Reconocimiento Automático de Artículos Científicos

Pablo J. Lavallén, Fernando R.A. Bordignon, Gabriel H. Tolosa[1]
Universidad Nacional de Luján
{plavallen, bordi, tolosoft}@unlu.edu.ar

[1] Becario de Investigación. Secretaría de Investigación y Postgrado. Universidad Nacional de Luján

Resumen

En el presente trabajo se describe un modelo basado en un conjunto de reglas heurísticas que permite la detección automática de documentos de carácter científico a partir del análisis lógico de su estructura. En particular, se definen 4 categorías de reglas que se aplican en diferentes niveles de especificidad.

Se implementó un prototipo de software a los efectos de validar y ajustar el comportamiento del modelo. Se trabajó con un corpus de formatos heterogéneos de 600 documentos relacionados al área de las ciencias de la computación y la informática, donde el 50% fueron artículos y el resto textos varios sobre el dominio del conocimiento en cuestión. Se realizaron diferentes experimentos con la intención de probar las reglas y ajustar empíricamente el valor umbral. De los experimentos realizados se obtuvieron resultados, medidos en términos de precisión, que oscilan entre 0.85 y 0.94.

1 – Introducción

El área de reconocimiento de documentos supone un análisis estructural que se puede dividir funcionalmente en dos niveles [3]. En el nivel bajo ó físico la tarea de reconocimiento intenta determinar elementos como páginas y párrafos hasta identificar tipos de fuentes, caracteres y mapas de bits [1] [2]. En el nivel alto ó lógico, el objetivo es encontrar “partes” tales como resumen, introducción y conclusiones.

En este trabajo se presenta un método basado en análisis lógico de documentos con el objetivo de construir un módulo de reconocimiento de literatura científica, en particular, de artículos de investigación del área de las ciencias de la computación y la informática. Mediante la implementación de heurísticas se pretende realizar el proceso de manera eficiente empleando para ello un bajo costo computacional.

Esto permitiría generar herramientas que puedan ser utilizadas como módulos de sistemas más complejos destinados al reconocimiento, clasificación y recuperación de información. Si se piensa en sistemas que indexan automáticamente determinados tipos de documentos, la capacidad de realizar el proceso de forma automática puede ser una ventaja importante – dada la complejidad de la tarea – respecto de sistemas donde se requiere una mínima intervención humana. Por ejemplo, podría emplearse como filtro en línea de un motor de búsqueda especializado en literatura científica, similar a CiteSeer.

2 – Propuesta

El objetivo de este trabajo es presentar un modelo que sirva para la clasificación automática de documentos de la literatura científica referida a temas relacionados con la informática y ciencias de la computación.

Desde el punto de vista de las estructuras lógicas, se pueden distinguir – por ejemplo –

título, autor/es o introducción. Además, tales estructuras deben guardar cierta relación entre sí, ya sea de ubicación dentro del documento o de precedencia entre si.

El modelo que se presenta se basa en la utilización de un conjunto de reglas heurísticas mediante las cuales se realiza un análisis de la estructura lógica del documento. A partir de éste, se trata de identificar dentro del documento la aparición de ciertos "*objetos lógicos*" – por ejemplo *abstract* o *introduction*) como así también la relación entre los mismos y se establece un sistema de puntuación para determinar si el documento en cuestión puede ser considerado o clasificado como artículo científico.

3 – Definición de Reglas

En esta sección se define el conjunto de reglas del modelo, las cuales se dividen en cuatro categorías que plantean características claramente definidas. Cabe destacar que las mismas se encuentran en idioma inglés ya que el corpus utilizado para el análisis se encuentra en dicho idioma. Las reglas utilizadas se agrupan de la siguiente manera:

1. Reglas de aparición (RA): Corresponden a las reglas que contienen los términos a buscar dentro del documento junto con una serie de condiciones extra. Mediante estas reglas se intentan determinar los límites de las secciones. Por ejemplo, se busca la palabra *Introduction* al comienzo de una línea para determinar la sección introducción del artículo.

2. Reglas de ubicación (RU): Permiten definir un intervalo que indica la porción del documento donde debe aparecer cada término de una RA. Siguiendo el ejemplo anterior, se puede especificar que la palabra *introduction* se encuentre en el primer cuarto del documento.

3. Reglas de precedencia (RP): Definen la relación de precedencia entre los términos hallados mediante la aplicación de RA y RU. Además, permiten especificar una distancia mínima – medida en líneas – a la que deben encontrase.

4. Reglas de exclusión (RE): Indican cuáles términos se consideran mutuamente excluyentes, es decir, términos que no pueden puntuar individualmente en un mismo documento. Estas reglas impiden que se puntúe doblemente cuando aparecen – por ejemplo – términos como *references* y *bibliography* ya que esta situación no es común en los artículos científicos.

4 – Experimentos y evaluación

Para realizar la evaluación inicial del modelo propuesto se realizaron una serie de experimentos de clasificación donde el software debía determinar cuáles documentos correspondían a artículos científicos y cuáles no.

