Resumo do Traballo de Fin de Grao

El correo electrónico es una de las herramientas de comunicación más populares en Internet. Es por ello que muchas organizaciones y/o usuarios han descubierto su potencial como herramienta gratuita para enviar información no deseada (publicidad, cadenas de mensaje, etc.), dando lugar al nacimiento del correo spam. Este uso abusivo del correo limita la eficacia, la experiencia e incluso la seguridad de los usuarios. Algunos estudios recientes estiman que en torno al 60% de los correos enviados son spam. El despliegue de filtros semánticos de forma eficiente capaces de clasificar el correo en base a su contenido es una de las soluciones. Sin embargo, resulta complejo realizar filtrado semántico de los correos entrantes en tiempo real. Una de las principales alternativas para lograr el filtrado en tiempo real pasa por disponer de perfiles de usuario que permiten desplegar aproximaciones probabilísticas. El perfil de usuario (o de organización si se quiere filtrar a nivel de organización) define exactamente qué temáticas son de interés y cuáles no. La elaboración de este perfil se puede realizar de forma manual listando una serie de temáticas agrupadas jerárquicamente y dejando que el usuario escoja las que desea recibir y las que no. Además, este trabajo implementará un sistema de clasificación basado en probabilidades con el objetivo de medir su eficacia y eficiencia frente a otras aproximaciones semánticas.

Las consultas semánticas y procesos realizados mediante ontologías requieren mucho tiempo de procesamiento. Para evitar las consultas semánticas y mejorar el tiempo de clasificación, es necesario acudir a aproximaciones que combinan la idea de bolsa de términos y las probabilidades de que estos términos están escritos en un sentido conectado con una temática interesante o irrelevante para el usuario. Para ello, para cada usuario se recopilará un listado de términos conectados con aquellos los conceptos que el usuario acepta recibir y los que no. Cuando un término tiene múltiples acepciones (como es el caso de banco), la probabilidad deberá calcularse en función de la frecuencia de aparición de la acepción del usuario. Para aproximar esta probabilidad se podrían usar distintos textos tratando de determinar esta frecuencia (por ejemplo, son más comunes las acepciones de institución financiera o mobiliario que la de conjunto de peces), o bien emplear una aproximación más simplista basado en el número de distintas acepciones de la palabra. Recopilando esta información para cada uno de los términos posibles en una estructura eficiente de recuperación, será fácil combinar las probabilidades obtenida junto con el interés del usuario por el término para calcular si un paquete/mensaje le interesará o no. El perfil de un usuario permitirá saber exactamente sus preferencias y, al haberlo determinado con anterioridad, evitará tener que compilarlo durante la etapa de clasificación ahorrando el tiempo correspondiente.

El objetivo de este trabajo de fin de grado es el desarrollo de una aplicación de configuración a través de la cual un usuario elabore o cambie manualmente su perfil de acuerdo a sus necesidades. La metodología a emplear para el desarrollo de la aplicación es Scrum(proceso en el que se aplican de manera regular un conjunto de buenas prácticas para obtener el mejor resultado posible de un proyecto. Estas prácticas se apoyan unas a otras y su selección tiene origen en un estudio de la manera de trabajar altamente productiva).

Para el desarrollo del trabajo se empleará un equipo que cuenta con: 16 Gb de RAM, un i7-4771 (3.5 GHz), un SSD de 240 Gb, una gráfica MSI GTX 1060 (6 GB) y una fuente de alimentación Plus Gold 800W modular.