Оптимизация распределенной обработки больших данных: Алгебраические основы и понятие информации

П. В. Голубцов*

Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра математики
Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2

(Поступила в редакцию 20.07.2023; подписана в печать 10.08.2023)

Рассматривается алгебраическая формализации распределенной обработки больших данных. Определяется понятие информационного пространства для заданной процедуры обработки данных и устанавливается критерий его минимальности. Доказывается существование минимального информационного пространства, обеспечивающего самую компактную форму представления информации, содержащейся в данных, и позволяющего наиболее эффективно распараллелить их обработку. Элемент этого пространства согласованным образом описывает информацию, содержащуюся в соответствующем наборе данных. Показано, что в терминах информационного пространства естественным образом выражаются понятия сложения информации и качества информации, отражающие интуитивные представление о самом понятии информации. Также рассматриваются преимущества использования минимального информационного пространства в модели распределенной обработки данных MapReduce. В контексте этой модели Map преобразует наборы исходных данных в элементы информационного пространства, а Reduce объединяет все эти фрагменты частичной информации в один элемент, представляющий все исходные данные. В качестве иллюстрации анализируется несколько примеров процедур обработки данных и описывается структура соответствующих минимальных информационных пространств.

PACS: 07.05.Kf, 89.70.-a. УДК: 519.722, 004.627, 519.254, 519.237.
Ключевые слова: большие данные, параллельная обработка, информационное пространство, алгебра информации, качество информации, MapReduce.

ВВЕДЕНИЕ

Данные в современных исследованиях нередко имеют огромный объём, 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 на многих компьютерах, извлекать из каждого набора исходных данных некоторую промежуточную компактную «информацию», постепенно объединять ее и, наконец, использовать накопленную информацию для получения окончательного результата. По мере поступления новых данных он должен иметь возможность добавлять содержащуюся в них информацию к накопленной и, в конечном итоге, обновлять результат.

Соответствующие технологии обработки данных получили мощный толчок с разработкой и реализацией модели распределённых вычислений MapReduce [1], которая активно используется, в частности, для масштабного анализа научных данных [2]. Наиболее популярной реализацией модели MapReduce является Hadoop [3]. Впоследствии архитектура систем распределённых вычислений получила новые реализации, например, Spark [4] и Flink [5]. Их технические особенности позволяют существенно повысить эффективность распределённых вычислений. В то же время, математические подходы, позволяющие исследовать и оптимизировать параллельные реализации алгоритмов, всё ещё находятся в зачаточном состоянии. Как правило, параллельные алгоритмы, используемые в таких системах, разрабатываются на основе эвристических соображений. Тем не менее, отметим работы [6, 7] и [8], в которых подчеркивается важность алгебраической формализации распределённых алгоритмов больших данных и, в частности, особую роль таких алгебраических структур, как моноиды.

В работах [9–13] были рассмотрены различные конкретные типы задач обработки данных и исследованы возникающие в них специальные виды представления информации, содержащейся в данных. Было показано, что для эффективной обработки распределённых данных ключевую роль играет возможность введения специальной промежуточной формы представления информации, обладающей определёнными алгебраическими свойствами. В рассмотренных задачах были введены соответствующие информационные пространства и исследованы их свойства.

Предлагаемый в данной статье подход подводит общий фундамент под эти исследования путаем построения алгебраической формализации распределённой обработки данных. Определяется понятие информационного пространства для заданной процедуры обработки и, в частности, минимального информационного пространства, реализующего максимально компактную форму представления информации и, как следствие, позволяющего наиболее эффективно распараллелить обработку данных. При этом в терминах информаци-

* golubtsov@physics.msu.ru

онного 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 [14], однако все они определяют меру количества информации в том или ином контексте. Напротив, минимальное информационное пространство приводит к понятию именно информации, содержащейся в данных. Его элементы реализуют максимально компактное представление набора данных, обеспечивающее тот же результат обработки что и этот набор. В результате, информация, извлеченная из данных, полностью заменяет эти данные.

