Detección de Noticias del Ámbito Educativo Sobre Múltiples Canales Dinámicos de Información

Fernando R. A. Bordignon y Gabriel H. Tolosa

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes de Datos
{bordi, tolosoft}@unlu.edu.ar

Resumen

Se presentan resultados preliminares, obtenidos de las pruebas experimentales, de un nuevo método de detección de noticias del ámbito educativo sobre múltiples canales dinámicos de información. La técnica descripta podría ser aplicada a canales de sindicación de contenidos afín de detectar automáticamente noticias del dominio en cuestión y generar pseudocanales especializados.

De los primeros experimentos realizados se puede concluir que el método de detección de noticias en el ámbito educativo funciona con una buena perfomance asociada (0,883 F-Score) y por lo tanto puede ser utilizado en aplicaciones de filtrado automático sobre canales de noticias.

1 – Introducción

En la actualidad, los servicios de información para la obtención de noticias no se encuentran limitados a periódicos electrónicos, agencias o portales, sino que se han expandido a diferentes medios de publicación. Esta tendencia se sostiene sobre la base de la posibilidad de que existan diferentes medios de publicación orientados a cualquier usuario como blogs y wikis. Además, tales medios cuentan con la posibilidad de entregar contenido por solicitud utilizando protocolos de sindicación de contenidos.

En el área de Recuperación de Información (RI) tradicional, se reconoce como alternativas para la obtención de información la recuperación inmediata (búsquedas en *Seach Engines* y *Browsing*) y la recuperación diferida (filtrado o ruteo). En este último caso, el usuario especifica sus necesidades y el sistema entregará de forma continua los nuevos documentos que le lleguen y concuerden con ésta. Un ejemplo es el servicio denominado GoogleAlert[1].

Sin embargo, el modelo de intercambio basado en la sindicación se ha popularizado últimamente mediante protocolos como RSS, dada la posibilidad de recibir información semiestructurada utilizando XML. Bajo este esquema de trabajo, prácticamente cualquier proveedor de información puede publicar su "nuevo" contenido en su sitio y en su archivo XML para que aquellos que lo deseen lo descarguen y procesen. Una característica interesante de este modelo es que la información está disponible justo cuando es publicada y la frecuencia de consulta la decide el consumidor.

Por otro lado, la amplitud temática de ciertos servicios de información genera ciertas veces que un usuario se encuentre sobrecargado en cuanto a la cantidad de información que recibe

[1] http://www.googlealert.com/

inclusive por este medio. Además, las características estructurales de las noticias recibidas son heterogéneas y dependen del que genera la información. Es posible recibir noticias cortas, de unas pocas líneas e – inclusive – solo títulos. De aquí que surge la necesidad de desarrollar técnicas que permitan filtrarlas por una determinada temática, teniendo en cuenta dichas características.

En este artículo se presenta un modelo de detección de noticias que se basa en la clara suposición de que existe un vocabulario propio asociado a cada área, junto con un uso particular de los términos del lenguaje. De esta manera, es posible construir una descripción específica que intente capturar alguna de sus particularidades a partir de ciertos parámetros. Aquí, se presenta y evalúa una metodología que permite detectar noticias en español relacionadas con la educación provenientes de múltiples fuentes informativas, en particular que propagan sus contenidos por medio de la sindicación (por ejemplo, mediante el protocolo RSS).

Este trabajo es parte del proyecto de investigación "*Modelos y Servicios de Información sobre Sistemas Complejos en Ambientes Académicos y Científicos*" desarrollado en el Laboratorio de Redes[2] de la Universidad Nacional de Luján. En el mismo, se estudian diferentes espacios de información educativos con la finalidad de caracterizar, extraer y organizar la información para facilitar su acceso y utilización, a partir de la aplicación de técnicas de Recuperación de Información y Minería de Datos.

