Секция 3.

КЛАСТЕРИЗАЦИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ КОЛИЧЕСТВЕННОГО КОНТЕНТ-АНАЛИЗА И Q-СОРТИРОВКИ

Порошенко А.И., Иващенко Г.С.
Харьковский национальный университет радиоэлектроники, Харьков

Среди задач анализа больших данных существенное место занимает анализ текстовой информации, для которой характерно отсутствие ключевых слов, определенных самим автором, что затрудняет определение тематики работы. Кроме того, часто не определено авторство текста. В работе представлен сравнительные анализ таких методов, как определение авторского инварианта, подготовка матрицы переходов на основе разбиения слов на фрагменты по 2-3 символа и использование цепей Маркова. Рассмотренные методы применяются для решения задачи атрибуции текста, но не обеспечивают необходимую точность и имеют требования по объему анализируемого текста.
Предлагается гибридный подход на основе метода авторского инварианта, количественного контент-анализа и Q-сортировки для фиксации определенных единиц содержания, что позволяет формировать описание анализируемого текста набором выявленных ключевых слов. Предложенный подход позволяет выполнять кластеризацию данных, определяя основную и дополнительные темы анализируемого текста, с разной степенью принадлежности.

Антон Ігорович Порошенко, 066-210-86-32, студент кафедри ЕОМ,
Харківський національний університет радіоелектроніки, 61166, пр. Науки, 14.
anton.poroshenko@nure.ua

Георгій Станіславович Іващенко, 067-751-85-90, ст. викладач кафедри ЕОМ,
Харківський національний університет радіоелектроніки, 61166, пр. Науки, 14.
heorhii.ivashchenko@nure.ua