Categorización automática de documentos en español: algunos resultados experimentales

Carlos G. Figuerola, Angel F. Zazo, José Luis Alonso Berrocal

Universidad de Salamanca, Facultad de Documentación,
C/ Fco. Vitoria, 6-16, 37008-Salamanca (ESPAÑA)
e-mail: [figue|afzazo|berrocal]@gug.usal.es

Resumen. La categorización automática puede contemplarse como un proceso de aprendizaje, durante el cual un programa capta las características que distinguen cada categoría o clase de las demás, es decir, aquéllas que deben poseer los documentos para pertenecer a esa categoría. De otro lado, pocos experimentos se han efectuado todavía con documentos en español. Se muestran las posibilidades de elaborar vectores patrón que recojan las características de distintas clases o categorías de documentos, utilizando técnicas basadas en aquéllas aplicadas en la expansión de consultas por relevancia. Al mismo tiempo, se describe un experimento consistente en la aplicación de esas técnicas a una colección de noticias de prensa en español, para su categorización. Los resultados obtenidos son, en conjunto, homologables o incluso mejores que los obtenidos en experimentos similares; para algunas de las categorías, estos resultados han sido muy favorables.

1 Introducción

La clasificación automática de documentos ha sido ampliamente estudiada por diversos investigadores. Su utilidad se basa en la posibilidad de poder efectuar posteriormente una adecuada recuperación, asumiendo que aquellos textos que tratan de la misma materia están clasificados juntos, o en apartados cercanos. Diversas técnicas han sido propuestas, desde hace ya bastantes años. Así, Fairthorne [1] y Hayes [2] sugirieron, separadamente, la posibilidad de utilizar sistemas de clasificación como un modo de aumentar la eficacia en la recuperación de información. Aunque el propio Salton [3] cree interesante la agrupación de documentos, estima que resta efectividad a la recuperación.

Buena parte de tales técnicas se basan en la utilización de medidas de semejanza (o de disparidad, dependiendo del punto de vista) entre dos documentos. Una exposición de las más importantes, tanto de unas como de otras, puede encontrarse en [4]. Utilizando algún sistema que permita asociar documentos entre sí, se pueden constituir esquemas clasificatorios complejos de forma automática.

Sin embargo, en muchas ocasiones, el problema es, en cierta medida, el contrario. Se parte de un esquema de clasificación previo, ya establecido, y la cuestión es decidir en qué lugar de este esquema debe ir cada documento. De otro

lado, el sistema de clasificación, sus distintas categorías y subcategorías, no siempre se basan exclusivamente en el contenido temático de los documentos. En la elaboración del cuadro clasificatorio pueden influir presunciones apriorísticas, desde luego, pero también cuestiones históricas, administrativas, incluso de tipo legal. En estos casos, cuando se parte de un esquema de clasificación que debe respetarse, hablamos de categorización de documentos: es decir, de seleccionar la categoría, establecida de antemano, en la cual hay que insertar cada nuevo documento que llega a la colección.

2 La Categorización de Documentos y las técnicas clásicas de Recuperación de la Información

Desde este punto de vista, la categorización automática puede contemplarse como un proceso de aprendizaje, durante el cual el programa capta las características que distinguen cada categoría o clase de las demás, es decir, aquéllas que deben poseer los documentos para pertenecer a esa categoría. Estas características no tienen porqué indicar de forma absoluta la pertenencia a una clase o categoría, sino que más bien lo hacen en función de una escala o graduación. De esta forma, por ejemplo, documentos que posean una cierta característica tendrán un factor de posibilidades de pertenecer a determinada clase. De modo que la acumulación de dichas cantidades puede arrojar un resultado consistente en un coeficiente asociado a cada una de las clases existentes. Este coeficiente lo que expresa en realidad es el grado de confianza o certeza de que el documento en cuestión pertenezca a la clase asociada al coeficiente resultante.

Puede observarse, pues, cierta semejanza entre este proceso y el utilizado en los sistemas de recuperación basados en realimentación por relevancia

2.1 El Modelo Vectorial y la Representación de Categorías

El modelo vectorial, definido por Salton hace ya bastantes años [3], es empleado ampliamente en operaciones de IR, y puede utilizarse también para explicar el proceso de categorización automática. Así, un documento puede considerarse como un vector $D = (c_1, c_2, c_3 \ldots c_j)$, es decir, como un conjunto de características, hasta un total de j, y en el cual c_1 es un valor numérico que expresa en qué grado el documento D posee la característica 1, c_2 lo mismo para la característica 2, y así sucesivamente.