4.1 – Juego de pruebas

El corpus utilizado para las pruebas estaba compuesto por 600 documentos relacionados con el área Informática y las Ciencias de la Computación. Del total, 300 documentos corresponden a artículos científicos descargados de: 200 de CiteSeer en formato Postcript y 100 de la Tenth International World Wide Web Conference, en formato html.

De los 300 documentos que no son artículos científicos 100 correspondieron a documentos que presentan una estructura muy similar en cuanto a los objetos lógicos que en ellos se encuentran, como tesis y reportes técnicos. Éstos fueron seleccionados con la intención de chequear la capacidad de reconocimiento del software ante objetos de texto que ofrecen una posibilidad de confusión. El resto del corpus se completó con documentos de diversa naturaleza dentro de la temática: ayudas, libros, RFCs, memorandos, noticias, índices de libros y demás.

4.2 – Métricas

Las métricas de performance utilizadas para la evaluación corresponden a las medidas clásicas de la literatura: Precision y Recall [4] y su relación a través de la medida F [5]. Como se trata de una clasificación binaria se plantea una tabla de contingencia en cual constan todos los resultados posibles obtenidos por el clasificador:

		Correctos	
		Si	**No**
Asignados	**Si**	**a**	**b**
	No	**c**	**d**

Luego, se definen las relaciones entre los documentos científicos reconocidos correctamente y el total de documentos clasificados como científicos (P); y los documentos científicos reconocidos de manera correcta y el total de documentos científicos en el corpus (R), donde:

$$P = \frac{a}{a+b} \qquad R = \frac{a}{a+c} \qquad F = \frac{2PR}{P+R}$$

4.3 – Resultados iniciales

Se realizaron diferentes experimentos iniciales para validar el funcionamiento del modelo. Para éstos, se modificó la composición del corpus y el valor umbral de corte que determina el puntaje que debe sumar para considerase a un artículo como reconocido correctamente.

El primer experimento se realizó con el corpus completo (denominado c600), es decir, con los 600 documentos y se varió el umbral de corte entre 2 y 5. Los resultados se resumen en la siguiente tabla:

	AC		**NAC**		**C600**		
Umbral	**Bien**	**Mal**	**Bien**	**Mal**	**R**	**P**	**F**
2	284	16	250	50	0.95	0.85	0.90
3	278	22	256	44	0.93	0.86	0.89
4	249	51	284	16	0.83	0.94	0.88
5	233	67	285	15	0.78	0.94	0.85

AC – Artículos científicos, NAC – No Artículos científicos

En un segundo experimento se quitaron del corpus los 100 documentos que no son artículos científicos pero presentaban una estructura muy similar en cuanto a los objetos lógicos que en ellos se encuentran, reduciéndose el corpus a 500 (denominado c500). De manera similar al experimento anterior, se varió el umbral de corte entre 2 y 5, obteniéndose los siguientes resultados:

	AC		NAC	
Umbral	**Bien**	**Mal**	**Bien**	**Mal**
2	284	16	181	19
3	278	22	185	15
4	249	51	191	9
5	233	67	191	9

C500		
R	**P**	**F**
0.95	0.94	0.94
0.93	0.95	0.94
0.83	0.97	0.89
0.78	0.96	0.86

En el gráfico 1 se muestra la performance del sistema – para ambas configuraciones del experimento – en términos de la medida F.

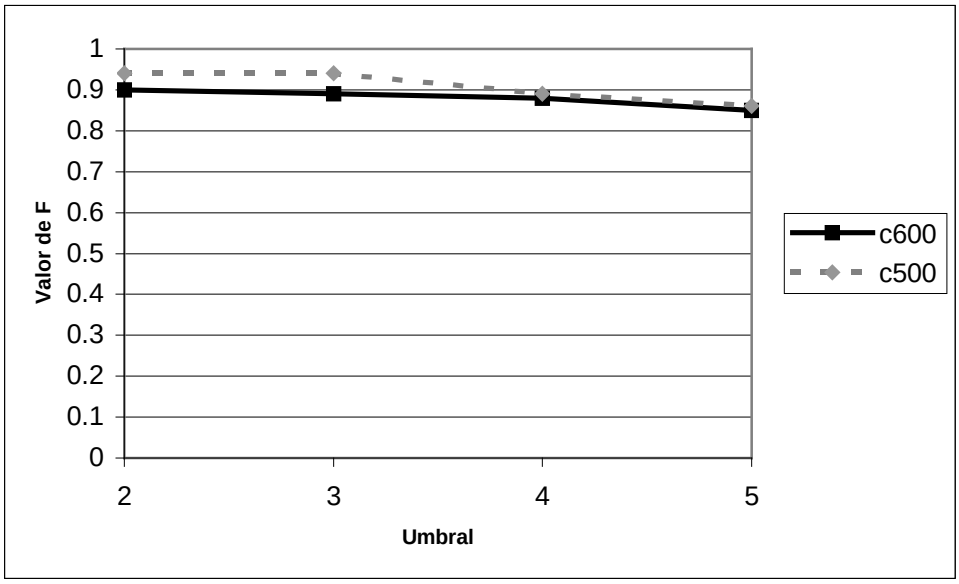

Gráfico 1 – Performance del sistema para ambos experimentos

En el gráfico 2 se presenta una comparación – en términos de porcentajes – sobre el total de los documentos. En éste, se puede apreciar la variación que determina el valor umbral sobre la cantidad de documentos correctamente clasificados.

