Эффективная Распределенная Обработка Больших Данных на Основе Наименьшего Информационного Пространства

П.В. Голубцов
Московский государственный университет имени М. В. Ломоносова
Москва, Россия
golubtsov@physics.msu.ru

***Аннотация*—Рассматривается алгебраическая формализации распределенной обработки больших данных. Определяется понятие информационного пространства для заданной процедуры обработки данных и доказывается существование наименьшего информационного пространства, обеспечивающего самую компактную форму накопления информации и позволяющего наиболее эффективно распараллелить обработку. Показано, что в терминах информационного пространства естественным образом выражаются понятия сложения информации и качества информации.**

Ключевые слова— большие данные, параллельная обработка, алгебра информации, качество информации, MapReduce

1. Введение

Данные в современных исследованиях нередко имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате собрать все относящиеся к исследованию данные на одном компьютере, как правило, невозможно и непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. Подходящий алгоритм анализа данных должен, параллельно работая в распределенной системе, извлекать из каждого набора исходных данных некоторую промежуточную компактную информацию, постепенно объединять ее и, наконец, использовать накопленную информацию для получения результата.

В предыдущих работах автора (напр., [1]) были рассмотрены конкретные типы задач обработки данных и исследованы возникающие в них специальные виды представления информации, содержащейся в данных. Было показано, что для эффективной обработки распределенных данных ключевую роль играет возможность введения специальной промежуточной формы представления информации, обладающей определенными алгебраическими свойствами. В рассмотренных задачах были введены соответствующие информационные пространства и исследованы их свойства.

Данная работа призвана подвести общий фундамент под эти исследования путем построения алгебраической формализации распределенной обработки данных. Определяется понятие информационного пространства для заданной процедуры обработки и, в частности, наименьшего информационного пространства, предоставляющего максимально компактную форму представления информации и, как следствие, позволяющего наиболее эффективно распараллелить обработку данных. При этом в терминах информационного пространства естественным образом выражаются бинарная операция сложения фрагментов информации и упорядочение, отражающее понятие качества информации.

Следует отметить, что существует довольно много подходов к понятию информация, например, комбинаторный, вероятностный, алгоритмический [2], однако все они определяют *меру количества информации* в том или ином контексте. Напротив, наименьшее информационное пространство приводит к понятию именно *информации*, содержащейся в данных, как максимально компактное представление набора данных, обеспечивающее тот же результат обработки что и этот набор. В результате, информация, извлеченная из данных, полностью заменяет эти данные.

2. Процедура обработки и информационные пространства

Пусть D – множество возможных значений входных данных, а R – множество значений результатов обработки. В задачах больших данных на вход процедуры обработки поступают наборы элементов из D, причем эти наборы могут быть распределены по многим компьютерам. Для математического представления множества всех таких наборов с операцией их слияния обычно используется свободный моноид D^* с операцией конкатенации. Однако, поскольку результат обработки как правило не должен зависеть от порядка поступления данных, удобно представлять пространство всевозможных наборов исходных данных *свободным коммутативным моноидом* D^+ с множеством образующих D. Его элементами являются конечные мультимножества на множестве D (в которых элемент может повторяться несколько раз) с операцией *сложения* мультимножеств (при которой кратности одинаковых элементов складываются).

Определение. *Процедурой обработки* с наборами данных из множества данных D и результатами из множества R будем называть отображение p из свободного коммутативного моноида D^+ в множество R, т.е. $p: D^+ \to R$.

Определение. *Информационное пространство* (ИП) (U, q, r) для процедуры $p: D^+ \to R$ это коммутативный моноид U, сюръективный гомоморфизм (СГ) моноидов $q: D^+ \to U$ и отображение $r: U \to R$ такие, что $r \circ q = p$.

Фактически, гомоморфизм q сжимает исходные данные без потери информации, представляя различные наборы данных одним и тем же элементом. Его гомоморфность означает, что объединению наборов данных отвечает сумма соответствующих фрагментов информации, а его сюръективность обеспечивает отсутствие в U элементов, которые не отвечают никаким наборам данных. Эффект от использования ИП

определяется тем, насколько оно позволяет сжать данные.

3. Наименьшее информационное пространство

Определение. Будем говорить, что ИП (U,q,r) *меньше,* чем (U',q',r') и обозначать это как $(U,q,r) \ll (U',q',r')$, если существует такое отображение $h\colon U' \to U$, что $h \circ q' = q$.

Поскольку q' – СГ, такое *преобразование* информационных пространств h единственно и, т.к. q – СГ, также является СГ. При этом $r \circ h = r'$, т.е. (U,h,r) можно рассматривать как ИП для процедуры $r'\colon U' \to R$. Отношение $\ll$ является предпорядком, причем если $U' \ll U$ и $U \ll U'$, то эти ИП изоморфны. *Наименьшее* в смысле этого упорядочения ИП (U,q,r) обладает тем свойством, что любое ИП (U',q',r') для p факторизуется через него, т.е. существует (единственный) СГ $h\colon U' \to U$ для которого $h \circ q' = q$ и $r' = r \circ h$.

