Procesamiento del Lenguaje Natural, nº39 (2007), pp. 63-70 recibido 22-05-2007; aceptado 22-06-2007

Integración de conocimiento en un dominio específico para categorización multietiqueta

María Teresa Martín Valdivia
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
maite@ujaen.es

Arturo Montejo Ráez
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
amontejo@ujaen.es

Manuel Carlos Díaz Galiano
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
mcdiaz@ujaen.es

L. Alfonso Ureña López
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
laurena@ujaen.es

Resumen: En este artículo se presenta un estudio sobre el uso e integración de una ontología en un corpus biomédico. Nuestro objetivo es comprobar cómo afectan distintas maneras de enriquecimiento e integración de conocimiento sobre un corpus de dominio específico cuando se aplica sobre un sistema de categorización de textos multietiqueta. Se han realizado varios experimentos con distintos tipos de expansión y con diferentes algoritmos de aprendizaje. Los resultados obtenidos muestran una mejora en los experimentos que realizan expansión sobre todo en los casos en los que se utiliza el algoritmo SVM.

Palabras clave: Ontología MeSH, corpus biomédico (CCHMC), categorización multietiqueta, integración de conocimiento, aprendizaje automático

Abstract: In this paper, we present a study on the integration of a given ontology in a biomedical corpus. Our aim is to verify the effect of several approaches for textual enrichment and knowledge integration on a domain-specific corpus when dealing with multi-label text categorization. The different reported experiments vary the expansion strategy used and the set of learning algorithms considered. Our results show that for SVM algorithm the expansion performed produces best results in any case.

Keywords: MeSH ontology, biomedical corpus (CCHMC), multi-label text categorization, knowledge integration, machine learning.

1 Introducción

Las técnicas de procesamiento de lenguaje natural se están aplicando cada vez con mayor eficiencia en el dominio biomédico. Muchas investigaciones recientes exploran el uso de técnicas de procesamiento de lenguaje natural aplicadas al dominio biomédico (Karamanis 2007, Müller et al 2006). La necesidad de etiquetar y categorizar automáticamente textos médicos se hace cada vez más evidente.

Es innegable la importancia en la investigación y desarrollo de sistemas de búsqueda y recuperación de información en el dominio de la biomedicina que faciliten la tareas de los especialistas dando soporte y ayuda en su trabajo diario.

En este trabajo se presenta un estudio sobre la influencia en un sistema de categorización de una ontología específica del dominio biomédico: la ontología MeSH (MeSH 2007). Concretamente, se ha utilizado dicha ontología para expandir los términos de un documento que se quiere categorizar con el fin de mejorar los resultados sobre un sistema categorizador multi-etiqueta. Pensamos que la incorporación de conocimiento mediante la integración de recursos tales como las ontologías puede

ISSN: 1135-5948

mejorar significativamente los resultados obtenidos con los sistemas de información.

Por otra parte, para llevar a cabo la experimentación se han utilizado distintas configuraciones tanto de algoritmos de aprendizaje automático utilizados como de parámetros para cada uno de ellos. Concretamente, se ha utilizado el algoritmo SVM (Support Vector Machine), una red neuronal tipo perceptrón denominada PLAUM y el algoritmo de regresión bayesiana BBR. Los experimentos muestran que el uso de SVM mejora los resultados prácticamente en todos los casos.

El artículo se organiza de la siguiente manera: en primer lugar, se describe brevemente la tarea de categorización de textos multietiquetados así como el sistema categorizador utilizado TECAT. A continuación, se presentan los dos recursos biomédicos integrados (el corpus CCHMC y la ontología MeSH). En la siguiente sección se muestran los experimentos y resultados obtenidos. Finalmente, se comentan las conclusiones y trabajos futuros.

2 Categorización multietiqueta

La asignación automática de palabras clave a los documentos abre nuevas posibilidades en la exploración documental (Montejo, 2004), y su interés ha despertado a la comunidad científica en la propuesta de soluciones. La disciplina de recuperación de información, junto con las técnicas de procesamiento del lenguaje natural y los algoritmos de aprendizaje automático son el substrato de donde emergen las áreas de Categorización Automática de Textos (Sebastiani, 2002). En esta última área de investigación es donde se enmarca el presente trabajo y donde vierte sus principales aportaciones.