El concepto 'característica' suele concretarse en la ocurrencia de determinas palabras en el documento, aunque nada impide tomar en consideración otros factores. En el caso más simple, pueden aplicarse valores binarios exclusivamente; de forma que si en el documento D aparece la palabra 1, el valor de c_1 sería 1 y en caso contrario, 0. Como, naturalmente, una palabra puede aparecer más de una vez en el mismo documento, y, como, además, unas palabras pueden considerarse como más significativas que otras, el valor numérico de cada uno de los componentes del vector obedece a cálculos algo más sofisticados que tienen en cuenta más factores, además de la simple ocurrencia o no de un término.

Se han propuesto diversos sistemas para calcular dicho valor numérico, es decir, el peso de cada término contemplado, para cada documento. En general, se tiene en cuenta para esto la frecuencia inversa (IDF), combinándola de alguna forma con la frecuencia del término dentro del documento [5]. Salton y Buckley [6] experimentaron con más de 200 sistemas de asignación o cálculo de pesos, así que hay bastante dónde elegir.

En operaciones clásicas de búsqueda de documentos, la consulta efectuada puede representarse igualmente por medio de un vector, con igual número de elementos, y en el cual el valor de cada uno de éstos expresaría el grado en que cada uno de los términos representa las necesidades de información de la persona que hace la consulta.

Así, la resolución de la consulta consiste en un proceso de establecer el grado de similaridad entre el vector consulta y cada uno de los vectores de cada uno de los documentos. Para una consulta determinada, pues, cada documento arroja un grado de similaridad determinado; aquéllos cuyo grado de similaridad sea más elevado, se ajustarán mejor a las necesidades expresadas en la consulta; es decir, serán más relevantes respecto de esa consulta.

El modo más simple de calcular esta similaridad es el producto de ambos vectores, consulta y documento. Habitualmente, se desea una normalización en los resultados, a fin de obviar distorsiones producidas por los diferentes tamaños de los documentos. Hay también unos cuantos métodos propuestos para calcular la similaridad; un cuadro con los más importantes puede encontrarse en [7].

Partiendo de estas ideas, y volviendo al campo de la categorización, podemos intentar establecer un vector para cada una de las clases o categorías posibles, que recoja las características de cada una de dichas clases. Para operaciones de clasificación y categorización, el mecanismo básico consiste en medir la similaridad del vector de cada documento con cada uno de los vectores patrón que contienen las características de las clases o categorías. Obviamente, aquel vector patrón de clase que ofrezca mayor similaridad con el vector del documento será el que con más confianza indique la clase o categoría a la cual pertenece o debe asignarse el documento en cuestión.

2.2 La Realimentación de Consultas y la Construcción Automática de Patrones de Categorías

Pero la cuestión está en cómo construir los vectores patrón representativos de cada categoría o clase. Nuevamente, podemos tomar prestadas ideas de la recuperación de documentos. Muchos sistemas aplican un mecanismo de realimentación, a través del cual, después de una primera consulta y sus correspondientes resultados, utilizan aquellos documentos señalados por el usuario como más relevantes para reformular de forma automática la consulta, extrayendo términos de estos documentos más relevantes y añadiéndolos a la consulta original y recalculando los pesos de los términos.

Así pues, si disponemos de una colección de documentos categorizados manualmente, y adscritos a una clase determinada, es posible aplicar dichos mecanismos de realimentación para construir un vector patrón, representativo de esa

clase. Los nuevos documentos a categorizar pueden ser confrontados con ese vector patrón, calculando la similaridad entre ambos y, en función del grado de ésta, se puede determinar su asignación o no a esa clase.

Diversos sistemas se utilizan en los procesos de realimentación, para construir un nuevo vector de consulta [8], y que pueden ser aplicados a la categorización, para construir los vectores patrón de cada clase o categoría.

Uno de los más utilizados es el algoritmo de Rocchio [9], el cual, en su forma estándar responde a la siguiente fórmula

$$C_1 = C_0 + \beta \sum_{i=1}^{nr} \frac{R_i}{n_r} - \gamma \sum_{i=1}^{nnr} \frac{NR_i}{n_{nr}} \quad (1)$$

donde

C_0 es el vector de la consulta original

R_i es el vector del documento relevante *i*

NR_i es el vector del documento no relevante *i*

n_r es el número de documentos relevantes

n_{nr} es el número de documentos no relevantes

β y γ son constantes que permiten ajustar el impacto de los documentos relevantes y los no relevantes.

