ИСПОЛЬЗОВАНИЕ МЕТОДОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ ПРИ ПОИСКЕ ТЕКСТОВЫХ ДАННЫХ В БОЛЬШИХ КОЛЛЕКЦИЯХ ДОКУМЕНТОВ

к.т.н., доц. Т.В. Гладких, студент Н.М. Красиков, НТУ "ХПИ", г. Харьков

На протяжении более чем 50-ти лет наблюдается неуклонный рост объема текстовой документации, которая используется в различных областях человеческой деятельности. Следствием этого, а также в связи со стремительным развитием Интернет, особую актуальность приобретает автоматизация процесса систематизации текстовых массивов, в частности, для поиска наиболее релевантной информации в больших коллекциях текстовых данных, к которым, например, можно отнести документы, выдаваемые поисковыми системами в ответ на запросы пользователей. Согласно данным журнала Reuters, в России 38% менеджеров "тратят много времени на поиск нужной информации". 79% журналистов обращаются за информацией в Интернет и всего лишь 20% находят то, что им необходимо.

Все это обуславливает важность такой задачи, как разработка специальных методов кластеризации текстовых документов, без решения которой невозможна эффективная работа с текстовой информацией. На сегодня наиболее востребованной является смысловая кластеризация текстовых документов, которая выполняет разделение текстовых коллекций на множества текстов (кластеры), такие, что тексты в пределах одного и того же кластера максимально схожи между собой по смыслу, а тексты, относящиеся к разным кластерам, имеют различный смысл.

В докладе рассмотрено применение нечетких методов кластеризации при поиске отдельных фрагментов текстовых данных в больших коллекциях документов, а также систематизация данных внутри текстовых массивов. Целью является повышение релевантности поиска, а также возможное создание математических алгоритмов, использующих методы нечеткой кластеризации для поиска или систематизации текстовых данных. Затронута область рубрицирования данных, в частности, тематическое рубрицирование текстовых данных.

Рассмотрен метод индексирования текстовых данных, а так же методы их классификации. Проанализированы результаты обзоров и практических исследований в этой и смежных областях специалистов из стран СНГ, Европы, США. Подготовлен теоретический материал, позволяющий создать собственный или усовершенствовать существующий алгоритм нечеткой кластеризации для поиска различных текстовых данных в больших коллекциях документов.