En la clasificación de documentos se distinguen tres casos:

1. Clasificación binaria. El clasificador debe devolver una de entre dos posibles categorías, o bien una respuesta SI/NO. Estos son los sistemas más simples, y al mismo tiempo los sistemas más conocidos en Aprendizaje Automático.
2. Clasificación multi-clase. En este caso el clasificador debe proporcionar una categoría de entre varias propuestas. Este sistema puede basarse en el anterior.
3. Clasificación multi-etiquetado. El documento se etiqueta no con una única clase, como en el caso anterior, sino que puede tomar varias de entre las categorías disponibles. Es el problema más complejo, pero puede simplificarse si utilizamos clasificadores binarios cuya repuesta pueda combinarse (por ejemplo, mediante un ranking de clases) o entrenando sobre cada clase un clasificador binario de repuesta SI/NO (como el sistema que se describe en este trabajo).

Hemos utilizado el software TECAT[1], que implementa un algoritmo para la clasificación multi-etiqueta basado en clasificadores base binarios. El algoritmo usado se muestra a continuación (*Algoritmo 1*), y consiste en entrenar un clasificador binario para cada clase seleccionando aquel que mejor rendimiento aporta dada una medida de rendimiento sobre el que se evalúa al clasificador. Además, aquellas clases para las que no es posible entrenar un clasificador con un rendimiento mínimo se descarta.

Entrada:
- un conjunto D_t de documentos multi-etiquetados para entrenamiento
- un conjunto D_v de documentos de validación
- un umbral α sobre la una medida de evaluación determinada
- un conjunto L de posibles etiquetas (clases)
- un conjunto C de clasificadores binarios candidatos

Salida:
- un conjunto C' = {c1, ..., ck, ..., c|L|} de clasificadores binarios entrenados

Pseudo-código:

```
C' ← ø
Para-cada l_i en L:
   T ← ø
   Para-cada c_j en C:
      entrena(c_j, l_i, D_t)
   T ← T ∪ {c_j}
  Fin-para-cada
   $c_mejor ← mejor(T, D_v)
  Si evalua(c_mejor) > α
   C' ← C' ∪ {c_mejor}
  Fin-si
  Fin-para-cada
```

Algoritmo 1. Entrenamiento de clasificadores base

[1] Disponible en http://sinai.ujaen.es/wiki/index.php/TeCat

3 Recursos utilizados

Nuestro objetivo principal consiste en estudiar la influencia que tiene el uso de una ontología médica sobre un corpus biomédico cuando se desea desarrollar un sistema automático de categorización de textos multi-etiquetados. Para ello, hemos utilizado dos recursos que describimos a continuación.

3.1 Corpus CCHMC

Se trata de un corpus desarrollado por "The Computational Medicine Center"[2]. Dicho corpus incluye registros médicos anónimos recopilados en el departamento de radiología del Hospital infantil de Cincinnati (the Cincinnati Children's Hospital Medical Center's Department of Radiology – CCHMC) (CMC, 2007).

La colección está formada por 978 documentos consistentes en informes radiológicos que están etiquetados con códigos del ICD-9-CM[3] (Internacional Classification of Diseases 9th Revision Clinical Modification). Se trata de un catálogo de enfermedades codificadas con un número de 3 a 5 dígitos con un punto decimal después del tercer dígito. Los códigos ICD-9-CM están organizados de manera jerárquica en los que se agrupan varios códigos consecutivos en los niveles superiores.

El número de códigos asignados a cada documento varía de 1 a 7. La Tabla 1 muestra la distribución del número de etiquetas por documento. El total de etiquetas distintas utilizadas en la colección es 142.

Clases	Documentos
1	389
2	368
3	162
4	46
5	12
7	1

Tabla 1. Número de clases asignadas por documento

La Figura 1 muestra un ejemplo de documento. Como se puede observar, la cantidad de información suministrada en cada documento es muy escasa pero muy relevante y bien estructurada. La colección se encuentra anotada manualmente por tres expertos. Por lo tanto, en cada documento existen tres conjuntos de anotaciones, una por cada uno de los expertos. Adicionalmente, se ha añadido un conjunto de etiquetas que unifica la mayoría de los tres expertos. Por otra parte, cada informe contiene dos partes de texto fundamentales: la historia clínica y la impresión o diagnóstico del médico.