1. 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, причём 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. Такая алгебраическая структура известна как свободный моноид

Заметим, однако, что поскольку результат обработки обычно не должен зависеть от порядка поступления данных, более удобно описывать 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 элементов складываются).

Формально мультимножество x удобно отождествлять отображением $x : D \to \mathbb{Z}_+$. Тогда $x(d)$ характеризует кратность элемента $d \in D$ в мультимножестве x, а $|x| = \sum_{d\in D} x(d)$ определяет число элементов (мощность) мультимножества x. Такое суммирование корректно (и $|x| < \infty$) тогда и только тогда, когда число элементов $d \in D$ для которых $x(d) > 0$ конечно. Таким образом,

$$D^+ = \big\{x : D \to \mathbb{Z}_+ \;\big|\; |x| < \infty\big\}.$$

Операция *сложения* мультимножеств $x, y \in D^+$ описывается суммой соответствующих отображений: $(x + y)(d) = x(d) + y(d)$ для любого $d \in D$. Говорят что мультимножество x *содержится* в y, $x \subseteq y$ если кратности элементов в x не превосходят кратностей этих элементов в y, т.е. $x(d) \leqslant y(d)$ для всех $d \in D$. Заметим, что $x \subseteq y \iff \exists! z : y = x + z$.

Для конечных мультимножеств также будем использовать обозначение типа $x = \{d_1, \ldots, d_n\}$, считая, что среди элементов $d_1, \ldots, d_n \in D$ могут быть одинаковые, а порядок элементов не имеет значения. Очевидно, в данном случае $|x| = n$.

Теперь мы можем формально определить понятие процедуры обработки данных.

Определение 1 Процедура обработки с наборами данных из множества данных D и результатами из множества R определяется как отображение p из свободного коммутативного моноида D^+ в множество R, т.е. $p : D^+ \to R$.

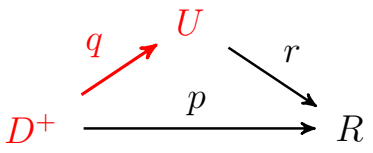
Для обработки большого числа распределённых наборов данных с использованием процедуры p, необходимо сначала собрать все эти наборы в одном месте в виде одного большого набора и применить к нему процедуру обработки. Чтобы избежать необходимости передачи и хранения потенциально огромных объёмов данных, возникает потребность модифицировать исходный алгоритм обработки, а именно, сначала подвергнуть данные предварительной обработке, чтобы уменьшить занимаемый ими объём, а потом осуществлять передачу и накопление данных в «сжатом» виде. Однако, при таком «сжатии» данных не должно происходить потери информации, а именно, должна иметься возможность получить из «сжатых» данных такой же результат обработки, какой даёт исходная процедура на оригинальных данных. Для математического описания этого подходы введем понятие информационного пространства.

Напомним, что *коммутативный моноид* $(U, +, 0)$ это множество U с коммутативной и ассоциативной бинарной операцией $+$ и нейтральным элементом 0, т.е. имеют место свойства:

$$x + y = y + x, \qquad (x + y) + z = x + (y + z),$$
$$x + 0 = x \qquad \forall x, y, z \in U.$$

Пусть $(U, +, 0)$ и $(U', +, 0)$ — два моноида. Отображение $h : U \to U'$ называется *гомоморфизмом* если $h(x + y) = h(x) + h(y)$ и $h(0) = 0$.

Определение 2 Информационное пространство (U, q, r) для процедуры $p : D^+ \to R$ — это коммутативный моноид U, сюръективный гомоморфизм $q : D^+ \to U$ и отображение $r : U \to R$, такие что $r \circ q = p$.



Иными словами, процедура обработки p пропускается через моноид U посредством гомоморфизма q. На приведённой выше диаграмме коммутативные моноиды и гомоморфизмы между ними выделены красным.

Фактически, гомоморфизм q «сжимает» исходные данные при сохранении возможности получить тот же результат обработки (с помощью отображения r), что и исходная процедура обработки p. Таким образом, «сжатие» данных происходит без потери информации. Гомоморфность этого отображения означает, что сумме наборов данных отвечает сумма соответствующих фрагментов информации, а его сюръективность обеспечивает отсутствие в U «лишних» элементов, которые никогда не могут возникнуть.

