Параллельная распределенная обработка данных и информационные пространства

П. В. Голубцов*
Московский государственный университет имени М. В. Ломоносова, физический факультет, кафедра математики
Россия, 119991, Москва, Ленинские горы, д. 1, стр. 2
(Статья поступила 25.06.2018; Подписана в печать 10.09.2018)

Данные в современных исследованиях нередко имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В таких случаях 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 результата. По мере поступления новых данных он должен иметь возможность добавлять их к накопленной информации и, при необходимости, обновлять результат. В работе рассмотрены особенности такой хорошо организованной промежуточной формы информации и ее естественные алгебраические свойства. В качестве примера исследована проблема трансформации процедуры оптимального линейного оценивания так, чтобы отдельные фрагменты исходных данных могли обрабатываться независимо и параллельно. Предложена каноническая форма информации, позволяющая алгоритму извлекать такую информацию параллельно из каждого набора исходных данных, объединять ее и использовать для получения результата. Показано, что на построенном информационном пространстве кроме алгебраической структуры также индуцируется согласованное с ней упорядочение, отражающее понятие качества информации.

PACS: 07.05.Kf, 89.70.-a. УДК: 519.254.
Ключевые слова: большие данные, каноническая информация, распределенные системы сбора и обработки данных, линейное оценивание, алгебра информации, информационное пространство.

ВВЕДЕНИЕ

Особый интерес к параллельной и распределенной обработке данных [1, 2] в последнее время стимулируется бурно развивающейся проблематикой больших данных (Big Data). С одной стороны, это связано с изобилием рутинно собираемых данных. С другой, было обнаружено, что большие объемы данных могут содержать ценную информацию, возможность извлечения которой из такого рода данных ранее даже и не предполагалась. Много интересных примеров можно найти в [3]. Можно сказать, что в задачах больших данных, как правило, речь идет об извлечении спрятанной информации и представлении ее в форме, пригодной для интерпретации или принятия решений. Такого рода процессы обычно проходят через несколько стадий, в которых информация извлекается из исходных данных, преобразуется, передается, накапливается и, в конце концов, трансформируется к удобному для интерпретации виду.

Отметим, использование термина «информация», в последнее время заметно возросло, особенно, в контексте анализа данных. Обычно он понимается слишком широко и неформально. Однако, по мнению автора, такая возросшая частота употребления этого термина свидетельствует о возрастающей потребности в более точном и формальном понимании феномена информации. Как будет показано ниже, проблематика больших данных наталкивает на новые подходы к понятию информации.

Исследования, связанные с системы больших данных, нацелены на проблемы обработки больших объемов распределенных данных и имеют, как правило, ярко выраженную практическую и техническую направленность. В то же время, основная масса исследований по теории информации проводится в контексте теории вероятностной и математической статистики и представляет преимущественно теоретический интерес.

Пожалуй, наиболее прикладная часть теории информации, берущая начало в работах Шеннона, связана с передачей сообщений при наличии помех [4–6]. При этом речь идет не столько о «смысле» или «качестве» информации, сколько о ее количестве. Особое место в математической статистике занимает информация Фишера, описываемая матрицами [7, 8]. Она обеспечивает более детальное отражение понятия информации и, в частности обладает важной аддитивной структурой, в рамках которой объединению независимых статистик отвечает сумма их информационных матриц. Несмотря на многочисленные исследования по теории информации, проблема формализации этого понятия, отражающей именно смысл информации, содержащейся в данных, представляется еще далекой от удовлетворительного решения. В связи с этим, упомянем работы [9–11], в которых вместо определения информации, содержащейся в данных, исследуется информативность систем, преобразующих данные. В рамках такого подхода естественным образом возникает

*E-mail: golubtsov@physics.msu.ru

алгебраическая структура на классе источников информации и частичный порядок, позволяющий сравнивать их информативность.

На данный момент сферы интересов больших данных и различных подходов к понятию информации слабо пересекаются. Однако, как уже было отмечено выше, проблематика больших данных требует более четкого, формального описания самого понятия информации и информационных процессов. Это необходимо для построения эффективных инструментов преобразования информации, опирающихся на математические (например, алгебраические) свойства информации. В связи с этим, по мнению автора, большие данные станут в ближайшее время основным двигателем и потребителем общей теории информации. В настоящей работе мы попытаемся показать, как некоторая формализация понятия информации и ее алгебраические свойства могут следовать просто из рассмотрения задачи в контексте больших данных.

