***М. А. Баранов**, аспирант Московского института электроники и математики НИУ ВШЭ*

Модификация жадного алгоритма кластеризации

Задача создания эффективных по времени и релевантности результата алгоритмов документального поиска сохраняет свою актуальность. Возможности для усовершенствования предоставляют жадные алгоритмы кластеризации.

Введение

Наблюдаемый рост количества создаваемых документов делает все более актуальной задачу создания эффективных алгоритмов информационного поиска, одной из составляющей которых является кластеризация — разбиение исходного множества документов на группы, состоящие из схожих документов. Из факта принадлежности документа кластеру и его релевантности по отношению к некоторому информационному запросу с высокой вероятностью следует, что и остальные документы данного кластера также релевантны этому запросу [13].

В настоящее время разработано множество алгоритмов кластеризации, использующих различные подходы к решению задачи кластерного анализа. Их классификация подробно изложена в работе [3]. Из всего многообразия используемых при кластеризации подходов стоит выделить так называемые жадные методы, суть работы которых сводится к тому, что на каждом шаге они делают локально оптимальный выбор в расчете на то, что это приведет к оптимальному решению всей задачи [5]. Стоит отметить, что жадные алгоритмы часто используются при решении задач кластеризации (см., например, [1, 2, 9, 12, 18]).

Одним из таких алгоритмов является алгоритм, предложенный в [2]. Суть его состоит в следующем. Предварительно строится матрица схожести документов размером $N \times N$, где N — число документов в коллекции, а документу с номером k соответствуют k-й столбец и k-я строка. В матрице в каждой ячейке (i, j) задается мера схожести между i-м и j-м документами. Предлагается подбирать меру схожести таким образом, чтобы ее значение лежало в диапазоне от 0 до 1, где 0 соответствует полному различию документов, а 1 — полному сходству.

На первом шаге в матрице схожести находится строка, сумма элементов которой является максимальной. Документ, соответствующий ей, объявляется центром очередного кластера, а сама строка содержит коэффициенты схожести этого документа со всеми остальными документами коллекции.

На втором шаге в кластер добавляются все документы, коэффициент схожести которых с центром кластера больше некоторого заранее заданного значения или равен ему (параметр threshold). После этого из матрицы удаляются все строки и столбцы, соответствующие попавшим в кластер документам.

Шаги 1 и 2 повторяются до тех пор, пока не останется документов, не включенных в какой-либо кластер.

В данной работе предлагается модифицированная версия описанного алгоритма. Суть модификации состоит в том, что при добавлении нового документа в кластер учитываются коэффициенты схожести этого документа по отношению к другим, уже включенным в кластер документам.