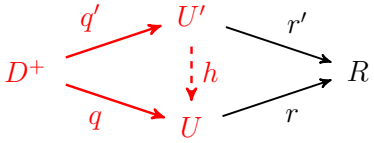
Заметим, что само множество наборов данных D^+, точнее, тройка (D^+, id, p), где $\mathrm{id} : D^+ \to D^+$ — тождественный гомоморфизм, также является информационным пространством для p. Будем называть его *исходным информационным пространством*.

2. МИНИМАЛЬНОЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО

Эффект от использования информационного пространства определяется тем, насколько сильно оно позволяет «сжать» данные. Чем «меньше» информационное пространство, тем лучшее «сжатие» информации оно обеспечивает.

Определение 3 Пусть (U, q, r) и (U', q', r') два информационных пространства для процедуры $p : D^+ \to R$. Будем говорить, что информационное пространство (U, q, r) *меньше* (лучше), чем (U', q', r') и обозначать это как $(U, q, r) \ll (U', q', r')$, если существует отображение $h : U' \to U$ такое, что $h \circ q' = q$ и $r \circ h = r'$.

Поскольку q — сюръективный гомоморфизм, такое *преобразование* информационных пространств h единственно и также является сюръективным гомоморфизмом. Иными словами, $(U, q, r) \ll (U', q', r')$ если существует единственный гомоморфизм h, для которого диаграмма коммутативна.



В определении 3 требуется коммутативность левого и правого треугольников диаграммы. Однако, согласно следующей теореме, это условие может быть ослаблено, а именно, достаточно коммутативности лишь одного из треугольников.

Теорема 1 *Следующие условия эквивалентны:*

1. $(U, q, r) \ll (U', q', r')$.
2. *Существует отображение* $h : U' \to U$ *такое, что* $q = h \circ q'$ *(*q *пропускается через* q'*).*
3. *Существует сюръективный гомоморфизм* $h : U' \to U$ *для которого* $r \circ h = r'$*, т.е.* (U, h, r) *является информационным пространством для* $r' : U' \to R$.

Отношение $\ll$ является предпорядком (рефлексивно и транзитивно), причём если $U \ll U'$ и $U' \ll U$, то эти информационные пространства изоморфны. Очевидно, исходное информационное пространство (D^+, id, p) является максимальным для p. Особый интерес представляет *минимальное* в смысле этого упорядочения информационное пространство (U, q, r), обеспечивающее максимальное сжатие информации. Оно обладает тем свойством, что любое информационное пространство (U', q', r') для p факторизуется через него, т.е. существует (единственный) сюръективный гомоморфизм $h : U' \to U$ такой что $h \circ q' = q$ и $r' = r \circ h$.

Для доказательства существования минимального информационного пространства дадим следующее

Определение 4 Будем говорить, что элементы x и y из коммутативного моноида U *неразличимы* относительно отображения $r : U \to R$ и обозначать это как $x \sim_r y$, если

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

Теорема 2 *Отношение* $\sim_r$ *является* конгруэнцией*, т.е. отношением эквивалентности, согласованным с алгебраической структурой на* U*:*

$$x \sim_r x' \;\&\; y \sim_r y' \implies x + y \sim_r x' + y'.$$

Из этого сразу следует, что множество классов эквивалентности $U/{\sim_r} = \left\{ [x]_{\sim_r} \;\middle|\; x \in U \right\}$ является коммутативным моноидом, в котором $[x]_{\sim_r} + [y]_{\sim_r} = [x+y]_{\sim_r}$, $0 = [0]_{\sim_r}$ (факторизация по конгруэнции порождает фактормоноид) и каноническое отображение $h : U \to U/{\sim_r}$, определённое как $h(x) = [x]_{\sim_r}$, является гомоморфизмом.

