Extracción automática de términos MeSH-DeCS en repositorios de ciencias de la salud: el caso de RUNA

Autores: Carmen Rodríguez Otero [1] , Emilio Lorenzo Gil [2], Adán Román Ruiz [2]
1. Bibliosaúde. Biblioteca Virtual do Sistema Sanitario Público de Galicia
2 Arvo Consultores y Tecnología

Introducción

Para administrar y mejorar las búsquedas en la literatura biomédica, la Biblioteca Nacional de Medicina de EE.UU. (NLM®) desarrolló el vocabulario controlado Medical Subject Heading (MeSH). La clasificación temática basada en vocabularios se ha identificado como uno de los factores principales en las estrategias de búsqueda y recuperación de documentos.

Desafortunadamente, dada su naturaleza especializada, la asignación manual de términos MeSH a artículos biomédicos es una tarea compleja, subjetiva y que requiere mucho tiempo, por lo que los sistemas de extracción automatizada de palabras clave (AKE) se convierten en soluciones evidentes para su incorporación a sistemas que necesitan describir y manejar miles de documentos, como son los repositorios.

En el trabajo se muestra la solución incorporada en el repositorio RUNA, repositorio institucional del Sistema Público de Salud de Galicia para facilitar la clasificación temática sobre vocabularios temáticos (MeSH-DeCS).

Se describe de forma específica el sistema de extracción automatica de términos de documentos y cómo se ha integrado dicha solución en el flujo de archivo de documentos en el repositorio para posibilitar el complemento por catalogadores expertos y así mejorar la calidad de la descripción temática efectuada.

Metodología

El sistema construido se integra en el flujo de autoarchivo de los documentos del repositorio, con el fin de unir las ventajas del procesamiento automático con la existencia de un experto que realice la selección de los términos efectivamente usados. En este sentido el subsistema extractor automatizado se visiona como un pre-tratamiento del documento que propone términos de clasificación, que deberán luego ser validados y rechazados por el usuario experto del repositorio.

En primer lugar el documento, normalmente en formato PDF o en formatos tipo WORD, etc.. , es convertido a formato simple textual (txt). Este paso del proceso no sirve únicamente para normalizar la entrada documental al sistema extractor sino que el fichero transformado es usado también por el indexador a texto completo del repositorio RUNA.

A partir de ese fichero „simple“ se realiza una primera selección de términos candidatos, con extracción de todas de las frases, palabras, términos y conceptos susceptibles de ser descriptores.

Sigue un proceso de puntuación y selección de términos. Todos los términos candidatos son puntuados combinando las propiedades de los términos (p.ej, su pertenencia al título del documento) con tecnicas de aprendizaje-máquina (machine learning techniques) para determinar la probabilidad de que un elemento sea un término clave. El sistema está configurado para proponer, a la finalización de este proceso un número determinado de términos (10 términos en RUNA). En la implementación