2 – Los Canales de Noticias

Una de las tecnologías que contribuye a la diseminación de información proveniente de diferentes fuentes está basada en el concepto de feed", utilizando protocolos derivados de XML. Uno de los casos más comunes es el protocolos RSS (Really Simple Syndication) [Richardson, 2005] [Hammersley, 2005] el cual permite poner a disposición de los usuarios pequeños textos que pueden ser extraídos por un software lector y presentados si necesidad de visitar un determinado sitio web u otra clase de repositorio. Esta modalidad de trabajar es opuesta a la idea original de publicar en un sitio web que los usuarios deban obligatoriamente visitar [Hammond, 2004], sino que provee un *snapshot* de la información del sitio en forma de resumen, con texto y enlaces. Aquí es importante la idea que un consumidor de información puede solicitar el resumen con la frecuencia que desea a los efectos de estar altamente actualizado.

En la actualidad existen múltiples fuentes de información que brindan resúmenes utilizando esta tecnología como por ejemplo los blogs, periódicos, agencias de noticias, wikis e inclusive páginas de empresas. Inclusive, existen algunos servicios de búsqueda como Feedster cuyas fuentes de información son cientos de miles de canales RSS. Otro ejemplo es Moreover.com, el cual almacena noticias de miles de fuentes, las categoriza y genera resúmenes también en RSS. Si bien resultan útiles, en muchos casos se requiere identificar unos pocos temas (o solo uno) de fuentes seleccionadas, en general, por el usuario final de acuerdo a parñametros propios.

3 – Detección de Noticias de Educación

Como se ha mencionado, se trata con la suposición de que existe un vocabulario propio relacionado con la educación, producto del uso frecuente de ciertos términos que caracterizan al dominio en estudio. A los efectos de obtener una lista de palabras que representa el vocabulario antedicho se

[2] http://www.tyr.unlu.edu.ar/investigacion.html

utilizó la técnica de *log-likelihood ratio* [Rayson 2000] [Rayson 2004], la cual provee un criterio que permite identificar el léxico propio del discurso educativo

El método de Rayson compara dos corpus, uno general – también llamado corpus de referencia de la lengua - y otro específico del dominio en estudio. A partir de analizar las frecuencias normalizadas de aquellos términos que están en el conjunto intersección de ambos corpus, para cada palabra se obtiene un valor de significancia (se excluyen las palabras vacías). Cuanto más alto sea el valor asociado a un término mayor será su peso o importancia en el vocabulario propio del dominio educativo.

En los experimentos realizados, como corpus de referencia de la lengua (corpus-español) se utilizó un conjunto de artículos de diversa temática extraídos del espacio web. El corpus de educación (corpus-educa) se construyó con capítulos de libros, noticias, artículos en línea y ensayos, entre otros documentos. En la tabla 1 se presentan las descripciones de cada uno.

	Corpus Educa	**Corpus Español**
Tokens total	685.667	6.414.694
Tokens distintos	41.104	188.496
Tamaño	8 Mb	31 Mb

Tabla 1 – Composición de las descripciones de los corpus

En primera instancia, se confeccionaron cinco listas (L1, ..., L5). con los primeros 100, 200, 300, 400 y 500 términos más representativos (valores de significancia más elevados) del vocabulario educativo. Luego se desarrolló una métrica que permita medir el contenido de lenguaje educativo de una noticia. Para ello se decidió medir la proporción de términos relativos al dominio en estudio. Dada una lista L_i y una noticia N_j la proporción se calcula como:

$$\text{Proporción}(N_j, L_i) = Q_{ij} / | N_j |$$

Donde:

Q_{ij} es la cantidad de términos de la lista i que hay en la noticia j (se cuentan las repeticiones)

$| N_j |$ es el total de términos en la noticia j (se excluyen las palabras vacías)

A continuación, es necesario determinar un umbral de corte para el valor de proporción que define que una noticia corresponde al dominio educativo o no. Este parámetro se determinó empíricamnete a partir de una serie de experimentos sobre un corpus de prueba.

