М. А. Баранов, *аспирант Национального исследовательского университета «Высшая школа экономики», г. Москва, thenorthcat@gmail.com*

Параллельная версия жадного алгоритма кластеризации

В статье предлагается параллельная версия одного из алгоритмов кластеризации, принцип работы которого основан на так называемом жадном подходе. Для решения задачи распараллеливания алгоритма была выбрана технология CUDA, разработанная компанией NVIDIA. Приводятся программный код и результаты вычислительных экспериментов для матриц схожести разного размера.

Ключевые слова: кластеризация, жадный алгоритм, параллельные вычисления.

Введение

Одной из наиболее важных задач в информационном поиске является кластеризация — разбиение исходного множества объектов на группы, состоящие из схожих объектов. Кластеризация нашла широкое применение в различных областях знаний: в биологии, социологии, информатике, астрономии, медицине, археологии, маркетинговых исследованиях.

Кластерный анализ относится к процессам обучения без учителя, его основная цель — разбить множество объектов на группы таким образом, чтобы объекты внутри одной группы были максимально похожими друг на друга, но в то же время максимально отличались от объектов другой группы.

В настоящее время разработано множество алгоритмов кластеризации, использующих различные подходы к решению задачи кластерного анализа. Их классификация подробно изложена в работе [4]. Из всего многообразия используемых при кластеризации подходов стоит выделить так называемые жадные алгоритмы, 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]. Жадные алгоритмы часто используются при решении задач кластеризации [7, 10].

Одним из таких алгоритмов является алгоритм, предложенный в работах [1, 2] (далее он будет именоваться Greedy). Там же показана эффективность данного алгоритма при кластеризации коллекций текстовых документов. На вход алгоритма подаются матрица схожести документов и пороговое значение степени схожести (параметр threshold). Блок-схема алгоритма представлена на рис. 1.

Целью данной работы является сравнительный анализ параметров параллельной и последовательной версий алгоритма Greedy при кластеризации коллекций документов разумного размера (несколько тысяч документов).

Краткое описание технологии CUDA

Для решения задачи распараллеливания алгоритма была выбрана аппаратно-программная архитектура CUDA (Compute Unified Device Architecture), разработанная компанией NVIDIA. Отметим, ранее в ряде работ данная технология была успешно использована для создания параллельных версий алгоритмов кластеризации [8, 9].

Архитектура CUDA основана на архитектуре вычислительных систем SIMD (single