Теорема 3 (Существование) *Минимальное информационное пространство для процедуры* $p : D^+ \to R$ *существует и с точностью до изоморфизма совпадает с фактормоноидом* $(D^+/{\sim_p}, q, r)$ *по конгруэнции неразличимости на* D^+ *относительно* p*. При этом* $q : D^+ \to D^+/{\sim_p}$ *— соответствующий канонический гомоморфизм,* $q(x) = [x]_{\sim_p}$*, а отображение* $r : D^+/{\sim_p} \to R$ *определяется как* $r([x]_{\sim_p}) = p(x)$ *для любого* $x \in D^+$.

Во многих практических задачах (см., напр. [9, 10, 12, 13]) анализ процедуры обработки нередко позволяет предложить естественный вариант удобного информационного пространства. Следующее утверждение позволяет установить его минимальность.

Теорема 4 (Критерий минимальности) *Информационное пространство* (U, q, r) *является минимальным тогда и только тогда, когда все его элементы различимы относительно конгруэнции* $\sim_r$.

Отметим, что в минимальном информационном пространстве, различные наборы данных, содержащие

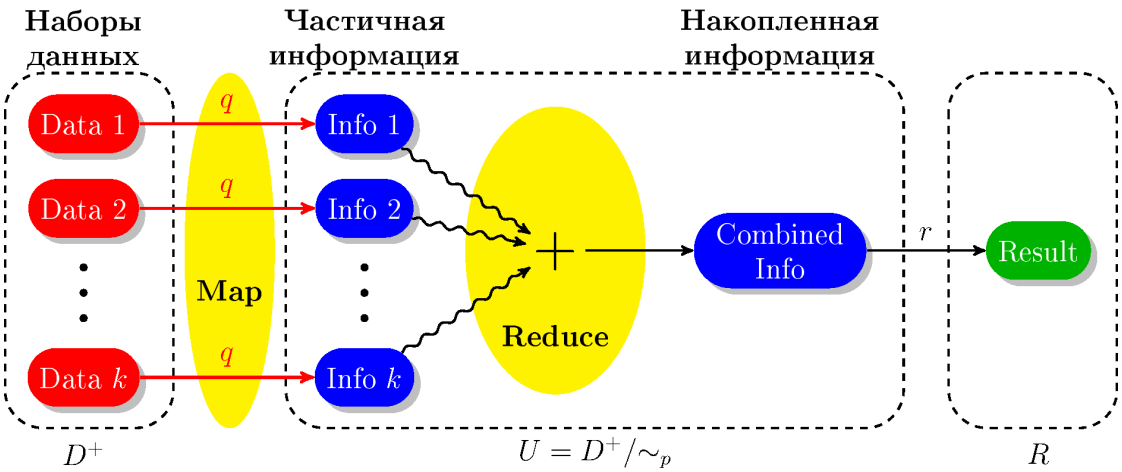


Рис. 1. Параллельная обработка распределённых данных с использованием минимального информационного пространства в модели MapReduce

одинаковую информацию относительно процедуры p (т.е. неразличимые относительно $\sim_p$), представляются одним и тем же элементом. Поэтому мы можем с полным правом считать, что элемент минимального информационного пространства представляет не что иное, как информацию, содержащуюся в соответствующем наборе данных. Это приводит к возможности математического определения информации, содержащейся в данных, как элемента минимального информационного пространства.

3. КАЧЕСТВО ИНФОРМАЦИИ

Алгебраическая структура информационного пространства позволяет естественным образом определить понятие *качества информации*.

Определение 5 Для элементов x и y информационного пространства U будем говорить, что x *лучше* (представляет больше информации), чем y и обозначать $x \succcurlyeq y$ если $\exists z \in U\ x = y + z$.

Это определение отражает то, что больший (в смысле включения) набор данных содержит больше информации.

Отношение качества на информационном пространстве U является отношением предпорядка, согласованным с алгебраической структурой, т.е.

$$x' \succcurlyeq x\ \&\ y' \succcurlyeq y \implies x' + y' \succcurlyeq x + y, \qquad x \succcurlyeq 0.$$

Более того, если $h : U' \to U$ — преобразование информационных пространств, то h сохраняет упорядочение качества: $x \succcurlyeq y \implies h(x) \succcurlyeq h(y)$.