Чем же выделяются задачи «больших данных» на фоне задач анализа данных? Данные в таких задачах, как правило, имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате даже самый простой анализ больших данных сталкивается с серьезными трудностями. Действительно, традиционные подходы к обработке информации предполагают, что данные, предназначенные для обработки, собираются в одном месте, организуются в виде удобных структур (например, матриц), и только тогда соответствующий алгоритм обрабатывает эти структуры и выдает результат анализа. В случае больших данных невозможно собрать все данные, необходимые для исследовательского проекта на одном компьютере. Более того, это было бы непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. В результате возникает необходимость в трансформации существующих алгоритмов, приводящих к их «распараллеливанию», или даже разработке новых подходов к обработке данных, которые по самой формулировке проблемы смогли бы обрабатывать отдельные фрагменты данных независимо и параллельно. Соответствующий алгоритм анализа данных должен, параллельно работая на многих компьютерах, извлекать из каждого набора исходных данных некоторую промежуточную компактную «информацию», постепенно объединять и обновлять ее и, наконец, использовать накопленную информацию для получения результата. По мере поступления новых фрагментов данных он должен иметь возможность добавлять их к накопленной информации и, по мере необходимости, обновлять результат.

В работе [12] в общих чертах рассматривалась специфика обработки информации в распределенных системах. Здесь мы обсудим особенности такой хорошо организованной промежуточной формы информации и выявим ее естественные алгебраические свойства. В качестве примера, мы исследуем задачу линейного оценивания с точки зрения распределенных систем сбора и обработки информации. Подходы, развиваемые в этой статье, могут быть полезны для многих прикладных задач, например, при сборе и обработке информации в крупномасштабных экспериментах, где данные собираются в многочисленных исследовательских центрах, разбросанных по всему Земному шару. Однако для нас основной интерес представляют особенности информационного пространства, возникающего при необходимости распределенной обработки данных. На построенном информационном пространстве естественным образом порождается алгебраическая структура, описывающая композицию отдельных фрагментов информации, и согласованное с ней отношение предпорядка, отражающее феномен качества информации. Мы также увидим, такая специальная форма представления информации в некотором смысле отражает саму суть информации, содержащейся в данных. Это приводит нас к совершенно новому подходу к самому понятию информации.

1. ТРАНСФОРМАЦИЯ МЕТОДОВ ОБРАБОТКИ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ

1.1. Особенности обработки информации в системах больших данных

Выделим следующие особенности задач обработки информации в системах больших данных:

1. Как правило, речь идет об огромных объемах данных.
2. Такие данные обычно не собраны воедино, а распределены по многочисленным, возможно, достаточно удаленным компьютерам.
3. Постоянно могут возникать новые данные, которые необходимо оперативно включать в обработку.

Традиционные методы обработки обычно не учитывают такую специфику и требуют серьезного пересмотра при необходимости их применения в задачах больших данных.

1.2. Неприменимость традиционного подхода обработки данных в распределенных системах

Рассмотрим бегло и предельно упрощенно стандартный подход к обработке данных. К задачам такого рода относятся задачи оценивания, принятия решений, обучения, классификации, и т.п. Обычно в задачах с малым фиксированным набором данных обработка состоит в применении некоторого отображения (алгоритма, метода), определяющего обработку, к набору данных и получению результата обработки (например, оценки некоторой величины), рис. 1.

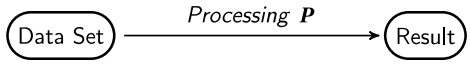

Рис. 1: Стандартный подход к обработке данных

Важным условием здесь является то, что все данные находятся в одном месте и готовы к применению к ним отображения обработки, например, представлены в виде подходящих структур, например, числовых массивов. Если же данные распределены по многим различным локациям, для применения обработки их требуется сначала собрать в одном месте, организовать комбинированные данные в виде подходящих структур, и применить к ним алгоритм обработки (рис. 2). Двойными пунктирными стрелками здесь и далее обозначается передача больших объемов данных в исходном или частично обработанном виде.

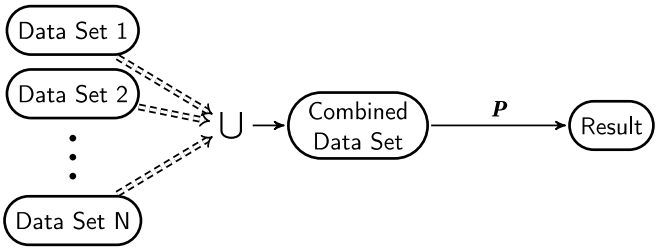

Рис. 2: Стандартный подход к обработке распределенных данных

Недостатки такого подхода к обработке распределенных данных достаточно очевидны:

1. Передача больших объемов исходных данных создаст чрезмерный траффик.
2. Хранение полного набора данных в одном месте потребует огромных объемов памяти.
3. Обработка всех данных на одном компьютере потребует чрезмерных вычислительных и временных ресурсов.
4. По мере поступления новых данных, комбинированный набор данных будет расти и, как следствие требовать постоянно возрастающих (потенциально бесконечных) ресурсов для хранения.
5. При этом, при поступлении новых данных, алгоритм обработки будет необходимо по новой применять к постоянно увеличивающемуся объему данных.

1.3. Выделение промежуточной информации в процессе обработки

Рассмотрим следующую модификацию процесса обработки, которая позволит преодолеть обозначенные выше недостатки. Предположим, что полный алгоритм обработки $\boldsymbol{P}$ допускает разбиение на две фазы $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$ (рис. 3):

1. $\boldsymbol{P}_1$ — выделение из исходных данных некоторой промежуточной информации.
2. $\boldsymbol{P}_2$ — вычисление результата на основании выделенной промежуточной информации.