3.2 Ontología MeSH

La ontología MeSH[4] (Medical Subject Headings) está desarrollada y mantenida por la National Library of Medicine y se utiliza como herramienta de indexación y búsqueda en temas

```
<doc id="97636670" type="RADIOLOGY_REPORT">
  <codes>
    <code origin="CMC_MAJORITY" type="ICD-9-CM">786.2</code>
    <code origin="COMPANY3" type="ICD-9-CM">786.2</code>
    <code origin="COMPANY1" type="ICD-9-CM">204.0</code>
    <code origin="COMPANY1" type="ICD-9-CM">786.2</code>
    <code origin="COMPANY1" type="ICD-9-CM">V42.81</code>
    <code origin="COMPANY2" type="ICD-9-CM">204.00</code>
    <code origin="COMPANY2" type="ICD-9-CM">786.2</code>
  </codes>
  <texts>
    <text origin="CCHMC_RADIOLOGY" type="CLINICAL_HISTORY">
      Eleven year old with ALL, bone marrow transplant on Jan. 2, now with
three day history of cough.</text>
    <text origin="CCHMC_RADIOLOGY" type="IMPRESSION">
      1. No focal pneumonia. Likely chronic changes at the left lung base.
2. Mild anterior wedging of the thoracic vertebral bodies.</text>
  </texts>
</doc>
```

Figura 1. Ejemplo de documento de la colección CCHMC

[2] http://www.computationalmedicine.org/

[3] http://www.cdc.gov/nchs/icd9.htm

relacionados con la medicina y la salud. Consiste en un conjunto de unos 23.000 términos denominados descriptores que se encuentran distribuidos de manera jerárquica permitiendo la búsqueda a varios niveles de especificidad. Un descriptor puede aparecer en varias ramas.

Existen varios estudios que demuestran que el uso y la integración de información procedente de ontologías y recursos con un vocabulario controlado, puede mejorar significativamente los sistemas de tratamiento de información (Chevallet, Lim y Radhouani, 2006, Guyot, Radhouani, y Falquet, 2005, Navigli, Velardi y Gangemi, 2003). Concretamente, nosotros haremos uso de la ontología MeSH con el fin de expandir los documentos del corpus CCHMC que se desean categorizar. De esta manera, se pretende incorporar conocimiento a la colección utilizada con el fin de mejorar los resultados en un sistema de categorización multietiqueta.

4 Descripción de los experimentos

4.1 Expansión con MeSH

Debido a que la cantidad de información en cada documento de la colección es escasa, se ha utilizado la ontología MeSH para expandir, con información médica dichos documentos. Se pretende incorporar información de calidad que ayude a mejorar la categorización de documentos.

