Procesamiento del Lenguaje Natural, núm. 36 (2006), pp. 61-62 recibido 01-03-2006; aceptado 01-06-2006

Clasificación Automática de Textos en el Dominio de la Física de Altas Energías

Arturo Montejo Ráez
Departamento de Informática
Universidad de Jaén
amontejo@ujaen.es

Resumen: Tesis doctoral en Informática realizada por Arturo Montejo Ráez bajo la dirección de los doctores L. Alfonso Ureña López (Univ. de Jaén) y Ralf Steinberger (Joint Research Centre, Comisión Europea). El acto de defensa de tesis tuvo lugar en marzo de 2006 en Granada ante el tribunal formado por los doctores Manuel Palomar Sanz (Univ. de Alicante), Ruslan Mitkov (Univ. de Wolverhampton, GB), Patricio Martínez Barco (Univ. de Alicante), Horacio Rodríguez Hontoria (Univ. Pol. de Catalunya) y Ramón López-Cózar Delgado (Univ. de Granada). La calificación obtenida fue Sobresaliente *Cum Laudem* por unanimidad.
Palabras clave: clasificación automática de documentos multi-etiquetados, aprendizaje automático, bibliotecas digitales.

Abstract: PhD thesis in Computer Science written by Arturo Montejo Ráez under the supervision of Dr. L. Alfonso Ureña López (Univ. of Jaén) and Dr. Ralf Steinberger (Joint Research Centre, European Commission). The author was examined in March 2006 in Granada by the commitee formed by Manuel Palomar Sanz (Univ. of Alicante), Ruslan Mitkov (Univ. of Wolverhampton, UK), Patricio Martínez Barco (Univ. of Alicante), Horacio Rodríguez Hontoria (Univ. Pol. of Catalunya) and Ramón López-Cózar Delgado (Univ. of Granada). The grade obtained was *Sobresaliente Cum Laudem*.
Keywords: multi-label text categorization, machine learning, digital libraries.

1. Introducción

Este trabajo constituye una propuesta de solución al problema del multi-etiquetado masivo de documentos en general, y el de documentos en el dominio de la Física de Altas Energías en particular. El resultado de esta investigación es una respuesta *real* y operativa a este problema. Dicho problema se identificó como un problema de *categorización de textos* (o *clasificación de textos*), en el que palabras clave predefinidas son consideradas categorías a ser asignadas a documentos en función del contenido semántico de los mismos. Durante el desarrollo de esta investigación, realizada principalmente en el Laboratorio Europeo para la Investigación Nuclear (CERN), la colección de documentos manejada desveló problemas no cubiertos con anterioridad por la literatura especializada. La necesidad expresa de una solución al manejo de datos de esta índole que debía ir más allá del mero análisis científico y del prototipado ha marcado la hipótesis planteada a lo largo de todo el trabajo.

La asignación automática de palabras clave a los documentos abre nuevas posibilidades en la exploración documental, y su interés ha despertado en la comunidad científica la búsqueda de soluciones. La disciplina de *Recuperación de Información* (RI), junto con las técnicas para el *Procesamiento del Lenguaje Natural* (PLN) y los algoritmos de *Aprendizaje Automático* (*Machine Learning*, ML) son el sustrato de donde emergen las tareas de *Categorización Automática de Textos*. Este último dominio de investigación es donde se enmarca el presente trabajo y es al mismo donde vierte sus principales aportaciones.

El trabajo se enfrenta ante tres tareas principales:

1. Estudiar la estructura y características de la colección de Documentos de Física de Altas Energías (*High Energy Physics, HEP)*. Se identifica de esta manera una colección de documentos multi-etiquetados con alto grado de desbalanceo, proponiendo una nueva métrica para su análisis global, el *grado de desbalanceo interno*.

