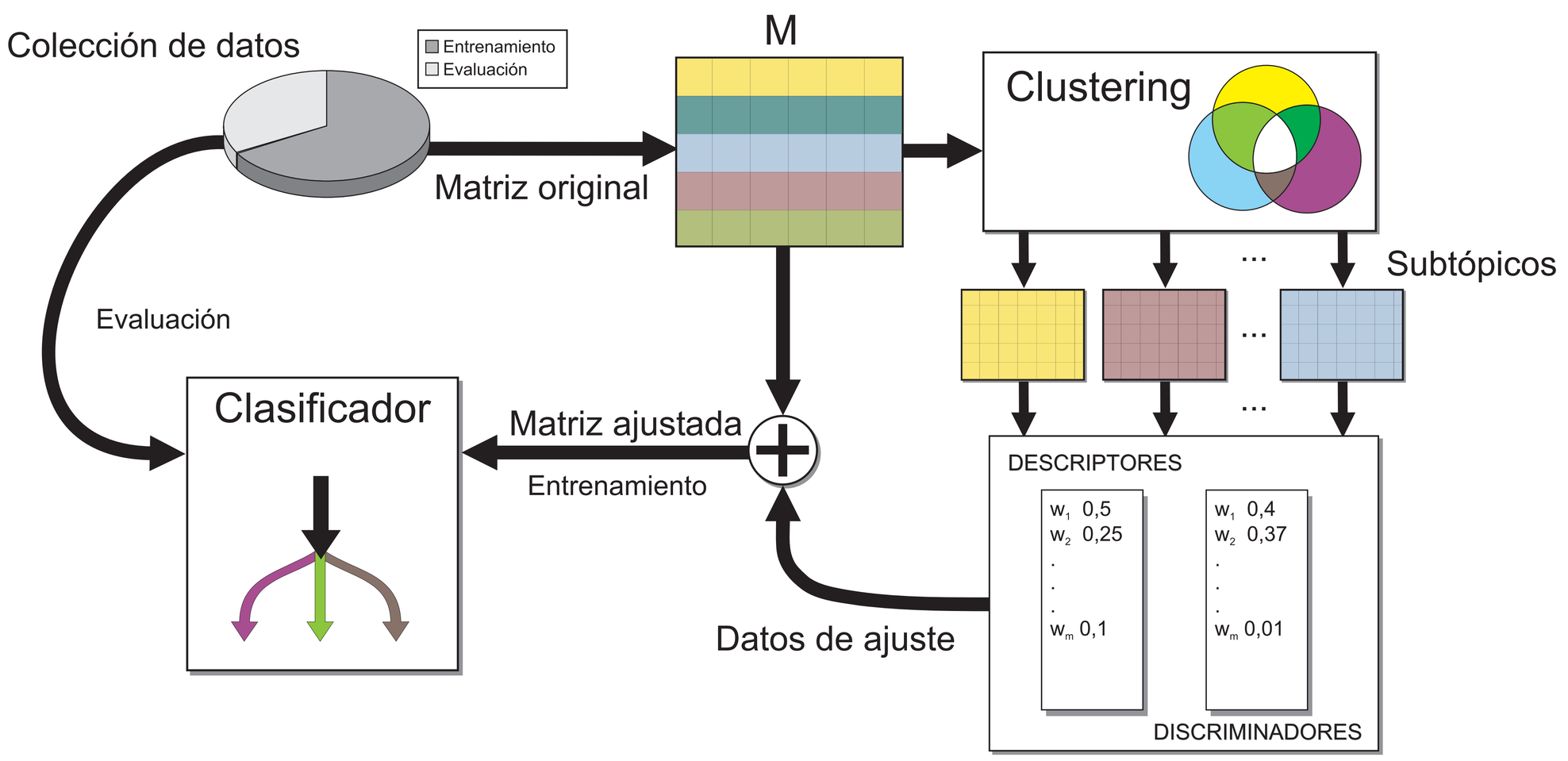
UNIVERSIDAD NACIONAL DEL SUR
DEPARTAMENTO DE CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