Existen otros algoritmos utilizables, algunos de los cuales pueden verse en el trabajo de Harman, ya citado [8]. En [10] puede encontrarse una revisión de varios algoritmos aplicados directamente a la categorización.

3 Descripción del Experimento

Hemos llevado a cabo un experimento de categorización automática de textos, utilizando, tanto para el entrenamiento, como para la categorización propiamente dicha dos colecciones de noticias de prensa extraídas del periódico en español 'EL MUNDO'. La razón para la elección de estas colecciones estriba en que dicho periódico edita desde 1994 una versión semestral en CD ROM completa de todo lo publicado diariamente en papel; esto ha permitido disponer de forma fácil de un número elevado de textos o documentos.

De otro lado, cada noticia se encuentra ya categorizada, al constar en el CD la sección del periódico en que fue publicada. Esto simplifica las operaciones de entrenamiento, ya que no es necesario efectuar una categorización manual; y también la comprobación de los resultados del experimento.

Las noticias se encuentran, obviamente, en español, lo cual proporciona un interés añadido al experimento. En efecto, la investigación de tipo práctico en IR sobre documentos en español es escasa [11], si bien en los últimos años comienzan a verse trabajos sobre textos en esta lengua. Particularmente notable es la inclusión

entre las colecciones que sirven de base a los experimentos TREC de documentos en español [12, 13, 14], así como en Conferencias CLEF [15].

3.1 La colección de entrenamiento

Como colección de entrenamiento se han utilizado 2.741 noticias publicadas en enero de 1994. Se trata de documentos de una extensión media de 3.603 caracteres, con notable uniformidad en cuanto a su tamaño. Es preciso destacar que hemos trabajado exclusivamente con noticias, habiendo desechado materiales como artículos de opinión, editoriales, etc...

Las noticias, por otra parte, corresponden a diferentes secciones del periódico. El número de noticias o documentos utilizados de cada una de las secciones ha sido aproximadamente similar, aunque no exactamente el mismo. Debe tenerse en cuenta que se ha buscado abarcar un rango temporal (referido a las fechas de las noticias) compacto, en la idea de que las características de cada sección pueden variar notablemente con el tiempo, al tratarse de noticias de un diario. La Tabla 1 recoge las secciones de las cuales se han extraído noticias, así como el número de ellas, tanto para la fase de entrenamiento como para las pruebas del sistema.

Se ha partido de la base de que cada una de estas secciones constituye una clase o categoría. Y, aunque se trata de áreas temáticas diferenciadas, nótese que entre algunas de ellas pudieran darse solapamientos: por ejemplo, entre Bolsa y Economía, o entre Campus (Educación) y Cultura.

La única operación de preprocesado efectuada ha sido la conversión en mayúsculas de todas las letras, así como la eliminación de acentos. Aunque, en español. los acentos son un elemento importante, hasta el punto de que pueden definir por sí solos palabras completamente diferentes, es un hecho cierto que cada vez se tiende más a no utilizarlos, a hacerlo indebidamente, o, al menos, a ser poco cuidadoso con su uso. Esto hace que, desde el punto de vista del procesado de cadenas de caracteres, constituyan un elemento de distorsión.

De otro lado, no hemos empleado ningún sistema de lematización, que nos permitiera trabajar con términos normalizados en lugar de palabras en bruto. En efecto, la lematización es altamente dependiente del idioma particular [16]; y el español es una lengua especialmente rica y compleja desde un punto de vista morfológico. Trabajos experimentales ha mostrado el fracaso de los sistemas que se utilizan con el inglés cuando se aplican al español [17].

No obstante, esto no parece haber sido una dificultad importante, debido, probablemente, a la sencillez sintáctica y a la limitada variedad morfológica que parecen caracterizar a los texto periodísticos [18, 19].