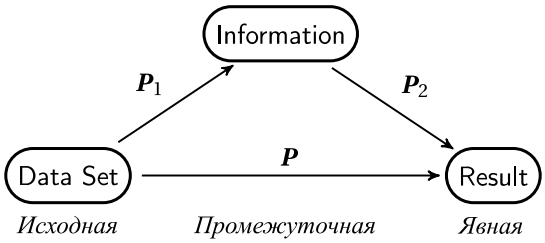

Рис. 3: Разбиение процесса обработки данных на две фазы

Выбор подходящей промежуточной формы представления информации определяется рассматриваемой задачей обработки данных. Будем называть некоторую выбранную форму представления промежуточной информации канонической формой информации или короче, **канонической информацией**.

В определенном смысле узлы диаграммы на рис. 3 отражают представления информации в разных формах:

1. Data Set — информация в сырой (исходной) форме.
2. Result — информация в явной (удобной для интерпретации) форме.
3. Information — информация в промежуточной (удобной для обработки) канонической форме.

Ниже мы более подробно обсудим желательные свойства канонической информации. Но сейчас отметим, что такая форма представления информации должна быть полна, то есть содержать всю необходимую для вычисления результата информацию (в этом и состоит коммутативность диаграммы на рис. 3) и компактна, то есть иметь минимально возможный размер, в идеале, не зависящий от объема представленных данных.

1.4. Композиции фрагментов канонической информации

Кроме того, будем предполагать существование операции композиции (сложения) отдельных фрагментов канонической информации, отвечающее объединению отдельных наборов исходных данных, рис. 4. Одиночными пунктирными стрелками обозначается передача компактных фрагментов канонической информации.

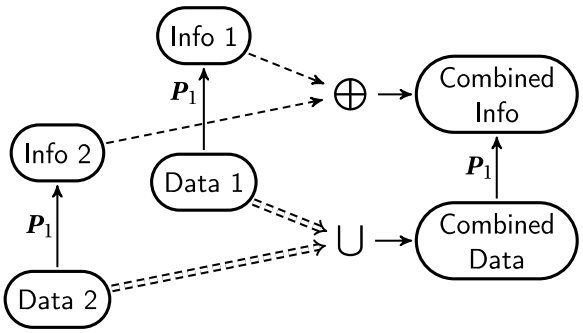

Рис. 4: Соответствие композиции фрагментов канонической информации и объединения наборов исходных данных

Это можно записать как $\boldsymbol{P}_1(D_1) \bigoplus \boldsymbol{P}_1(D_2) = \boldsymbol{P}_1(D_1 \cup D_2)$, где под $D_1 \cup D_2$ понимается объединение двух наборов данных в один.

1.5. Модифицированная схема обработки данных

Рассмотрим, как может быть трансформирована схема обработки, за счет предварительного выделения канонической информации из каждого набора исходных данных, рис.5.

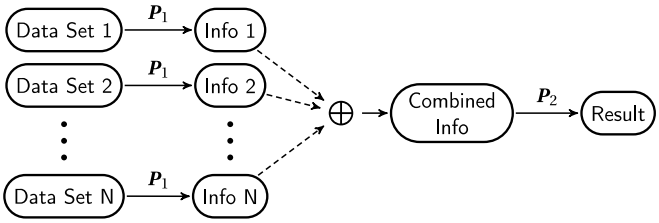

Рис. 5: Модифицированная схема обработки распределенных данных

Такая схема позволяет преодолеть все недостатки стандартной схемы обработки распределенных данных, отмеченные выше:

1. Передаются лишь компактные фрагменты выделенной промежуточной информации.

2. Хранение комбинированной информации требует небольших объемов памяти, возможно, таких же, как и объемы, требуемые для хранения отдельных частей промежуточной информации.

3. Промежуточная информация выделяется параллельно из каждого отдельного набора данных (фаза $\boldsymbol{P}_1$). Если основная часть обработки сосредоточена в первой фазе, то вторая фаза $\boldsymbol{P}_2$, состоящая в построении результата по компактной накопленной информации, не потребует серьезных вычислительных и временных ресурсов.

4. По мере поступления новых данных, потребуется лишь выделить из них промежуточную информацию и «добавить» ее к накопленной.

5. При этом, алгоритм обработки будет необходимо снова применять к компактной информации фиксированного объема.

Отметим, что благодаря введению специальной промежуточной формы представления информации, появляется возможность избавиться от необходимости накопления больших объемов исходных данных в одном месте и обеспечить высокую степень параллелизации и масштабирования. В результате этого, схема обработки распределенных данных, представленная на рис. 5, органично «вписывается» в архитектуру систем распределенного хранения и анализа данных, таких как, например, Hadoop MapReduce [13–16] или Spark [17].