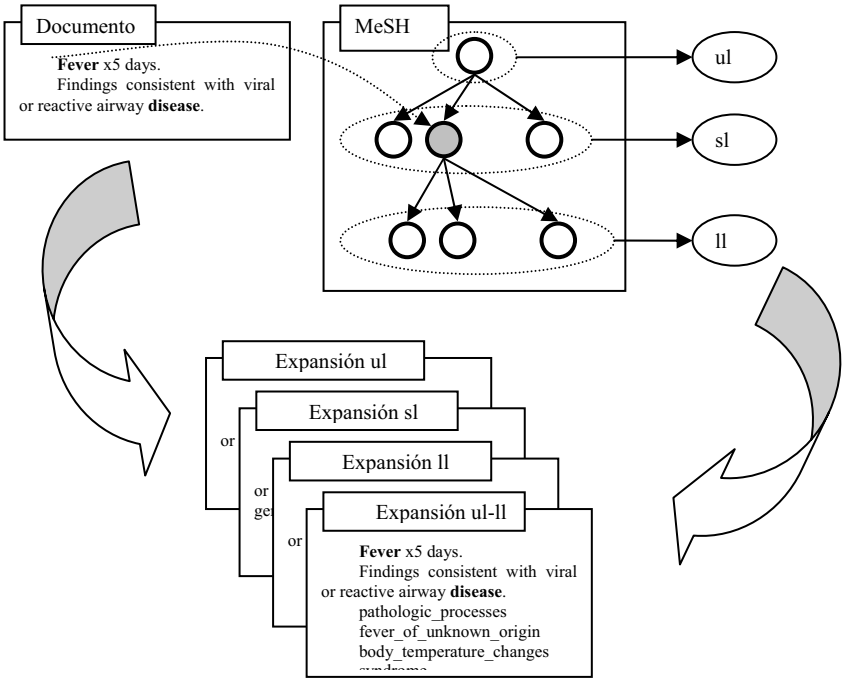

Figura 2. Estrategias de expansión con MeSH

Sin embargo, el uso indiscriminado de todos los términos extraídos de la ontología pueden empeorar los resultados puesto que incorporarían demasiado ruido. Así se pone de manifiesto por ejemplo en (Chevallet, Lim y Radhouani, 2006) donde se demuestra que seleccionar aquellas categorías de MeSH más acordes a la temática de los documentos, mejora la calidad de la expansión.

Con el fin de limitar el número de términos expandidos, se ha filtrado el número de categorías utilizadas para realizar la expansión. Así, aunque el primer nivel de MeSH incluye 16 categorías generales, se han seleccionado solo las siguientes tres:

- A: Anatomy
- C: Diseases
- E: Analytical, Diagnostic, and Therapeutic Techniques and Equipment

El motivo para elegir precisamente estas tres categorías es que el corpus incluye casos clínicos de niños con enfermedades relacionadas con el aparato respiratorio por lo

[4] http://www.nlm.nih.gov/mesh/

que dichas categorías deberían incluir la mayoría de los términos usados en el corpus.

Al realizar la expansión se busca el primer nodo de la ontología que coincide con la palabra a expandir. Una vez encontrado el nodo, la selección de términos que formarán parte de la selección se puede realizar de tres maneras distintas (ver Figura 2):

- Upper level (ul): se selecciona el término que está en un nivel superior a dicho nodo, es decir, el nodo padre.
- Same level (sl): se selecciona los términos que están al mismo nivel que dicho nodo, es decir, los nodos hermanos.
- Lower level (ll): se seleccionan los términos inmediatamente inferiores de dicho nodo, es decir, los nodos hijos.

Las palabras existentes dentro de los nodos seleccionados para formar parte de la expansión, han sido consideradas como entidades. Por lo tanto, si un nodo contiene una multipalabra (varias palabras separadas por espacios), dichas palabras se han incluido en la expansión formando un único término.

Con el fin de realizar un estudio para comprobar el comportamiento del sistema con varios tipos de expansión, se han diseñado distintas combinaciones con las tres expansiones anteriores. De esta forma, se han generado expansiones del tipo: ul+sl, ul+ll, ul+sl+ll… En la primera columna de la tabla 3 se pueden ver todas las expansiones realizadas.

4.2 Configuraciones de TECAT

Una vez realizada la expansión, cada experimento se ha realizado ajustando los distintos parámetros de TECAT:

- Se han eliminado las *palabras vacías* (*stop-words*).
- Se han obtenido las raíces de las palabras usando el *stemmer* de Porter (Porter 1980).
- Se han filtrado las características así obtenidas mediante *ganancia de información* (Shannon 1948), limitándonos a considerar 50,000 características.
- Se ha usado un pesado según el esquema TD.IDF.
- Se ha normalizado usando la función coseno.

