Ф. И. Третьяков, Л. В. Серебряная
(БГУИР, Минск)

КЛАСТЕРИЗАЦИЯ ТЕКСТОВЫХ ДАННЫХ НА ОСНОВЕ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ

Стремительное развитие информационных технологий способствует резкому увеличению доступных информационных ресурсов, вследствие чего постоянно растут объемы информации, которые приходится обрабатывать человеку. При этом информация часто оказывается разнородной, слабо структурированной и избыточной, имея высокую динамику обновления. Хранение больших объемов информации практически оправдано только при условии, что ее поиск и обработка осуществляются быстро и выдается она в доступной для понимания форме, что стимулирует создание эффективных методов обработки данных. Настоящая работа посвящена организации параллельной обработки данных, которые сначала требуется разбить на кластеры, а затем объединить полученные решения в единое пространство.

Схему организации параллельных вычислений для выполнения процедуры кластеризации можно представить следующим образом. Имеется набор текстов произвольной тематики, поступающих с целью их кластеризации в динамическом режиме, где общее количество текстов неизвестно. Одновременно в обработке может находиться достаточно большое число текстов. Поэтому предлагается разбить все имеющиеся тексты на некоторое число отдельных подмножеств, содержащих равное число элементов. На каждом подмножестве в параллельном режиме выполняется построение собственных кластеров. Эффективная реализация параллельных вычислений могла бы существенно уменьшить общее время решения задачи кластеризации [1].

Главным условием эффективности параллельных вычислений при решении поставленной задачи является возможность максимального использования результатов, полученных на текущем уровне, на последующих уровнях вычислений. В противном случае применение параллельных вычислений для решения задачи кластеризации может потерять всякий смысл [2].

Литература

1. Theodoridis, S. Pattern Recognition [Text] / S. Theodoridis, K. Koutroumbas.— 4th edition.— Athens: Academic Press, 2009.— 874 p.— ISBN 978-1-59749-272-0.

2. Таненбаум, Э. Распределенные системы [Текст]: принципы и парадигмы / Э. Таненбаум, М. ван Стеен.— Санкт-Петербург: Питер, 2003.— 877 с.— (Классика computer science).— ISBN 5-272-00053-6.