2. ЗАДАЧА ОПТИМАЛЬНОГО ЛИНЕЙНОГО ОЦЕНИВАНИЯ

2.1. Линейный эксперимент

В качестве примера намеченного выше подхода, исследуем возможность подобного распараллеливания в задаче оптимального линейного оценивания.

Рассмотрим схему линейного эксперимента [18, 19] вектора $x \in \mathcal{D}$ вида

$$y = Ax + \nu, \tag{1}$$

где $y \in \mathcal{R}$ — результат измерения, $A : \mathcal{D} \to \mathcal{R}$ — линейное отображение, описывающее искажения измерительной системы, и $\nu \in \mathcal{R}$ — случайный вектор шума с нулевым средним $\mathrm{E}\nu = 0$ и заданным ковариационным оператором $\mathrm{D}\nu = S > 0$. Ковариационный оператор случайного вектора $\mu \in \mathcal{R}$ является многомерным обобщением понятия дисперсии и определяется как $(\mathrm{D}\mu)(z) = \mathrm{E}(\mu - \mathrm{E}\mu, z)(\mu - \mathrm{E}\mu)$ для любого $z \in \mathcal{R}$ и является самосопряженным положительно полуопределенным оператором, которому отвечает ковариационная матрица координат вектора μ в ортонормированном базисе.

2.2. Оптимальное линейное оценивание

Задача линейного оценивания вектора x состоит в построении такого линейного отображения $R : \mathcal{R} \to \mathcal{D}$, что оценка $\hat{x} = Ry$ максимально близка к x, а именно, R доставляет минимум функционалу

$$H(R) = \sup_{x \in \mathcal{D}} \mathrm{E}\|Ry - x\|^2 .$$

Эта задача имеет решение [18, 19] тогда и только тогда, когда отображение A невырождено, т.е.,

$\mathcal{N}(A) = \{0\}$. (Ядро и образ линейного отображения $A : \mathcal{D} \to \mathcal{R}$, будем обозначать $\mathcal{N}(A) \subseteq \mathcal{D}$ и $\mathcal{R}(A) \subseteq \mathcal{R}$). При этом оценка

$$\hat{x} = Ry = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y$$

является несмещенной и обладает наименьшим ковариационным оператором [20] (Частичный порядок на пространстве $\mathbb{S}_{\mathcal{D}}$ всех самосопряженных операторов, действующих в $\mathcal{D}$, определяется следующим образом: $\tilde{Q} \geq Q \;\Leftrightarrow \tilde{Q} - Q \geq 0$.)

$$Q = \mathrm{D}\hat{x} = RSR^* = \left(A^* S^{-1} A\right)^{-1}.$$

Отсюда, в частности, следует, что оценка $\hat{x} = Ry$ обладает минимальными дисперсиями координат $\hat{x}_j$ в некотором ортонормированном базисе: $\mathrm{D}\hat{x}_j = Q_{jj}$ и $\mathrm{E}\|\hat{x} - x\|^2$ достигает минимального значения $\mathrm{E}\|\hat{x} - x\|^2 = \mathrm{tr}Q$.

Итак, пусть заданы результат измерения y и модель измерения (A, S). Тогда исходные данные для линейного оценивания представляются тройкой (y, A, S) и процедура обработки $\boldsymbol{P}$ состоит в преобразовании исходных данных в результат оценивания: $\hat{x}$ — оптимальную оценку вектора x, рис. 6.

$$(y, A, S) \xrightarrow{\boldsymbol{P}} \hat{x} = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y$$

Рис. 6: Оптимальное линейное оценивание для одного набора данных

При этом преобразование $\boldsymbol{P}$ определено не всюду, а лишь тогда, когда отображение A невырождено (и, следовательно, $A^* S^{-1} A$ обратим).

2.3. Линейное оценивание в случае многих независимых измерений

Пусть теперь имеется много независимых измерений одного и того же неизвестного вектора $x \in \mathcal{D}$:

$$y_i = A_i x + \nu_i, \quad \mathrm{D}\nu_i = S_i, \quad i = 1, \ldots, n, \qquad (2)$$

где $y_i \in \mathcal{R}_i$ — результаты измерений, $A_i : \mathcal{D} \to \mathcal{R}_i$ — линейные отображения, и $\nu_i \in \mathcal{R}_i$ — независимые случайные векторы с нулевыми средними $\mathrm{E}\nu_i = 0$ и ковариационными операторами $\mathrm{D}\nu_i = S_i : \mathcal{R}_i \to \mathcal{R}_i$. В общем случае пространства измерений $\mathcal{R}_i$ могут быть различными.