4 – Experimentos y Resultados

A los efectos de validar la hipótesis de trabajo y ajustar los parámetros del modelo se diseñaron una serie de experimentos. Se dispone de una colección de prueba de noticias cortas donde cada una posee un título y un texto asociado al mismo. La colección tiene en total 5.085 noticias, donde 325 son de educación y 4.760 pertenecen a otros temas como deportes, espectáculos, arte, política y actualidad internacional. A cada noticia se le eliminaron las palabras vacías.

Para evaluar la eficiencia de la técnica propuesta se utilizaron las medidas clásicas del área de recuperación de información: *Precision* (P) y *Recall* (R), las cuales se calcularon como:

$$\text{Precision(umbral\_n)} = \frac{\text{cantidad de noticias de educación en conjunto resultados}}{\text{Cantidad de noticias totales en conjunto resultados}}$$

$$\text{Recall(umbral\_n)} = \frac{\text{cantidad de noticias de educación en conjunto resultados}}{\text{Cantidad de noticias de educación en el corpus}}$$

Luego, se utilizó la métrica F-Score o F-Measure la cual integra P y R (es su media armónica).

$$\text{F-Score(umbral\_n)} = 2 / ((1 / \text{precision(umbral\_n)}) + (1 / \text{recall(umbral\_n)}))$$

El primer experimento consistió en calcular los valores de proporción para cada noticia del corpus de prueba y obtener la métrica F-Score para valores de umbral comprendidos entre 0,06 y 0,25 (6% y 25% de *tokens* de cada documento). La tabla 2 muestra los resultados obtenidos de eficiencia para distintos valores de umbral y distintos largos de lista (en itálicas se indica el máximo valor de perfomance alcanzado) .

	Listas de palabras				
Umbral	**L1 (100 palabras)**	**L2 (200 palabras)**	**L3 (300 palabras)**	**L4 (400 palabras)**	**L5 (500 palabras)**
0,06	0,774	0,627	0,693	0,659	0,675
0,07	0,832	0,730	0,778	0,753	0,765
0,08	0,835	0,810	0,823	0,816	0,820
0,09	0,815	0,868	0,841	0,854	0,847
0,10	0,767	***0,883***	0,821	0,851	0,835
0,11	0,713	0,868	0,783	0,824	0,803
0,12	0,673	0,844	0,749	0,794	0,771
0,13	0,619	0,821	0,706	0,759	0,731
0,14	0,583	0,787	0,670	0,724	0,696
0,15	0,548	0,742	0,630	0,682	0,655
0,16	0,459	0,709	0,557	0,624	0,588
0,17	0,402	0,680	0,505	0,580	0,540
0,18	0,363	0,646	0,465	0,540	0,500
0,19	0,325	0,596	0,420	0,493	0,454
0,20	0,248	0,562	0,344	0,427	0,381
0,21	0,209	0,523	0,299	0,380	0,335
0,22	0,159	0,488	0,239	0,321	0,274
0,23	0,127	0,422	0,195	0,267	0,225
0,24	0,094	0,354	0,148	0,209	0,174
0,25	0,077	0,307	0,123	0,176	0,145

Tabla 2 – F-Score para distintos valores de umbral y de longitud de lista de palabras

Como se aprecia en la tabla 2, la máxima perfomance se obtiene utilizando una lista de 200 términos, lo cual arroja un valor de F-Score de 0,88 con una *Precision* de 0,91 y *Recall* de 0,86. También se observa que a medida que se eleva el valor de umbral, como es lógico, la *Precision* aumenta y el *Recall* se reduce. Los resultados anteriores implican que el filtro debería configurarse con una lista de 200 palabras y un umbral de corte de 0,1 a los efectos de lograr la máxima perfomance.