3.2 El Entrenamiento

En el proceso de entrenamiento se construyeron los vectores patrón para cada una de las clases. Para ello, se utilizó un sistema binario (1 para la ocurrencia de un término *i*, 0 para la no ocurrencia), así como otro de pesos, calculados a partir de las propuestas de Salton [7], de manera que el peso del término *t* en el documento *d* se obtiene en base a la fórmula

$$\frac{(ftd * \log(N/n))}{\left[\sum_{i=1}^{nt} ftd_i * \log/(N^2/nd_i)\right]^{1/2}} \quad (2)$$

donde

ftd es la frecuencia del término *t* en el documento *d*

nt es el número de términos en el documento *d*

N es el número de documentos en toda la colección

n es el número de documentos en que aparece el término *t*

nd_i es el número de documentos en que aparece el término *i*

Se formaron, pues, dos vectores para cada documento de la muestra de entrenamiento, uno binario, y otro con pesos. A continuación, para cada una de las nueve clases contempladas, se construyeron dos vectores patrón (binario y con pesos), a partir del algoritmo de Rocchio, obteniendo el peso de cada término de dichos vectores patrón mediante la fórmula comentada más arriba, teniendo en cuenta que, en este caso, no existe consulta inicial, con lo que C_0 fue puesto a 0.

De otro lado, y teniendo en cuenta los problemas derivados de trabajar con cantidades negativas, se modificaron los vectores de entrada, de forma que aquellos pesos con valor negativo fueron puestos también a 0.

Las constantes β y γ, por su parte, y siguiendo las recomendaciones de Buckley et al. [19] fueron puestos a 16 y 4, respectivamente.

3.3 El Test de Categorización

Para probar el sistema se empleó una colección de 4.147 noticias del mismo periódico y, aproximadamente, de las mismas fechas. Las características generales de estos documentos son similares a las de los utilizados para el entrenamiento. Todas las noticias que se intentaron categorizar, por otra parte, pertenecían a alguna de las secciones o clases contempladas en la fase de entrenamiento.

Para estimar el grado de similaridad entre los documentos a categorizar y los patrones de cada una de las clases se empleó el coeficiente del coseno, ampliamente aplicado en operaciones de Recuperación de Información [5, 7]:

$$SIM(P_x, D_y) = \frac{\sum_{i=1}^{n} p_{xi} d_{yi}}{\sqrt{\sum_{i=1}^{n} p_{xi}^2 \cdot \sum_{i=1}^{n} d_{yi}^2}} \quad (3)$$

P_x es el vector patrón de la clase *x*

D_y es el vector del documento *y*

p_{xi} es el elemento *i* de P_x

d_{yi} es el elemento *i* de D_y

n es el número de elementos o términos en los vectores

De la confrontación entre los documentos a categorizar y los patrones de las clases contempladas se obtuvieron (por cada una de las modalidades, binaria y con pesos) nueve coeficientes de similaridad en cada documento, uno por cada clase contemplada. En una situación de trabajo manual, estos coeficientes podrían presentarse al usuario en orden decreciente, de manera que éste pudiera, manualmente, determinar la clase o clases más adecuadas.

Si se opera de forma totalmente automática, sin embargo, es preciso definir un umbral en los citados coeficientes, de manera que los de las clases situadas por encima del umbral indicarían en qué categorías puede ubicarse el documento a categorizar. El establecimiento de ese umbral debe efectuarse de manera experimental, tendiendo a optimizar los resultados [10, 21].

Sin embargo, el uso de umbrales presupone que un documento puede ser adscrito a más de una clase. En algunas situaciones reales, y debido a condicionantes o restricciones externas, puede ser preciso elegir una sola clase, no varias. De hecho, los documentos utilizados en nuestro experimento pertenecen en la realidad a una sola sección del periódico. Con esta restricción, debería optarse por aquella clase cuyo coeficiente de similaridad fuese más alto.

3.4 Evaluación

Tradicionalmente, la efectividad de las operaciones en IR se miden utilizando las medidas clásicas de Precisión y Exhaustividad [7]. Esta práctica se ha seguido también en trabajos de categorización, aunque, en algunos casos, se ha optado por presentar los resultados en términos de porcentajes de aciertos y fallos. Por nuestra parte, y a efectos de posibles comparaciones, hemos preferido utilizar Precisión y Exhaustividad, calculándolas de acuerdo a las siguientes fórmulas [10]:

$$R = \frac{a}{a+c} \tag{4}$$

$$P = \frac{a}{a+b} \tag{5}$$

donde:

R es la exhaustividad

P es la precisión

a es el número de documentos pertenecientes a una clase y adscritos a esa clase

b es el número de documentos no pertenecientes a una clase pero asignados a esa clase

c es el número de documentos pertenecientes a una clase no asignados a esa clase

Naturalmente, se trata de evaluar resultados para cada una de las clases. Adicionalmente, se han propuesto algunas medidas que unifican en un único resultado precisión y exhaustividad. Una de ellas es la medida F_β [4]:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{6}$$

β es un parámetro que permite ajustar la influencia relativa de ambos componentes, precisión y exhaustividad. $\beta = 1$ proporciona igual peso a ambos componentes de la medida. F_β ha sido utilizada con frecuencia en trabajos de categorización [22].