Чтобы воспользоваться результатом предыдущего раздела в случае серии измерений (2), представим эту серию в виде одного измерения вида (1) [21], где

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{R}, \quad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} : \mathcal{D} \to \mathcal{R}, \qquad (3)$$
$$S = \begin{pmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_n \end{pmatrix} : \mathcal{R} \to \mathcal{R},$$

$$\mathcal{R} = \mathcal{R}_1 \times \mathcal{R}_2 \times \cdots \times \mathcal{R}_n,$$

$$\dim \mathcal{R} = \sum_{i=1}^{n} \dim \mathcal{R}_i \ .$$

Таким образом, при наличии n независимых измерений (2) потребуется собрать соответствующие данные в одном месте, реорганизовать их в виде блочных матриц, возможно, очень больших размерностей и применить к объединенным данным отображение $\boldsymbol{P}$, рис. 7.

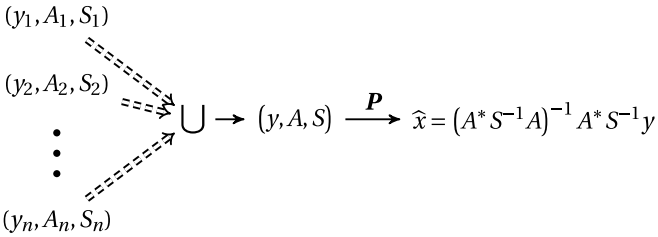

Рис. 7: Стандартная схема линейного оценивания для большого числа измерений

При большом числе измерений размерность объединенных данных может стать крайне большой, в результате чего данный подход может оказаться практически нереализуемым. Кроме того, добавление новых данных будет приводить к увеличению размерностей объединенных данных, что, в свою очередь, будет требовать все больше ресурсов для их хранения и обработки (применения преобразования $\boldsymbol{P}$).

2.4. Распараллеливание обработки за счет выделения канонической информации

Покажем, что обработку данных в задаче линейного оценивания можно разбить на две фазы $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, где первая фаза $\boldsymbol{P}_1$ состоит в выделении некоторой компактной промежуточной информации из исходных данных, а вторая $\boldsymbol{P}_2$ вычисляет результат оценивания на основании этой промежуточной информации. При этом, нашей целью будет найти такую факторизацию, что применение преобразования $\boldsymbol{P}_1$ к объединенному набору данных может быть заменено параллельным

применением $\boldsymbol{P}_1$ к отдельным данным и последующему «сложению» полученных фрагментов информации.

Пусть (y_i, A_i, S_i) $i = 1, \ldots, n$, образуют наборы данных, обеспечиваемах измерениями (3). Рассмотрим результат оценивания, отвечающий объединенному набору данных (y, A, S), где y, A и S определяются выражениями (3). Заметим, что для вычисления оптимальной оценки вектора x, $\hat{x} = \left(A^* S^{-1} A\right)^{-1} A^* S^{-1} y$, требуется вычислять выражения вида $A^* S^{-1} A$ и $A^* S^{-1} y$. Несложно убедиться, что

$$A^* S^{-1} y = A_1^* S_1^{-1} y_1 + \cdots + A_n^* S_n^{-1} y_n,$$

$$A^* S^{-1} A = A_1^* S_1^{-1} A_1 + \cdots + A_n^* S_n^{-1} A_n.$$

Это означает, что вся необходимая для дальнейшей обработки информация, относящаяся к -му измерению может быть представлена парой (v_i, T_i), где

$$v_i = A_i^* S_i^{-1} y_i \in \mathcal{D}, \quad T_i = A_i^* S_i^{-1} A_i : \mathcal{D} \to \mathcal{D},$$

и T_i — неотрицательно определенный оператор. При этом, объединенным данным будет отвечать пара $(v, T) = \bigoplus_{i=1}^n (v_i, T_i)$, в которой $v = \sum_{i=1}^n v_i$ и $T = \sum_{i=1}^n T_i$.

2.5. Каноническое информационное пространство

Будем называть пару $(v, T) = \left(A^* S^{-1} y, \ A^* S^{-1} A\right)$ **канонической информацией** для данных (y, A, S), а множество $\mathfrak{I}$ всех таких пар каноническим **информационным пространством** для задачи линейного оценивания вектора из пространства $\mathcal{D}$. Заметим, что $\mathcal{R}\left(A^* S^{-1}\right) = \mathcal{R}\left(A^* S^{-1} A\right) = \mathcal{N}^{\perp}(A)$ [18, 19]. Следовательно, измерениям вида (y, A, S) могу отвечать лишь такие пары (v, T), в которых $v \in \mathcal{R}(T)$. Таким образом,

$$\mathfrak{I} = \left\{(v, T) \ \mid \ T \in \mathbb{S}_{\mathcal{D}}^+, \ v \in \mathcal{R}(T)\right\}$$

где $\mathbb{S}_{\mathcal{D}}^+$ - множество неотрицательно определенных операторов на $\mathcal{D}$ — выпуклый конус в линейном пространстве $\mathbb{S}_{\mathcal{D}}$ самосопряженных операторов на пространстве $\mathcal{D}$. Если $\dim \mathcal{D} = m$ то $\dim \mathbb{S}_{\mathcal{D}} = \frac{m(m+1)}{2}$. Тогда $\mathfrak{I} \subset \mathcal{D} \times \mathbb{S}_{\mathcal{D}}^+$ представляет собой выпуклый конус в $\frac{m(m+3)}{2}$-мерном линейном пространстве $\mathcal{D} \times \mathbb{S}_{\mathcal{D}}$. Отсюда, в частности, следует, что любой элемент информационного пространства $\mathfrak{I}$ может быть задан $\frac{m(m+3)}{2}$ числами.