Debido a que TECAT nos permite aplicar varios algoritmos al mismo tiempo, hemos estudiado las configuraciones siguientes:

- *SVM-multi* indica que se han pasado a TECAT varias configuraciones simultáneas del algoritmo SVM (Joachims, T., 1998). Estas configuraciones son aquellas que dan un peso adicional a los ejemplos positivos (normalmente escasos) con los valores 1, 2, 5, 10 y 20, es decir, 5 configuraciones diferentes de SVM que TECAT usará como clasificadores base independientes.
- *PLAUM-multi* indica, también, varias configuraciones para el perceptrón PLAUM (Y. Li et al., 2002) con pesados para ejemplos positivos en {0, 1, 10, 100} y pesados para negativos en {-10, -1, 0, 1}. Esto implica pasar a TECAT 16 configuraciones diferentes de PLAUM simultáneamente.
- *BBR-multi*. De forma similar a los anteriores, aquí el algoritmo BBR (A. Genkin et al., 2006) ha sido parametrizado con valores de umbral {0, 1, 2, 3, 4, 5} y valores de utilidad {0, 1, 2, 3}, si bien no se han analizado las combinaciones de todos ellos, por lo que las configuraciones consideradas han sido 10 para este algoritmo.

Las configuraciones en las que intervienen varias algoritmos combinados han sido realizadas, bien usando la simple de cada uno de ellos, bien la combinación de las múltiples parametrizaciones comentadas en cada uno de estos algoritmos.

5 Evaluación

Para evaluar los resultados se han usado validación cruzada en 10 particiones. Es decir, se ha dividido la colección en 10 particiones diferentes. Se ha ido alternativamente tomando una partición para test y el resto para entrenamiento. Los resultados finales de evaluación se calculan haciendo el promedio de cada ejecución correspondiente a cada participación. De esta forma se reduce el efecto que la selección de un determinado grupo de documentos para entrenamiento o evaluación pudiera tener sobre el resultado final.

Con las respuestas de un sistema de clasificación automático, y disponiendo de las predicciones reales que un experto humano asignaría, podemos construir la siguiente tabla de contingencia:

	SI es correcto	NO es correcto
El sistema dice SI	A	B
El sistema dice NO	C	D

Tabla 2. Contingencias.

Las medidas consideradas son precisión (P), cobertura (R) y F1, siendo ésta última la que nos da una visión más completa del comportamiento del sistema. Estas medidas han sido obtenidas mediante *micro-averaging*, es decir, calculando los aciertos y fallos en cada clase de forma acumulativa y calculando los valores finales sobre dichos valores acumulados, tal y como se refleja en las ecuaciones siguientes a partir de las medidas correspondientes según la tabla de contingencia anterior:

$$P_\mu = \frac{\sum_{c \in C'} A_c}{\sum_{c \in C'} A_c + \sum_{c \in C'} B_c}$$

$$R_\mu = \frac{\sum_{c \in C'} A_c}{\sum_{c \in C'} A_c + \sum_{c \in C'} C_c}$$

$$F1_\mu = \frac{2 P_\mu R_\mu}{P_\mu + R_\mu}$$

Los resultados obtenidos se pueden observar en las tablas 3, 4, 5 y 6. Como se puede observar, la integración de la ontología MeSH mejora prácticamente en todos los casos excepto para el caso de PLAUM, si bien con el algoritmo SVM es con el que la mejora es mayor. De hecho, como se muestra en la tabla 3, con la configuración SVM-multi se obtienen los mejores resultados independientemente del tipo de expansión realizada.

Si observamos los resultados desde el punto de vista de la expansión de los documentos, el método con unos resultados más homogéneos es el que realiza la expansión con los nodos padre (ul). Con este tipo de expansión se obtienen términos más generales que pueden considerarse como puntos en común entre documentos.

En cuanto a los algoritmos de aprendizaje utilizados, se puede observar que la expansión funciona en todos los casos excepto con la red neuronal PLAUM cuyos resultados son mejores sin ningún tipo de expansión.

