КЛАСТЕРИЗАЦИЯ ИЗОБРАЖЕНИЙ В ГРИД СИСТЕМЕ

О.В. Лобанов

Научный руководитель: доцент, к.т.н. А.Ю. Дёмин

Национальный исследовательский Томский политехнический университет,

Россия, г. Томск, пр. Ленина, 30, 634050

E-mail: Lobanovov@tpu.ru

IMAGE CLUSTERING IN GRID COMPUTING

O.V. Lobanov

Scientific Supervisor: Associate Professor, Ph.D. A.Yu. Demin

Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050

E-mail: Lobanovov@tpu.ru

Abstract. *This paper shows the feasibility of creating software that allows you to distribute the highload on the main features identification of images for their further clustering in grid computing systems.*

The task of the images-clustering was distributed in a three-element computing cluster. An image-clustering algorithm was k-mean with Euclidean distances. The results were compared with the results of running the algorithm on a single element. To assess the effectiveness, the scalable acceleration of the Gustafson law was calculated. The numerical values of acceleration exceed 2, which shows the expediency of using the considered approach in a similar class of problems.

Введение. Человек существует и взаимодействует с окружающим миром в течение всей своей жизни. Для восприятия мира он, как и любое другое живое существо, использует органы чувств. Создавая искусственные вычислительные системы по своему образу и подобию, людям приходится продумывать аналогичные механизмы восприятия и для них. Каждый объект или событие окружающего мира при обработке компьютером представляются моделью с необходимыми параметрами.

Современные суперкомпьютеры ежегодно бьют рекорды по производительности, но даже они не способны воспринимать картину мира во всей полноте, доступной человеку. Это ограничение вызвано множеством факторов, один из которых – признаки объекта реального мира, которые используются при создании и обработки модели внутри компьютера. В то же время вопрос доступности суперкомпьютеров для большинства желающих может оказаться нерешаемым в силу дороговизны обслуживания. Куда более доступным инструментом перевода объектов реального мира в компьютерную модель является персональный компьютер или небольшой личный виртуальный сервер.

В данной работе рассматривается целесообразность создания программного обеспечения, позволяющего распределить нагрузку по выявлению отличительных признаков у изображений для их дальнейшей кластеризации в грид-системах [1];

Экспериментальная часть. Задачу кластеризации набора изображений определим как необходимость разбиения набора изображений P на несколько групп N [2], по признаку S, тогда:

Принадлежность P_i к группе N_j (причем i > j) будет определяться некоторой функцией принадлежности $N_j = F_1(S_i)$, где S_i – признак изображения P_i, полученный в результат применения функции $F_2(P_i)$ к данному изображению.

В контексте данной работы под функцией принадлежности (F_1) будет пониматься результат работы алгоритма кластеризации k-средних, где мера расстояния будет Евклидовым, а функция выделения признака изображения (F_2) следующая:

$$S_i = \begin{pmatrix} R_i \\ G_i \\ B_i \end{pmatrix} / (pxsum_i),$$

где *R, G, B* – суммарное количество цветовых компонент всех пикселей изображения, а *pxsum* – количество пикселей в изображении.

Первым этапом кластеризации набора изображений является создание набора признаков всей выборки изображений, данная задача может быть распределена в вычислительном кластере.

С целью создания вычислительного кластера типа grid было разработано программное обеспечение на языке C# для использования в кроссплатформенной программной среде .Net Core.

Вычислительный кластер был развёрнут в двух сценария использования: на одной типовой машине и на трёх вычислительных элементах. Типовая конфигурация машины: 1 ядерный процессор с таковой частотой 1 ГГц, 1 Гб ОЗУ и 10 Гб ПЗУ (HDD).

Результаты. Для каждого сценария проведены экспериментальные исследования с различным набором изображений, в таблице ниже представлены усредненные значения их 50 однотипных запусков. Под временем обработки понимается получение результатов выделения признаков изображения и применение алгоритма кластеризации к ним, в том числе включая все затраты на пересылку данных и другие непараллельные операции.

Таблица 1

Временные затраты на кластеризацию

Количество вычислительных элементов	Размер выборки, шт.	Время выделения признаков, сек	Время обработки, сек	Суммарное время, сек
1	60	1,122	0,734	**1,856**
3		1,342	0,789	2,131
1	120	1,462	0,791	2,253
3		1,408	0,801	**2,209**
1	240	2,054	1,002	3,056
3		2,006	1,039	**3,045**
1	500	2,442	2,073	4,515
3		2,112	2,114	**4,226**
1	1000	4,145	3,079	7,224
3		3,102	3,11	**6,212**

Дополнительно была рассмотрена количественная характеристика – масштабируемое ускорение, результат оценки закона Густафсона[3]:

$$speedup \le p + (1 - p) \times s ,$$

где p означает количество ядер. Для упрощения записи s является процентом времени последовательного выполнения в параллельном приложении для указанного размера набора данных.

Таблица 2

Масштабируемое ускорение для каждой выборки

Размер выборки, шт.	Время выделения признаков, сек	Время обработки, сек	Суммарное время, сек	P	S	Speedup
60	1,122	0,734	1,856	1	0,395474	1
	1,342	0,789	2,131	3	0,370249	2,259503
120	1,462	0,791	2,253	1	0,351087	1
	1,408	0,801	2,209	3	0,362608	2,274785
240	2,054	1,002	3,056	1	0,32788	1
	2,006	1,039	3,045	3	0,341215	2,31757
500	2,442	2,073	4,515	1	0,459136	1
	2,112	2,114	4,226	3	0,500237	1,999527
1000	4,145	3,079	7,224	1	0,426218	1
	3,102	3,11	6,212	3	0,500644	1,998712

Заключение. Исходя из полученных данных о суммарnov времени затрат на задачу по кластеризации изображений, можно судить о преимуществе подхода с использованием вычислений в грид-системе для выборок больше 60 изображений. Кроме того, масштабируемое ускорение больше единицы дополнительно указывает эффективность подобного подхода. По итогу можно говорить о возможности использования грид-вычислений в задачах по кластеризации изображений с целью распределения нагрузки и снижения временных затрат.

СПИСОК ЛИТЕРАТУРЫ

1. Грид-система // Национальная библиотека им. Н. Э. Баумана Bauman National Library [Электронный ресурс]. – Режим доступа: https://ru.bmstu.wiki/Грид-система (дата обращения 20.02.19).
2. Немировский В.Б., Стоянов А.К. Кластеризация изображений лиц // Компьютерная оптика. – 2017. – Т. 41, № 1. – С. 59-66
3. Оценка эффективности параллелизации // Intel Software [Электронный ресурс]. – Режим доступа: https://software.intel.com/ru-ru/articles/predicting-and-measuring-parallel-performance (дата обращения 22.02.19).