Очевидно, процесс линейного оценивания можно разбить на две фазы $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, где первая фаза $\boldsymbol{P}_1$ состоит в построении канонической информации:

$$(v, T) = \boldsymbol{P}_1 (y, A, S) = \left(A^* S^{-1} y, \ A^* S^{-1} A\right),$$

а вторая $\boldsymbol{P}_2$ вычисляет результат оценивания

$$\hat{x} = \boldsymbol{P}_2 (v, T) = T^{-1} v,$$

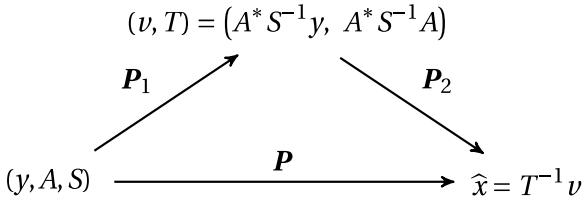

Рис. 8: Разбиение процесса обработки данных на две фазы

на основании этой информации (рис. 8).

Как было показано выше, объединению исходных данных (y_1, A_1, S_1) и (y_2, A_2, S_2) отвечает композиция соответствующих элементов канонической информации (v_1, T_1) и (v_2, T_2), определенная как

$$(v_1, T_1) \bigoplus (v_2, T_2) = (v_1 + v_2, T_1 + T_2).$$

Это можно записать как $\boldsymbol{P}_1 (y_1, A_1, S_1) \bigotimes \boldsymbol{P}_1 (y_2, A_2, S_2) = \boldsymbol{P}_1((y_1, A_1, S_1) \cup (y_2, A_2, S_2))$, где под $(y_1, A_1, S_1) \cup (y_2, A_2, S_2)$ понимается определяемое выражением (3), объединение двух наборов данных в один, рис. 9.

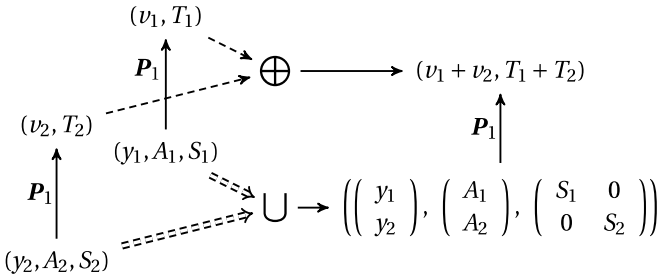

Рис. 9: Соответствие композиции фрагментов канонической информации и объединения наборов исходных данных

2.6. Распараллеливание обработки за счет использования канонической информации

В результате факторизации алгоритма $\boldsymbol{P}$ на две фазы и введения канонической информации, схема обработки распределенных данных, представленная на рис. 7 может быть трансформирована следующим образом (рис. 10). Из каждого отдельного фрагмента (y_i, A_i, S_i) данных выделяется каноническая информация (v_i, T_i), которая впоследствии объединяется и используется для вычисления результата оценивания.

$$\hat{x} = \boldsymbol{P}_2 (v, T) = \boldsymbol{P}_2 \left(\bigoplus\nolimits_{i=1}^n \boldsymbol{P}_1 (v_i, T_i)\right) = \\ = \boldsymbol{P}_2 \left(\bigoplus\nolimits_{i=1}^n \boldsymbol{P}_1 (y_i, A_i, S_i)\right).$$

Отметим основные особенности такой модифицированной схемы:

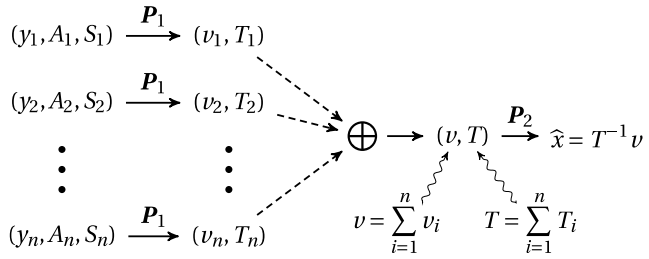

Рис. 10: Модифицированная схема обработки распределенных данных

1. Объем памяти, требуемый для хранения информации в каноническом виде не зависит от объема представляемых исходных данных и составляет $\frac{m(m+3)}{2}$ чисел (m-мерный вектор и самосопряженный оператор в m-мерном пространстве).

2. Выделение канонической информации (v_i, T_i) из i-того набора данных (преобразование $\boldsymbol{P}_1$) может проводиться на тех компьютерах, где эти данные находятся, причем, параллельно и независимо.

3. Передаются лишь компактные фрагменты выделенной канонической информации одинакового объема.

