КЛАССИФИКАЦИЯ ЭЛЕКТРОННОЙ ПОЧТЫ

Будник А.С.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Герман Ю.О. – канд. тех. наук

Рассматривается метод классификации электронной почты на основе выделения ключевых слов в письме и алгоритма кластеризации К-средних.

С недавнего времени электронная почта сменила устаревшую бумажную переписку по причине своей практичности. Актуальность создания методов классификации объясняется следующим: большое количество писем приходит на почтовые ящики ежедневно, сортировка этих писем поможет сэкономить время пользователя.

Для решения данной задачи можно использовать алгоритм нахождения ключевых слов в письме, а затем классифицировать письма на основе одного из методов кластеризации.

Любой текст можно охарактеризовать набором ключевых слов, которые этот текст представляют. Первоначально рассматривают только существительные. Подсчитываются частоты слов (сколько раз каждое слово вошло в текст). Поскольку слова входят в текст с разными окончаниями, то гласные выбрасываются и два слова считаются совпадающими, если они достаточно похожи друг на друга. При этом короткие слова должны быть похожими друг на друга в большей степени, чем длинные. Затем слова размещаются по убыванию частот, и берется примерно корень из общего числа слов с наибольшими значениями частот. При этом чем длиннее слова, тем меньше требуется допустимый процент совпадения [1]. Этот процент можно установить исходя из следующей эмпирически составленной таблицы 1.

Таблица 1 – Допустимое число неверных символов

Размер слова	Допустимое число неверных символов
До 5	0
5-6	1
7-8	2
9-10	3
Больше 10	4

На следующем этапе будет использоваться алгоритм К-средних.

Кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы [2].

Метод К-средних – это метод кластерного анализа, цель которого – разделение m наблюдений на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру которого оно ближе всего. Результат разбиения на кластеры представлен на рисунке 1.

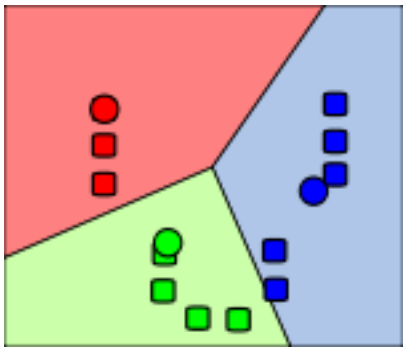

Рисунок 1 – Результат работы алгоритма К-средних

Таким образом, на основе тестовой выборки будут сформированы кластеры почтовых писем (например, деловые письма, поздравления, спам). Затем при получении нового письма будет определяться набор его ключевых слов, и письмо будет отнесено к тому или иному кластеру, тем самым будет выполнена процедура классификации писем.

Список использованных источников:

1. Герман, О.В. Искусственный интеллект: метод. пособие / О. В. Герман, Ю.О. Герман. – Минск : БНТУ, 2013. – 127с.
2. Мандель, И.Д. Кластерный анализ / И .Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.