3.5 Resultados

La Tabla 2 recoge los resultados obtenidos para cada clase, considerando sólo la posibilidad de asignación de una única clase para documento a categorizar, obviamente la de mayor similaridad. Los resultados son sensiblemente mejores con la utilización de pesos que con vectores binarios, cosa que cabría esperar.

Para algunas clases, los resultados son francamente buenos (por ejemplo, DEPORTES, con un F_1 de 0,91), pero incluso los resultados medios pueden considerarse interesantes.

4 Conclusiones

Se han mostrado las posibilidades de elaborar vectores patrón que recojan las características de distintas clases o categorías de documentos, utilizando técnicas basadas en aquéllas aplicadas en la expansión de consultas por relevancia. Al mismo tiempo, se ha descrito un experimento consistente en la aplicación de esas técnicas a una colección de noticias de prensa en español, para su categorización. Los resultados obtenidos son, en conjunto, homologables o incluso mejores que los obtenidos en experimentos similares; para algunas de las categorías, estos resultados han sido muy favorables.

Referencias

1. FAIRTHORNE, R.A.(1961):“The mathematics of the classification”, *Towards Information Retrieval*, Butterwoths, London (1961), 1-10
2. HAYES, R.M. (1963):“Mathematical models in information retrieval”, *Natural Language and the Computers* (P,L, Garvin, Ed,), McGraw-Hill, N,Y, (1963), 287
3. SALTON, G. (1968): *Automatic Information Organization and Retrieval*, McGraw-Hill, N,Y, (1968)
4. RIJSBERGEN, K. VAN (1979): *Information Retrieval*, Butterworths, London, 1979,

5. HARMAN, D. (1992): Ranking Algorithms, en Frakes, W,B, & Baeza-Yates, R,: *Information retrieval: Data Structures and Algorithms*, Prentice-Hall, Englewood Cliffs (NJ), 1992, pp, 363-392
6. SALTON, G. & BUCKLEY, C. (1988): Term-Witghting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24(5), 1988, 513-523
7. SALTON, G. & MCGILL, M.J. (1983): *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983
8. HARMAN, D. (1992): Relevance feedback and other query modification techniques, en Frakes, W,B, & Baeza-Yates, R,: *Information retrieval: Data Structures and Algorithms*, Prentice-Hall, Englewood Cliffs (NJ), 1992, pp, 241-263
9. ROCCHIO, J.J. (1971): Relevance feedback in Information Retrieval, en Salton, G, de,: *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Engelwood Cliffs (NJ), 1971, pp, 313-323
10. LEWIS, D.D., SCHAPIRE, R.E., CALLAN, J.P. & R. PAPKA (1996): Training Algorithms for Linear Text Classifiers, *SIGIR 96*, 298-306
11. FIGUEROLA, C.G. (2000): La investigación sobre Recuperación de la Información en español, en Gonzalo García, E, y García Yebra, V, Eds,: *Documentación, Terminología y Traducción*, Síntesis, Madrid, 2000,
12. HARMAN, D. (Ed.): *3º Text Retrieval Conference (TREC-3)*, NIST Special Publicatin 500-225, Gaithersburg, 1995
13. HARMAN, D. (Ed.): *4º Text Retrieval Conference (TREC-4)*, NIST Special Publication 500-236, Gaithersburg, 1996
14. HARMAN, D. (Ed.): *5º Text Retrieval Conference (TREC-5)*, NIST Special Publication, 500-238, Gaithersburg, 1997
15. PETERS, CAROL (Ed.) (2000). *First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign.* Working Notes for the CLEF 2000 Workshop. Lisboa, Sept, 2000.
16. PAICE, C.D. (1996): "Method for Evaluation of Stemming Algorithms Based on Error Counting", *JASIS*, 47(8), 632-649
17. GÓMEZ DÍAZ, R. (1998): *La Recuperación de Información en español: evaluación del efecto de sus peculiaridades lingüísticas*, Universidad de Salamanca, trabajo de Grado(tesina), Salamanca, 1998
18. EFE, AGENCIA (1991): *Manual de español urgente*, Madrid, Cátedra, 1991, pp, 18-22 y 36-60
19. ELENA GARCIA, P. (1994): “La traducción de textos informativos (noticias)”, en *Curso práctico de traducción general alemán - español*, Salamanca, Ediciones Universidad de Salamanca, 1994
20. BUCKLEY, C., SALTON, G. & ALLAN, J. (1994):The effect of adding relevance information in a relevance feedback environment, *SIGIR 94*, pp, 292-300
21. COHEN, W.W. & SINGER, Y. (1996): Context-sensitive learning methods for text categorization, *SIGIR 96*, pp, 307-315
22. LEWIS, D.D. & GALE, W. (1994): A sequential algorithm for training texts classifiers, *SIGIR 94*, pp, 3-12