4. Сложение частей канонической информации максимально упрощено и определяется покомпонентным сложением пар (v_i, T_i).

5. Ресурсоемкость второй фазы $\boldsymbol{P}_2$, состоящей в построении результата по компактной накопленной информации (v, T), определяется только размерностью m пространства неизвестных и не зависит от объема исходных данных.

6. По мере поступления новых данных, потребуется лишь выделять из них каноническую информацию и «добавлять» ее к накопленной. При этом окончательную обработку $\boldsymbol{P}_2$ будет необходимо снова применять к компактной информации фиксированного объема.

В результате, распределенность исходных данных способствует повышению эффективности обработки за счет естественного распараллеливания алгоритма.

В рассмотренной выше задаче линейного оценивания нашей целью было построение оценки $\hat{x}$, т.е., $\boldsymbol{P}(y, A, S) = \hat{x}$. При построении оценки $\hat{x}$ важно также охарактеризовать ее точность, исчерпывающую информацию о которой содержит ковариационный оператор $Q = \mathrm{D}\hat{x} = \left(A^* S^{-1} A\right)^{-1} = T^{-1}$. В частности, он позволяет определить погрешности оценивания отдельных компонент вектора x (в произвольном ортонормированном базисе) $\mathrm{E}(\hat{x}_j - x_j)^2 = \mathrm{D}\hat{x}_j = Q_{jj}$ и полную погрешность оценивания $\mathrm{E}\|\hat{x} - x\|^2 = \mathrm{tr}Q = \sum_{j=1}^{m} Q_{jj}$. Таким образом, каноническая информация вида (v, T) подходит также для построения результатов оценивания $\boldsymbol{P}(y, A, S)$ в виде $(\hat{x},\ \mathrm{D}\hat{x})$, $\left(\hat{x},\ \{\mathrm{D}\hat{x}_j\}_{j=1,\ldots,m}\right)$ или $\left(\hat{x},\ \mathrm{E}\|\hat{x} - x\|^2\right)$. В этих случаях отображение $\boldsymbol{P}_1$ остается прежним, а отображение $\boldsymbol{P}_2$ заменяется, соответственно, на $\boldsymbol{P}_2(v, T) = \left(T^{-1}v,\ T^{-1}\right)$, $\boldsymbol{P}_2(v, T) = \left(T^{-1}v,\ \left\{\left(T^{-1}\right)_j\right\}_{j=1,\ldots,m}\right)$ или $\boldsymbol{P}_2(v, T) = \left(T^{-1}v,\ \mathrm{tr}T^{-1}\right)$.

2.7. Качество информации и информативность источников информации

Как было отмечено выше, ковариационный оператор $Q = \mathrm{D}\hat{x} = \left(A^* S^{-1} A\right)^{-1} = T^{-1}$ полностью характеризует точность оценивания. При этом, чем меньше ковариационный оператор оценки $\hat{x}$ тем меньше погрешность оценивания: т.е., если $Q \le \tilde{Q}$ то $Q_{jj} \le \tilde{Q}_{jj}$ и $\mathrm{tr}Q \le \mathrm{tr}\tilde{Q}$.

Будем говорить, что информация (v, T) не хуже (не менее точна), чем $\left(\tilde{v}, \tilde{T}\right)$ если $T \ge \tilde{T}$ и писать $(v, T) \succcurlyeq \left(\tilde{v}, \tilde{T}\right)$. Если $(v, T) \succcurlyeq \left(\tilde{v}, \tilde{T}\right)$ и $\left(\tilde{v}, \tilde{T}\right) \succcurlyeq (v, T)$, то будем говорить, что (v, T) и $\left(\tilde{v}, \tilde{T}\right)$ имеют одинаковую точность и обозначать это $(v, T) \approx \left(\tilde{v}, \tilde{T}\right)$. Очевидно, это равносильно условию $T = \tilde{T}$. Нетрудно видеть, что более точная информация обеспечивает более точное оценивание. Действительно, пусть $T \ge \tilde{T}$ и (v, T) и $\left(\tilde{v}, \tilde{T}\right)$ позволяют построить соответствующие оценки, т.е., T и $\tilde{T}$ обратимы. Согласно [18] отсюда следует, что $T^{-1} \le \tilde{T}^{-1}$ и значит, $Q \le \tilde{Q}$, где Q и $\tilde{Q}$ — ковариационные операторы соответствующих оценок.