Теорема (Существование). *Наименьшее ИП для процедуры $p\colon D^+ \to R$ существует и единственно с точностью до изоморфизма.*

Для исследования структуры наименьшего ИП дадим следующее

Определение. Пусть U – коммутативный моноид. Будем говорить, что элементы x и y из U *неразличимы* относительно $r\colon U \to R$ и обозначать $x \sim_r y$, если

$$\forall z \in U \quad r(x+z) = r(y+z).$$

Теорема (Конструкция). *ИП $(D^+/\sim_p, q, r)$ является наименьшим ИП для процедуры $p\colon D^+ \to R$. Здесь $D^+/\sim_p$ – фактормоноид по конгруэнции неразличимости на D^+ относительно p, гомоморфизм $q\colon D^+ \to D^+/\sim_p$ – соответствующий канонический эпиморфизм, $q(x) = [x]_{\sim_p}$ для $x \in D^+$, а отображение $r\colon D^+ \to R$ определяется как $r([x]_{\sim_p}) = p(x)$ для $x \in D^+$.*

В практических задачах (см., напр. [1]) анализ процедуры обработки нередко позволяет предложить естественный вариант ИП. Следующее утверждение дает критерий проверки того, что ИП является наименьшим.

Теорема (Критерий). *ИП (U,q,r) является наименьшим если все его элементы различимы относительно $r\colon U \to R$.*

4. Качество информации

Алгебраическая структура ИП позволяет естественным образом определить упорядочение, характеризующее *качество информации*.

Определение. Для элементов x и y из ИП U будем говорить, что x содержит *больше информации*, чем y и обозначать $x \succcurlyeq y$ если

$$\exists z \in U \quad x = y + z.$$

Отношение $\succcurlyeq$ на ИП U является отношением предпорядка, согласованным с алгебраической структурой, т.е., $x' \succcurlyeq x \wedge y' \succcurlyeq y \implies x' + y' \succcurlyeq x + y$ и $x \succcurlyeq 0$. Более того, преобразование ИП $h\colon U' \to U$ сохраняет упорядочение качества: $x \succcurlyeq y \implies h(x) \succcurlyeq h(y)$.

5. Накопление информации в MapReduce

Использование наименьшего ИП позволяет максимально эффективно распараллеливать процесс накопления информации в рамках модели распределенного анализа данных MapReduce [3] и организовать эффективную обработку без необходимости передачи и накопления самих исходных данных. В контексте этой модели Map преобразует наборы исходных данных в элементы ИП путем применения отображения q, а Reduce складывает все эти фрагменты частичной информации в один элемент, представляющий все исходные данные, Рис. 1.

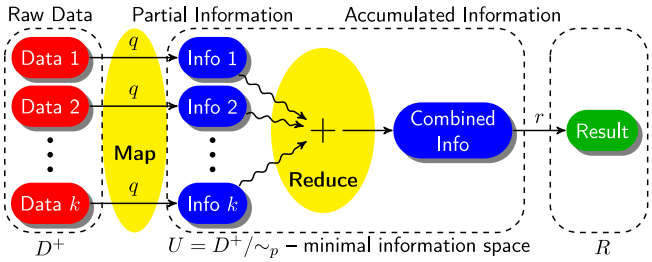

Рис. 1. Параллельная обработка распределенных данных с использованием наименьшего информационного пространства в модели MapReduce

При этом наименьшее информационное пространство определяет наиболее эффективную математическую структуру для представления информации, содержащейся в данных, и описывает «теоретический предел» компактности представления информации.

6. Заключение

Как показано в этой работе, проблема оптимизации распределенной обработки данных приводит к математическому представлению информации, содержащейся в данных, как элементу наименьшего ИП. При этом в терминах ИП естественным образом выражаются сложение и качество информации.

Понятие информации всегда было предметом преимущественно теоретического интереса. Сейчас проблематика больших данных требуют компактных, эффективных и хорошо организованных форм представления информации. Такие идеальные формы могут отражать самую суть информации, содержащейся в данных. Поэтому изучение таких форм и их свойств может приблизить нас к адекватному математическому описанию самого понятия информации.

Благодарности

Работа выполнена при финансовой поддержке РФФИ, грант № 19-29-09044.

Литература

[1] Golubtsov, P. Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras / P. Golubtsov // Advances in Intelligent Systems and Computing, Springer. – 2020. – Vol. 1127. – P. 274–298.

[2] Колмогоров, А.Н. Три подхода к определению понятия “количество информации / А.Н. Колмогоров // Пробл. передачи информ. – 1965. – Том 1, № 1. – С. 3–11.

[3] Dean, J. MapReduce: simplified data processing on large clusters / J. Dean, S. Ghemawat // Comm. of the ACM. – 2008. – Vol. 51(1). – P. 107–113.