Tipo de Expansión	SVM simple	SVM-multi
ll	0,724912	0,7675
ul	**0,739461**	**0,7957**
sl	0,734283	0,7697
ul-ll	0,739327	0,7766
ul-sl	0,726128	0,7669
ul-sl-ll	0,713533	0,7557
Sin expansión	0,737024	0,7699

Tabla 3. Expansión con SVM

Tipo de Expansión	BBR simple	BBR multi
ll	0,7290	0,732330
ul	0,7267	0,734653
sl	0,7400	0,737367
ul-ll	0,7314	**0,744386**
ul-sl	**0,74462**	0,735738
ul-sl-ll	0,7253	0,737014
Sin expansión	0,7250	0,724841

Tabla 4. Expansión con BBR

Tipo de Expansión	PLAUM simple	PLAUM multi
ll	**0,7284**	0,7228
ul	0,7233	**0,7372**
sl	0,7163	0,7262
ul-ll	0,7230	0,7263
ul-sl	0,7213	0,7210
ul-sl-ll	0,7177	0,7206
Sin expansión	0,7323	0,7311

Tabla 5. Expansión con PLAUM

Tipo de Expansión	SVM-BBR-PLAUM simple	SVM-BBR-PLAUM multi
ll	0,7562	0,7490
ul	**0,7704**	**0,7814**
sl	0,7642	0,7633
ul-ll	0,7611	0,7757
ul-sl	0,7513	0,7719
ul-sl-ll	0,7569	0,7479
Sin expansión	0,7478	0,7682

Tabla 6. Expansión combinando los tres algoritmos utilizados

6 Conclusiones y trabajos futuros.

En este trabajo se ha presentado un estudio en categorización multietiqueta enriqueciendo e integrado conocimiento. Para ello, se expande el corpus utilizado (CCHMC) en el proceso de categorización multietiqueta, con la ontología médica MeSH.

Para realizar el estudio se ha utilizado un categorizador multi-etiqueta TECAT disponible libremente y que permite la configuración y utilización simultánea de varios algoritmos de aprendizaje. Nuestro trabajo utiliza SVM, PLAUM y BBR además de una combinación de ellos. Los resultados muestran la conveniencia de integrar conocimiento externo proceden de una ontología específica biomédica. Sin embargo, las diferencias entre los distintos tipos de algoritmos utilizados no son excesivamente significativas.

En el futuro se pretende estudiar el uso de otros tipos de expansión utilizando dicha ontología, como por ejemplo la selección automática de las categoría que se utilizan para expandir, o el uso de sinónimos y palabras similares en lugar de nodos padres y/o hijos. Además se intentarán aplicar estas técnicas de expansión a otro tipo de tareas textual para comprobar el rendimiento de dicha técnica.

7 Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología a través del proyecto TIMOM (TIN2006-15265-C06-03).

Bibliografía

Chevallet, J. P., J. H. Lim y S. Radhouani. 2006. A Structured Visual Learning Approach Mixed with Ontology Dimensions for Medical Queries. Lecture Notes in Computer Science. Volume 4022/2006. Pages 642-651

CMC. 2007. The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. Disponible en http://www.computationalmedicine.org/challenge/cmcChallengeDetails.pdf

Genkin, A., D.D. Lewis and D. Madigan. 2006. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*

Guyot, J., Radhouani, S., y Falquet, G. 2005 Ontology-based multilingual information retrieval. In CLEF Workhop, Working Notes Multilingual Track, Vienna, Austria, 21–23. September 2005.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning, N. 1398, Springer Verlag, pp. 137-142.*

Karamanis, N. 2007. Text Mining for Biology and Biomedicine. Computational Linguistics. Volume 33. Pages 135-140.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor y J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. *Proceedings of the International Conference of Machine Learning (ICML'2002).*

MeSH. 2007. Medical Subject Headings. Accesible desde la página web: http://www.nlm.nih.gov/mesh/

Montejo-Ráez, A. y R. Steinberger. 2004. Why keywording matters. *High Energy Physics Libraries Webzine*. *Num. 10. Diciembre.*

Müller, H., T. Deselaers, T. Lehmann, P. Clough y W. Hersh. 2006. *Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks*. Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS 2006.

Navigli, R. Velardi, P. y Gangemi, A., 2003. Ontology learning and its application to automated terminology translation. Intelligent Systems, volume 18, issue 1, pp 22-31.

Porter, M. 1980. An Algorithm for Suffix Stripping. Program,Vol. 14 (3), pp. 130-137, 1980.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Survey, Vol. 34, Num. 1, pp. 1-47.*

Shannon, C. E. 1948.A mathematical theory of communication. *Bell System Technical Journal, vol. 27, pp. 379-423 y 623-656*.