan Kaufmann, pp. 312-318.

LEWIS, David D., y RINGUETTE, Marc (1994). « A comparison of two learning algorithms for text categorization". En: *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, pp. 81–93.

MANIEZ, Jacques (1992). *Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales*. Madrid: Fundación Germán Sánchez Ruipérez.

MITCHELL, Tom (1997). *Machine Learning*. New York: McGraw-Hill.

SALTON, Gerard y BUCKLEY, Christopher (1988). "Term-weighting approaches in automatic text retrieval." *Information Proccessing and Management*, Vol. 5, no 24, pp. 323–328.

SEBASTIANI, Fabrizio (2003). "Research in Automated text Classification: trends and perspectives". En: *Actas del 6º Congreso del Capítulo Español de ISKO*. Salamanca.

VAN RIJSBERGEN, Keith (1979). *Information Retrieval*. London: Rutterworths.

YANG, Yiming (1999). "An evaluation of statistical approaches to text categorization." *Information Retrieval*, Vol. 1, no 1/2, pp. 69-90.

YANG, Yiming y PEDERSEN, Jan O. (1997). "A comparative study on feature selection in text categorization". En: *Proceedings of ICML-97, 14th International Conference on Machine Learning*. Morgan Kaufmann Publishers, San Francisco, pp. 412–420.

YANG, Yiming, ZHANG, Jian y KIESEL, Bryan (2003). "A scalability analysis of classifiers in text categorization". En: *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval*. New York: ACM Press, pp. 96-103.