4. НАКОПЛЕНИЕ ИНФОРМАЦИИ В МОДЕЛИ MAPREDUCE

Минимальное информационное пространство обеспечивает наиболее экономичную форму хранения информации, содержащейся в данных, а соответствующая бинарная операция позволяет единообразно складывать фрагменты информации, полученные из разных источников. В результате этого, использование минимального информационного пространства позволяет максимально эффективно распараллеливать процесс накопления информации в рамках модели распределённого анализа данных MapReduce [1] и организовать эффективную обработку без необходимости передачи и накопления самих исходных данных. В контексте этой модели Map преобразует наборы исходных данных в элементы информационного пространства, а Reduce объединяет все эти фрагменты частичной информации в один элемент, представляющий все исходные данные, рис. 1.

В процессе обработки из каждого набора исходных данных — элементы моноида D^+ извлекается максимально компактная информация — элемент минимального информационного пространства $D^+/\sim_p$; далее все эти элементы частичной информации складываются и получается информация, представляющая все наборы данных; из которой вычисляется окончательный результат обработки — элемент множества R.

Отметим, что минимальное информационное пространство определяет «оптимальную» математическую структуру для представления информации, содержащейся в данных, и описывает «теоретический предел» компактности представления информации.

5. ПРИМЕРЫ

В данном разделе мы рассмотрим три примера процедуры обработки и соответствующие минимальные

информационные пространства. Доказательство минимальности этих пространств опирается на теорему 4.

5.1. Среднее

Пожалуй простейшим примером процедуры обработки является вычисление среднего для набора чисел. Благодаря своей прозрачности, он позволяет легко проиллюстрировать рассмотренные выше понятия.

Итак, пусть $D = R = \mathbb{R}$ и

$$p\big(\{x_1, \dots, x_n\}\big) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

где $x_1, \dots, x_n \in \mathbb{R}$. Тогда минимальное информационное пространство (U, q, r) определяется моноидом

$$U = \big\{(n, S) \;\big|\; n \in \mathbb{N}, S \in \mathbb{R}\big\} \cup \{(0, 0.)\}$$

с операцией сложения

$$(n, S) + (n', S') = (n + n', S + S'),$$

гомоморфизмом

$$q\big(\{x_1, \dots, x_n\}\big) = \left(n, \sum_{i=1}^{n} x_i\right)$$

и отображением

$$r(n, S) = \frac{S}{n}.$$

То, что (U, q, r) является информационным пространством для p, легко проверяется, а его минимальность доказывается с помощью теоремы 4.

Таким образом, независимо от объёма набора данных $\{x_1, \dots, x_n\}$, вся информация о нём, необходимая для вычисления среднего, представляется парой чисел, натурального n и вещественного S. Дополнительный элемент $(0, 0.)$ является нейтральным в U. Он описывает отсутствие информации и отвечает пустому набору данных.

Несложно проверить, что частичный порядок, описывающий качество информации,

$$(n, S) \succcurlyeq (n', S') \iff n > n' \;\vee\; (n, S) = (n', S'),$$

т.е., чем больше элементов в наборе, тем лучше информация.

5.2. Медиана

Пусть, как и в предыдущем примере, $D = R = \mathbb{R}$ и $p\big(\{x_1, \dots, x_n\}\big) = \text{медиана}\big(\{x_1, \dots, x_n\}\big)$.

Оказывается, относительно такой процедуры обработки все элементы исходного (максимального) информационного пространства $(\mathbb{R}^+, \mathrm{id}, p)$ различимы и следовательно, по теореме 4, оно является также и минимальным. Это означает, что для процедуры вычисления медианы, никакое более компактное представление информации, чем исходные данные, принципиально невозможно. Таким образом, с точки зрения возможности компактного представления информации и распараллеливания, процедуры вычисления среднего и медианы представляют диаметрально противоположные примеры.