Sobre los artículos científicos si se escoge un valor umbral bajo – 2 por ejemplo – aumenta la cantidad de documentos reconocidos como tales y a medida que se eleva el valor umbral disminuye el reconocimiento de los artículos científicos. Lo contrario ocurre con los documentos que no son artículos científicos, donde disminuye la cantidad reconocida correctamente a medida que disminuye el valor umbral elegido.

Por lo tanto – como se muestra en el gráfico 2 – un valor umbral entre 3 y 4 es el valor que mejor ajusta el rendimiento del software clasificador si el mismo actuará como filtro sobre un corpus conformado proporcionalmente por artículos científicos y documentos que no lo son. Un valor umbral bajo es conveniente sobre corpus donde predominen los artículos científicos y – caso contrario – será conveniente un valor umbral mayor.

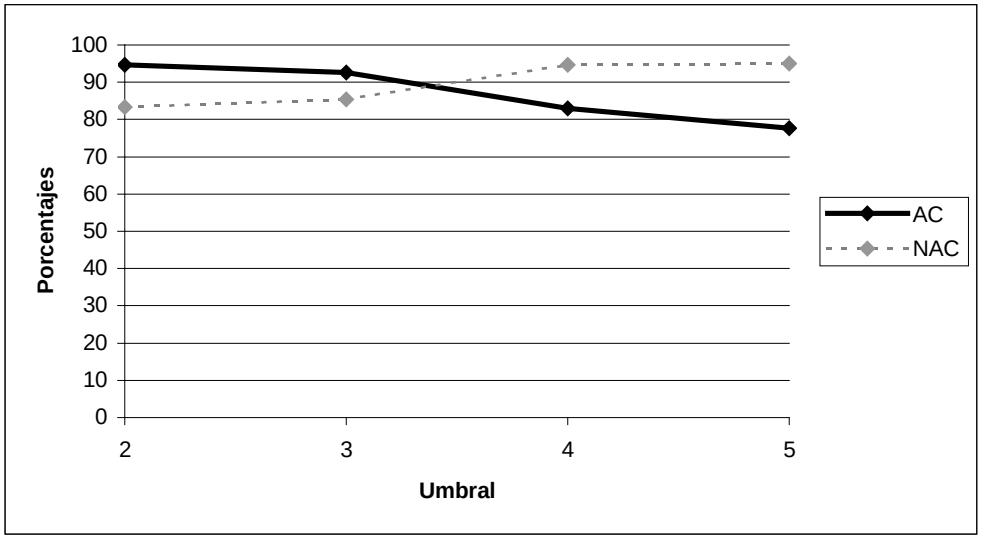

Gráfico 2 – Relación entre valor umbral y la capacidad del filtro

5 – Consideraciones

En este trabajo se presenta un modelo de desarrollo de un filtro destinado al reconocimiento automático de artículos científicos basado en la utilización de reglas heurísticas, sobre documentos del área de la Informática y las Ciencias de la Computación.

Se confeccionó un corpus heterogéneo de 600 documentos en distintos formatos. Los resultados, medidos en términos de precisión, oscilan entre 0.85 y 0.94, utilizando valores umbrales desde 2 a 5, respectivamente. Cabe señalar que los valores de corte – o umbrales – aquí utilizados son los considerados más significativos desde el punto de vista de la precisión que se pretende que brinde el software clasificador.

En el caso de utilizar el software clasificador como filtro automático – sin conocer la naturaleza de los documentos a analizar – se considera un valor umbral apropiado de 3 o 4 para balancear el rendimiento del software.

6 – Referencias

[1] Klink, S.; Dengel, A.; Kieninger, T. "*Document Structure Analysis Based on Layout andTextual Features*". En: Proceedings of Fourth IAPR International Workshop on Document Analysis Systems, DAS2000, pp 99-111. 2000.

[2] Mao, Sa.; Rosenfeld, Aa.; Kanungo, Tb. "*Document Structure Análysis Algorithms: A literature Survey*". En: Proceedings of SPIE Electronic Imaging, pp 197-207. 2003.

[3] Marovac, N. "*Document recognition Concepts and Implementations*". ACM SIGOIS Bulletin. Vol. 13, I.3, pp. 28-38. 1992.

[4] Salton, G. y McGill, M. *Introduction to Modern Informatin Retrieval*. McGraw-Hill, 1983.

[5] van Rijsbergen, C.J. *Information Retrieval*. 2da. edición, Butterworths, 1979.