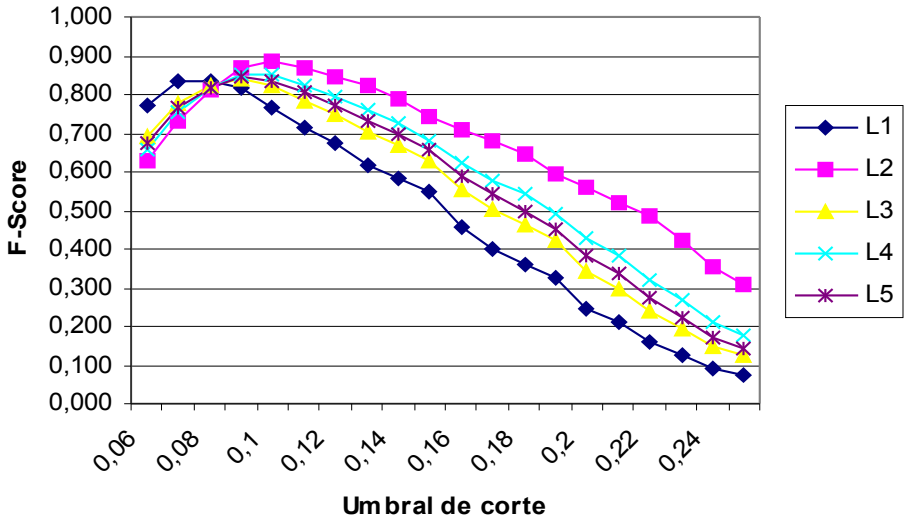

Gráfico 1 – F-Score para distintos valores de umbral y de longitud de lista de palabras

5 – Conclusiones y Trabajos Futuros

Se ha presentado un nuevo método destinado a la detección automática de noticias educativas sobre múltiples canales dinámicos de información. Su implementación es simple y requiere pocos recursos computacionales para operar. Se ha evaluado su perfomance bajo la métrica F-Score arrojando importantes valores de eficiencia (F-Score 0,88 con 0,91 en *Precision* y 0,86 en *Recall*).

En próximos experimentos se evaluará la eficiencia del método en función de dsitintos largos de noticias, esto se propone debido a que los canales de información donde opera (típicamente sobre sindicación de contenidos) suelen presentar algunas noticias de corto tamaño y esto podría afectar la perfomance. También se pretende evaluar posibles mejoras aplicando técnicas de preprocesamiento de textos, tales como *stemming* [Panessi, 2001] y uso de colocaciones en listas de palabras. Por otro lado, se considera extender el modelo a otros dominios del conocimiento para obtener filtros de propósito general que puedan ser utilizados en procesos de minería de datos como la detección automática de documentos de interés temático.

6 – Referencias

[Hammersley, 2005] Hammersley, B. Developing Feeds with RSS and Atom. O'Reilly, Cambridge, MA, 2005.

[Hammond, 2004] Hammond, T.; Hannay, T. and Lund B. The Role of RSS in Science Publishing. Syndication and Annotation on the Web. D-Lib Magazine. Vol. 10 No. 12. 2004.

[Panessi, 2001] Panessi, W. y Bordignon, F. Procesamiento de Variantes Morfológicas en Búsquedas de Textos en Castellano. Revista Interamericana de Bibliotecología. Vol 24, No. 1, pp 69-88. 2001.

[Rayson, 2000] Rayson, P. and Garside, R. Comparing Corpora Using Frequency Profiling. Proceedings of the Wworkshop on Comparing Corpora. Hong Kong, pp. 1-6. 2000.

[Rayson, 2004] Rayson P., Berridge D. and Francis B. Extending the Cochran rule for the comparison of word frequencies between corpora. Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Belgium, Presses universitaires de Louvain, pp. 926-936. 2004.

[Richardson, 2005] Richardson, W. The ABCs of RSS. Technology & Learning, 25, Issue 10, pp. 20-24. 2005.