Поскольку минимальное информационное пространство совпадает с исходным, частичный порядок, описывающий качество информации, имеет тривиальный вид:

$$\{x_1, \dots, x_n\} \succcurlyeq \{y_1, \dots, y_m\} \iff \\ \iff \{x_1, \dots, x_n\} \supseteq \{y_1, \dots, y_m\},$$

т.е. второй набор просто содержится в первом как подмультимножество.

5.3. Оптимальное линейное оценивание

Рассмотрим линейное измерение [15] неизвестного вектора x из евклидова пространства $\mathcal{D}$

$$y = Ax + \nu.$$

Здесь $A : \mathcal{D} \to \mathcal{R}$ — линейное отображение, $\nu \in \mathcal{R}$ — случайный вектор с $\mathsf{E}\nu = 0$ и корреляционным оператором $S > 0$, а $y \in \mathcal{R}$ — результат измерения.

Оптимальная линейная оценка вектора x даётся выражением [15]

$$\widehat{x} = p(y, A, S) = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y.$$

Рассмотрим теперь серию независимых линейных измерений одного и того же вектора x

$$y_i = A_i x + \nu_i, \qquad i = 1, \dots, n,$$

где $y_i \in \mathcal{R}_i$ — результаты измерений, а пространства $\mathcal{R}_i$ могут различаться; $A_i : \mathcal{D} \to \mathcal{R}_i$; $\nu_i \in \mathcal{R}_i$ — случайные векторы с корреляционными операторами $S_i > 0$. Вся исходная информация об i-том измерении представляется тройкой (y_i, A_i, S_i), а множество данных D представляет собой множество всех троек такого вида.

Как показано в [10], оптимальная линейная оценка для набора данных $\big\{(y_1, A_1, S_1), \dots, (y_n, A_n, S_n)\big\} \in D^+$ даётся выражением

$$\widehat{x} = p\Big(\big\{(y_1, A_1, S_1), \dots, (y_n, A_n, S_n)\big\}\Big) = \\ = \left(\sum_{i=1}^{n} A_i^* S_i^{-1} A_i\right)^{-1} \sum_{i=1}^{n} A_i^* S_i^{-1} y_i.$$

Отсюда несложно получить, что в качестве «естественного» информационного пространства удобно взять моноид

$$U = \big\{(u, T) \;\big|\; T : \mathcal{D} \to \mathcal{D},\; T \geqslant 0,\; u \in \mathrm{im} T\big\}$$

с операцией сложения

$$(u, T) + (u', T') = (u + u', T + T')$$

и нейтральным элементом $(0_{\mathcal{D}}, 0_{\mathcal{D}\to\mathcal{D}})$.

Гомоморфизм, переводящий набор данных в соответствующий элемент информационного пространства определяется как

$$q\Big(\{(y_1, A_1, S_1), \ldots, (y_n, A_n, S_n)\}\Big) = \\ = \left(\sum_{i=1}^{n} A_i^* S_i^{-1} y_i,\ \sum_{i=1}^{n} A_i^* S_i^{-1} A_i\right),$$

а отображение, строящее оценку $\widehat{x}$ на основе информации (u, T), определяется как

$$\widehat{x} = r(u, T) = T^{-1} u.$$

Теорема 4 позволяет установить, что это информационное пространство (U, q, r) является минимальным.

Частичный порядок, описывающий качество информации, определяется как

$$(u, T) \succcurlyeq (u', T') \iff T - T' \geqslant 0 \ \wedge\ u - u' \in \mathrm{im}(T - T').$$

Заметим, что в случае нормальных распределений компоненты u и T имеют интересный теоретико-статистический смысл. Вектор u является минимальной достаточной статистикой, а оператор T представляет информационную матрицу Фишера [16] для результата измерения (и достаточной статистики u), см. напр., [17]. Как известно, матрица Фишера описывает количество (возможно, правильнее сказать, качество) информации, содержащейся в измерении. Таким образом, информация (u, T) в данном контексте представляет собой минимальную достаточную статистику плюс детальную характеризацию её информативности.

ЗАКЛЮЧЕНИЕ

Как показано в данной работе, при весьма общих предположениях, проблема оптимизации распределенной обработки данных приводит к математическому представлению информации, содержащейся в данных, как элементу минимального информационного пространства. При этом в терминах информационного пространства естественным образом выражаются сложение и качество информации.