CLASES	*Entrenamiento*	*Test*
BOLSA	287	455
CAMPUS	292	464
CULTURA	301	452
DEPORTES	315	472
ECONOMIA	301	461
INTERNACIONAL	293	439
MOTOR	324	469
NACIONAL	338	471
SOCIEDAD	290	464
TOTAL docs.	2741	4147

Tabla 1. Número de documentos para entrenamiento y test, por secciones

	binario			pesos		
Clase	*Prec.*	*Exhaust.*	F_1	*Prec.*	*Exhaust.*	F_1
BOLSA	0,05	1	0,09	0,29	1	0,44
CAMPUS	0,21	1	0,35	0,28	1	0,43
CULTURA	0,90	0,63	0,74	0,89	0,83	0,86
DEPORTES	1	0,70	0,82	0,93	0,88	0,91
ECONOMIA	1	0,21	0,35	0,73	0,57	0,64
INTERNACIONAL	1	0,33	0,50	0,88	0,73	0,80
MOTOR	0,84	0,80	0,82	0,8	1	0,89
NACIONAL	0,68	0,49	0,57	0,83	0,75	0,80
SOCIEDAD	0,66	0,14	0,24	0,86	0,48	0,62

Tabla 2**, Resultados con asignación de clase única**

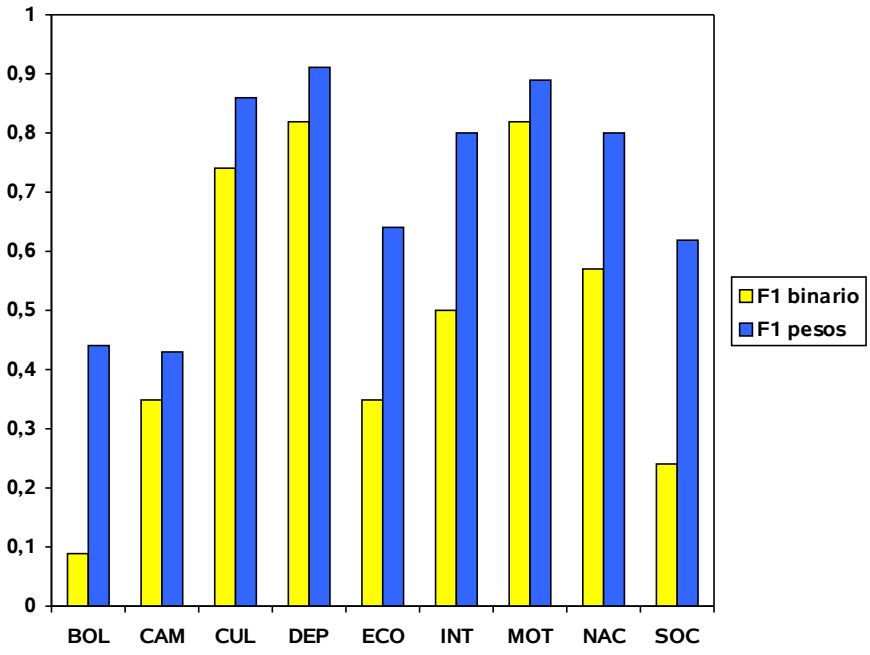

Figura 1**, Resultados (F_1) con asignación de clase única**