2. Proponer una estrategia de clasificación

multi-etiquetado que pueda construirse a partir de algoritmos de aprendizaje automático conocidos como clasificadores base binarios (*Support Vector Machines, Rocchio, PLAUM,* etc.). Esta estrategia debe ser robusta ante clases con alto grado de desbalanceo y debe proporcionar una asignación en tiempo real. La estrategia de integración se fundamenta en la hipótesis que considera cada clase como un problema de clasificación aislado. Dicho sistema ha sido usado para resolver clasificación multi-etiquetado en otros dominios distinos a HEP, en concreto sobre una colección de documentos de corte político y legal etiquetados con el tesauro EUROVOC, usado por la Comisión Europea. Los resultados experimentales demostraron el buen rendimiento de esta estrategia fuera del dominio para el cual fue diseñada.

3. Validar la hipótesis de que la integración de información bibliográfica mejora los sistemas de clasificación. Para ello se han estudiados distintas fuentes de información y se ha realizado un análisis estadístico de los resultados obtenidos que garanticen con un 95 % de fiabilidad de que, efectivamente, la introducción de información bibliográfica es muy beneficiosa para la clasificación.

2. Estructura de la tesis

La estructura de la tesis sigue un discurso clásico, introduciendo el problema en su paradigma, los sistemas aparecidos hasta la fecha dando respuesta a la necesidad de un multi-etiquetador automático, el análisis teórico de la tarea en cuestión y una posterior propuesta de solución seguida por un extenso trabajo experimental que valide dicha propuesta.

En el capítulo 2 se ofrece una detallada introducción al marco en el cual se ha desarrollado la investigación, describiendo el CERN y sus necesidades documentales, junto con una visión general del problema que el indexado manual representa. En el capítulo 3 se revisa el área de investigación donde se encuadra esta tarea, junto con las convenciones notacionales y la arquitectura que representa nuestra propuesta para el multi-etiquetado de colecciones de documentos con alto grado de desbalanceo.

Desvelar las diferentes técnicas involucradas en la clasificación automática de documentos ocupa el contenido del capítulo 3. El capítulo 4 introduce algunos métodos orientados a la extracción de características. El capítulo 5 expone algunas estrategias encaminadas a reducir la alta dimensionalidad propia de una representación del documento mediante modelos de espacio vectoriales. Finalmente, el capítulo 6 expone una estrategia de clasificación multi-etiquetado construida a partir de clasificadores base binarios.

En el capítulo 7 se proponen y discuten los aspectos relativos a la evaluacion de dichos sistemas. El capítulo 8 analiza en detalle algunas de las aplicaciones actuales y potenciales de estos sistemas automáticos, explorando el vasto rango de posibilidades de la clasificación computerizada. A continuación, en el capítulo 9, se ofrece una revisión histórica de los sistemas existentes que intentan dar una solución automática de este problema, remarcando sus aportaciones más relevantes y sus limitaciones principales.

La segunda parte del trabajo enfoca la solución propuesta al problema del multi-etiquetado. Comienza con el capítulo 10, donde el sistema propuesto es descompuesto para facilitar la descripción de cada componente. Los datos usados en el marco experimental así como sus características destacables son detalladas en el capítulo 11. En el capítulo 12 se recoge el vasto trabajo experimental llevado a cabo para estudiar la viabilidad y capacidad del modelo propuesto. Estos experimentos (desde la sección 12.1 a la sección 12.11) están diseñados para cubrir casi cada pequeño aspecto en el diseño de un sistema de multi-etiquetado completo, así como para evaluar las hipótesis introducidas en la solución propuesta. El último experimento documentado (sección 12.12) estudia el comportamiento del sistema propuesto sobre un conjunto de datos de dominio totalmente diferente al HEP.

Finalmente, el capítulo 13 resume las conclusiones principales encontrados a lo largo de la experimentación, y establece algunas de las líneas futuras aún por investigar. El trabajo finaliza con la bibliografía utilizada durante todo el proceso de investigación y se complementa con unos apéndices con información adicional sobre algunos aspectos que merecen mayor nivel de detalle y quedaron excluídos de la argumentación principal con vistas a mejorar la comprensión y claridad del texto.