Понятие информации всегда было предметом преимущественно теоретического интереса. Однако, в последнее время, потребность эффективно манипулировать огромными распределенными массивами данных выдвигает новые требования к осмыслению и формализации понятия информации. Бурное развитие проблематики больших данных приводит к необходимости построения 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.

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

[1] *Dean J., Ghemawat S.* // Communications of the ACM. **51**, N 1. 107. (2008).
[2] *Ekanayake J., Pallickara S., Fox G.* MapReduce for Data Intensive Scientific Analyses // Fourth IEEE International Conference on eScience – Indianapolis, IN. 277. (2008).
[3] *White T.* Hadoop: The Definitive Guide. O'Reilly, 2015.
[4] *Ryza S., Laserson U., Owen S., Wills J.* Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly, 2015.
[5] *Hueske F., Kalavri V.* Stream Processing with Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications. O'Reilly, 2019.
[6] *Lin J.* Monoidify! Monoids as a Design Principle for Efficient MapReduce Algorithms // arXiv:1304.7544 2013.
[7] *Parsian M.* Data Algorithms. Chapter 28. MapReduce and Monoids. O'Reilly, 2015.
[8] *Fegaras L.* // Journal of Functional Programming. **27**, E27. (2017).
[9] *Голубцов П.В.* Понятие информации в контексте задач обработки больших данных // НТИ Сер. 2. Информационные процессы и системы. №1. 31. (2018).
[10] *Голубцов П.В.* Задача линейного оценивания и информация в системах больших данных // НТИ Сер. 2. Информационные процессы и системы. № 3. 23.(2018).
[11] *Golubtsov P.* Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras // Advances in Intelligent Systems and Computing. Springer, 2020. **1127**. Pp. 274–298.
[12] *Golubtsov P.* Information Spaces for Big Data Problems in Fuzzy Bayesian Decision Making // Lecture Notes on Data Engineering and Communications Technologies. Springer, 2022. **121**. Pp. 102–114.
[13] *Golubtsov P.* Information Spaces and Efficient Information Accumulation in Calibration Problems. Lecture Notes on Data Engineering and Communications Technologies. Springer, 2023. **158**. Pp. 53–62.
[14] *Колмогоров А.Н.* // Пробл. передачи информ. **1**, № 1. 3. (1965).
[15] *Пытьев Ю.П.* // Мат. сб. **118**, № 5. 19. (1982).
[16] *Барра Ж.-Р.* Основные понятия математической статистики. М., 1974.
[17] *Пытьев Ю.П.* Методы математического моделирования измерительно-вычислительных систем. М., 2012.

Optimizing big data distributed processing: Algebraic foundations and the concept of information

P. V. Golubtsov

Department of Mathematics, Faculty of Physics Lomonosov Moscow State University. Moscow 119991, Russia
E-mail: golubtsov@physics.msu.ru

An algebraic formalization of distributed processing of big data is considered. The concept of information space is defined for a given data processing procedure and a criterion for its minimality is established. The existence of a minimal information space is proved, which provides the most compact form of representation of the information contained in the data and allows the most efficient parallelization of data processing. An element of this space describes in a consistent way the information contained in the corresponding data set. It is shown that in terms of the information space, the concepts of information addition and information quality are naturally expressed, reflecting the intuitive idea of the very concept of information. The advantages of using the minimal information space in the MapReduce distributed data processing model are also considered. In the context of this model, Map transforms the original data sets into information space elements, and Reduce combines all these pieces of partial information into a single element representing all the original data. By way of illustration, several examples of data processing procedures are analyzed and the corresponding minimal information spaces are presented.

PACS: 07.05.Kf, 89.70.-a.
Keywords: big data, parallel processing, information space, information algebra, information quality, MapReduce.
Received 20 July 2023.

Сведения об авторе

Голубцов Петр Викторович — докт. физ.-мат. наук, доцент, профессор; тел.: (903) 580-13-39, e-mail: golubtsov@physics.msu.ru.