Medical-Miner: Integración de conocimiento textual explícito en técnicas de minería de datos para la creación de herramientas traslacionales en medicina

Medical-Miner: Integrating explicit knowledge in data mining techniques for the development of translational medicine tools

Manuel de Buenaga
Enrique Puertas
Univ. Europea de Madrid
Grupo de Sistemas Inteligentes
{buenaga|enrique.puertas}@uem.es

Florentino Fdez-Riverola
Daniel Glez-Peña
Univ. de Vigo
SING - DI
{riverola|dgpena}@uvigo.es

Manuel J. Maña López
Jacinto Mata Vázquez
Univ. de Huelva
D. Tec. de la Información
{manuelj.mana|mata}@dti.uhu.es

Resumen: El proyecto plantea analizar, experimentar y desarrollar nuevas tecnologías de minería de texto y datos, de forma interrelacionada en sistemas inteligentes de información médica. Se implementará un sistema inteligente de acceso a la información basado en ellas, que ofrezca funcionalidades avanzadas con capacidad para interrelacionar la información médica, principalmente la información (texto y datos) de historiales clínicos y documentación científica, haciendo uso de recursos estándar del dominio (e.g. UMLS, SNOMED, Gene Ontology). Se plantea el desarrollo de una plataforma de código abierto integrando todos los elementos.
Palabras clave: Minería de texto, Minería de datos, Dominio Biomédico, Plataforma Opensource

Abstract: The project proposes to analyse, experiment and develop new text and data mining techniques in an interrelated way, in intelligent medical information systems. An intelligent information access system based on them will be developed, offering advanced functionalities able to interrelate medical information, mainly information (text and data) from clinical records and scientific documentation, making use of standard resources of the domain (e.g. UMLS, SNOMED, Gene Ontology). An open source platform will be developed integrating all the elements.
Keywords: Text Mining, Data Mining, Biomedical Domain, Opensource Platform

1 Introducción

El proyecto Medical-Miner se centra en la integración de técnicas de minería de datos y minería de texto en el marco del desarrollo de herramientas para la medicina traslacional. Es un proyecto financiado por el Ministerio de Ciencia y Tecnología en el Programa de Investigación Fundamental (TIN-2009-14057-C03) y se desarrolla desde enero del 2010 hasta diciembre del 2012.

La medicina traslacional es un esfuerzo emergente en la práctica médica que busca trasladar los descubrimientos científicos desde los laboratorios a la práctica clínica para el diagnóstico y tratamiento de los pacientes (conocido en inglés con la expresión “bench-to-bedside”). El cambio de perspectiva se ha producido recientemente como resultado de la revolución genómica y bioinformática. Sin embargo, esa misma revolución ha venido acompañada de un grave problema: la generación de gran cantidad de información que está produciendo un importante cuello de botella en la investigación médica y su aplicación. Esta información se encuentra tanto en formato estructurado, fundamentalmente relacionada con las investigaciones en biología molecular, como texto, procedente de resultados de investigación.

2 Objetivos del proyecto y beneficios esperados

El objetivo principal del proyecto es analizar, experimentar y desarrollar nuevas tecnologías

de minería de texto y datos, de forma interrelacionada en sistemas inteligentes de información médica. Se diseñarán nuevas técnicas de ambos tipos, más eficientes, interrelacionadas, y mejor adaptadas a problemas específicos del dominio. Se implementará un sistema inteligente de acceso a la información basado en ellas que ofrezca funcionalidades avanzadas con capacidad para interrelacionar la información médica, principalmente la información (texto y datos) de historiales clínicos y documentación científica, haciendo uso de recursos estándar del dominio como UMLS, SNOMED y Gene Ontology. Se desarrollará una plataforma de código abierto integrando todos los elementos. Se realizará una evaluación tanto de efectividad de las nuevas técnicas integradas, como del sistema en su conjunto, en entornos abiertos sobre usuarios finales.

Los beneficios esperados de la consecución de los objetivos del proyecto abarcan una doble vertiente: por un lado, se aportarán mecanismos específicos de utilidad para un ámbito profesional (biomédico), y por otro, se obtendrán técnicas innovadoras y avances significativos en la investigación en el ámbito de la minería de texto y la minería de datos, concretados sobre éste dominio, y basados de forma significativa en seguir este enfoque integrador. Se espera conseguir una mejora en la precisión de los modelos existentes para el análisis de texto y datos de tipo biomédico y la generación de explicaciones a partir de los resultados obtenidos, y poder ofrecer herramientas con funcionalidades avanzadas basadas en la interrelación de la información médica científica y del paciente.

3 Metodología

La metodología planteada busca conseguir el objetivo general del proyecto de forma fundamentada en las prácticas comunes en esta área de trabajo. Se distinguen tres tareas básicas que se pueden realizar en paralelo: (i) el diseño de técnicas de minería de textos, tarea orientada a la adaptación, mejora de la efectividad, y evaluación de las técnicas de análisis utilizadas (categorización de textos, reconocimiento de entidades nombradas, identificación de la negación y generación automática de resúmenes), (ii) una tarea de diseño y evaluación de técnicas de minería de datos que permitan el procesamiento de información médica, y (iii) el desarrollo de una plataforma *open source* que permita la integración de las técnicas de análisis anteriores.

Respecto a la primera de estas tareas, cabe destacar que las técnicas de clasificación de textos abordadas en este proyecto presentan importantes similitudes en cuanto a las técnicas utilizadas hasta la actualidad para su implementación. Lo más relevante es que todas ellas se pueden abordar como problemas de aprendizaje automático. En consecuencia, nuestro primer paso consistirá en definir sus aspectos comunes, y construir un conjunto de herramientas que permitan realizar su implementación y evaluación de manera sistemática.

La segunda de las tareas mencionadas, la minería de datos, tiene como objetivo principal la adaptación de las técnicas existentes de clasificación y agrupamiento para su aplicación sobre datos pertenecientes al dominio biomédico. Concretamente, la investigación girará en torno al desarrollo de un modelo híbrido de IA capaz de integrar conocimiento biológico con el fin de mejorar los resultados obtenidos por técnicas clásicas de análisis de microarrays en tareas de clasificación y agrupamiento. Esta mejora de resultados es de especial relevancia en dominios biológicos donde, debido al volumen de datos a tratar, se hace necesaria la utilización de técnicas para el filtrado de datos.

Puesto que la interpretación de genes co-expresados y patrones coherentes dependen de manera directa del dominio de conocimiento, el modelo a desarrollar deberá incorporar mecanismos que posibiliten la integración de nuevo conocimiento biológico a medida que vaya estando disponible. Esto permitirá a las técnicas clásicas de clasificación y agrupamiento la utilización de diversas fuentes de información según la naturaleza del problema.

La integración de las técnicas desarrolladas en el ámbito de la minería de textos, de la minería de datos y el acceso a fuentes de información biológica, se llevará a cabo a través de una plataforma *open source* que se irá desarrollando de modo paralelo a las anteriores. La plataforma dará soporte al trabajo colaborativo realizado por los integrantes de los tres grupos que componen el proyecto.