АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДЛЯ СИСТЕМЫ ФОРМИРОВАНИЯ КОЛЛЕКЦИЙ НАУЧНЫХ ПУБЛИКАЦИЙ

д-р техн. наук, доц., зав. каф. "Компьютерных наук" Е.Е. Федоров, магистр О.Н. Полякова, Донецкий Национальный технический университет, г. Красноармейск

Решаемая задача – создание коллекций (кластеров) тематически близких документов. Основной инструмент для решения задачи – кластерный анализ, использующий в качестве признаков ключевые слова документа. К достоинствам кластерного анализа относится то, что он не требует больших объемов обучающих данных и длительной процедуры обучения. В ходе решения были рассмотрены такие алгоритмы неиерархической кластеризации как алгоритм *k*-средних, нечетких *c*-средних, ожидания-максимизации, соответствующих метрической, нечеткой и байесовской кластеризациям. В качестве расстояния для алгоритма *k*-средних и нечетких *c*-средних бралось расстояние Хемминга. Проведен анализ эффективности применения указанных алгоритмов кластеризации для формирования коллекций документов. Был выбран наиболее эффективный алгоритм кластеризации для системы формирования коллекций научных публикаций по ключевым словам документа [1 – 4].

Получены новые теоретические результаты, обосновывающие выбор алгоритма кластеризации для создания коллекций научных публикаций. В итоге наилучший результат показал алгоритм ожидания-максимизации (95%).

Список литературы: 1. *Котов А.* Кластеризация данных / *А. Котов, Н. Красильников.* – 2006. – 16 с. **2.** *Воронцов К.В.* Алгоритмы кластеризации и многомерного шкалирования / *К.В. Воронцов* // Курс лекций. – М.: МГУ, 2007. **3.** *Леоненков А.В.* Нечеткое моделирование в среде MATLAB и fuzzyTECH / *А.В. Леоненков.* – СПб.: БХВ-Петербург, 2003. – 736 с. **4.** *Черноруцкий И.Г.* Методы оптимизации. Компьютерные технологии / *И.Г. Черноруцкий*. – СПб.: БХВ-Петерберг, 2011. – 384 с.