Métodos para la Selección y el Ajuste de Características en el Problema de la Detección de Spam

El correo electrónico es quizás la aplicación que más tráfico genera en la Internet y es una aplicación de misión crítica para muchos negocios. En la última década la avalancha de correo no deseado (Spam) ha sido el mayor problema para los usuarios del correo electrónico. El spam no solo es frustrante para muchos usuarios, sino que también compromete a la infraestructura tecnológica de las empresas, costando dinero a causa de la pérdida de productividad. En los últimos años, el spam ha evolucionado desde ser una molestia a ser un serio riesgo en la seguridad, llegando a ser el principal medio para el robo de información personal, así como también para la proliferación de software malicioso.

Los métodos de Aprendizaje Automatizado son atractivos para realizar la tarea de detección de spam ya que son capaces de adaptarse a las características evolutivas del spam, contándose además con disponibilidad de datos para entrenar tales modelos.

NUESTRA PROPUESTA PLANTEA UN AJUSTE EN LOS DATOS DE ENTRADA DEL CONJUNTO DE ENTRENAMIENTO DE UN CLASIFICADOR CON EL OBJETIVO DE MEJORAR SU RENDIMIENTO.



Ajuste de Características en la Detección de Spam

El algoritmo propuesto está compuesto por los siguientes pasos básicos:

- División de los datos de entrada en Entrenamiento y Testeo,
- Aplicación de un algoritmo de CLUSTERING,
- Cálculo de los DESCRIPTORES y DISCRIMINADORES de cada subtópico,
- Ajuste del Modelo de Datos,
- CLASIFICACION de los documentos,
- Evaluación.

En las etapas de Clustering y Clasificación se evaluarán distintos algoritmos existentes en la literatura analizando sus fortalezas y debilidades en las distintas colecciones de documentos disponibles.
El ajuste de los datos a través de la detección de buenos *descriptores* y buenos *discriminadores* mejorará los datos de entrada del clasificador final.

Integrantes de la Línea de Investigación

Maguitman Ana G.[1] agm@cs.uns.edu.ar
Benczur András A.[3] benczur@ilab.sztaki.hu
Cecchini Rocío L.[2] rlc@cs.uns.edu.ar
Lorenzetti Carlos M.[1] cml@cs.uns.edu.ar

GRUPO DE INV. EN RECUPERACIÓN DE INFORMACIÓN Y GESTIÓN DEL CONOCIMIENTO

El Grupo de Investigación forma parte del Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (**LIDIA**), el cual se formó en 1992 con el objetivo de nuclear a docentes investigadores y alumnos avanzados de diferentes Departamentos de la **UNIVERSIDAD NACIONAL DEL SUR** vinculados a la investigación en el área de Inteligencia Artificial. Dentro del LIDIA, se encuentran en progreso varias líneas de investigación orientadas al estudio formal de Razonamiento Rebatible para Agentes Inteligentes Autónomos.

Principales Líneas de Investigación del LIDIA

- Programación en Lógica Rebatible
- Robótica Cognitiva Aplicada a la Robótica Móvil
- Programación en Lógica Rebatible Posibilística
- Revisión de Creencias, Explicaciones y Razonamiento Rebatible
- Negociación y Programación en Lógica Rebatible
- Sistemas de Representación de Conocimiento Basados en Lógica
- Razonamiento en la Web
- Recuperación de Información y Minería de Datos en la Web

El Grupo de Investigación en Recuperación de Información y Gestión del Conocimiento se dedica al estudio del tratamiento computacional de grandes volúmenes de información obtenidos de la Web. Las herramientas de búsqueda tradicionales tienen limitaciones para responder a consultas teniendo en cuenta los deseos del usuario o su contexto de trabajo. Se está trabajando en el desarrollo de nuevos mecanismos de razonamiento en la Web que permitan integrar conceptos de inteligencia artificial con los algoritmos tradicionales usados por los buscadores.

[1] Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
[2] Laboratorio de Investigación y Desarrollo en Computación Científica
[3] Data Mining and Web search Research Group - MTA STAKI - Hungría