Как мы отмечали, построение оценки вектора x возможно лишь если $\mathcal{N}(A) = \{0\}$. Если $\mathcal{N}(A) \ne \{0\}$ то часть вектора x, принадлежащая $\mathcal{N}(A)$, обнуляется и не может быть оценена. Однако, как показано в [18, 19] можно оценить проекцию Px вектора x на подпространство $\mathcal{N}^{\perp}(A)$. Здесь $P = A^- A : \mathcal{D} \to \mathcal{D}$ — ортогональный проектор на $\mathcal{N}^{\perp}(A)$, а $A^- : \mathcal{R} \to \mathcal{D}$ — линейное отображение, псевдообратное к $A : \mathcal{D} \to \mathcal{R}$ [18, 21]. При этом оптимальная оценка вектора Px и ее ковариационный оператор определяются выражениями

$$\widehat{Px} = \left(A^* S^{-1} A\right)^- A^* S^{-1} y, \quad \mathrm{D}\left(\widehat{Px}\right) = \left(A^* S^{-1} A\right)^-,$$

или, в терминах канонической информации, $\widehat{Px} = T^- v$, $\mathrm{D}\left(\widehat{Px}\right) = T^-$. Более того, поскольку $\mathcal{N}(T) = \mathcal{N}\left(A^* S^{-1} A\right) = \mathcal{N}(A)$, проектор P также может быть выражен через T, $P = T^- T$.

Таким образом, рассматриваемая нами каноническая информация содержит всю информацию, необходимую

для построения оптимальной оценки и в этом более широком контексте.

Пусть информация (v,T) не хуже, чем $\left(\tilde{v},\tilde{T}\right)$, то есть $T \geq \tilde{T} \geq 0$. Тогда $\mathcal{N}(T) \subseteq \mathcal{N}\left(\tilde{T}\right)$ и $\tilde{T}^{-} \geq \tilde{P}T^{-}\tilde{P}$ [16], где $\tilde{P} = \tilde{T}^{-}\tilde{T}$ — ортогональный проектор на $\mathcal{N}^{\perp}\left(\tilde{T}\right)$. Включение $\mathcal{N}(T) \subseteq \mathcal{N}\left(\tilde{T}\right)$ влечет $\mathcal{N}^{\perp}\left(\tilde{T}\right) \subseteq \mathcal{N}^{\perp}(T)$, т.е., информация (v,T) позволяет оценить большую часть вектора x, чем $\left(\tilde{v},\tilde{T}\right)$, а неравенство $\tilde{T}^{-} \geq \tilde{P}T^{-}\tilde{P}$ означает, что информация (v,T) обеспечивает более точное оценивание вектора $\tilde{P}x$.

В работе [22] было введено понятие качества модели (A,S) для измерения типа (1). Считается, что качество модели (A,S) не ниже, чем $\left(\tilde{A},\tilde{S}\right)$, если

$$A^{-}A \geq \tilde{A}^{-}\tilde{A} \quad \text{и} \quad \tilde{A}^{-}\tilde{A}\left(A^{*}S^{-1}A\right)^{-}\tilde{A}^{-}\tilde{A} \leq \left(\tilde{A}^{*}\tilde{S}^{-1}\tilde{A}\right)^{-}. \tag{4}$$

Иными словами, лучшей модели отвечают более широкие возможности оценивания и при прочих равных условиях менее интенсивный шум оценки [22].

В [23] было рассмотрено понятие информативности моделей измерения (преобразователей информации). Согласно [23], информативность модели (A,S) не хуже, чем $\left(\tilde{A},\tilde{S}\right)$, если существует модель измерения (B,U), такая что последовательная композиция (A,S) и (B,U) дает $\left(\tilde{A},\tilde{S}\right)$. Последовательная композиция определяется выражением $(B,U)*(A,S) = (BA, U+BSB^{*})$ и описывает ситуацию, когда результат измерения с помощью (A,S) подвергается дальнейшему измерению с помощью (B,U). Таким образом, информативность модели (A,S) не хуже, чем $\left(\tilde{A},\tilde{S}\right)$ если

$$\exists B: \ \left(BA = \tilde{A} \ \ \& \ \ BSB^{*} \leq \tilde{S}\right). \tag{5}$$

В [23] было доказано что понятия качества модели и информативности эквивалентны. Из приведенных выше рассуждений следует, что условие (4), а следовательно, и (5) равносильны более простому условию $T \geq \tilde{T}$, т.е.,

$$A^{*}S^{-1}A \geq \tilde{A}^{*}\tilde{S}^{-1}\tilde{A}.$$

Таким образом, рассмотренное выше понятия точности информации, естественным образом появившееся ни информационном пространстве, приводит к такому же упорядочению на множестве моделей линейного измерения, как и понятие качества моделей измерений в [22, 24] или информативности преобразователей информации в[23].

2.8. Связь с достаточными статистиками и информационными матрицами

Заметим, что в случае нормальных распределений компоненты u и T канонической информации (u,T) имеют интересный теоретико-статистический смысл. Вектор u является *минимальной достаточной статистикой*, а оператор T представляет собой *информационную матрицу Фишера* [7, 8, 24] для измерения y (и достаточной статистики u), см. напр., [24]. Как известно, матрица Фишера описывает количество (возможно, правильнее сказать, качество) информации, содержащейся в измерении. Таким образом, каноническая информация (u,T) в данном контексте представляет собой минимальную достаточную статистику плюс детальную характеризацию ее информативности.

3. СВОЙСТВА КАНОНИЧЕСКОЙ ИНФОРМАЦИИ В ЗАДАЧЕ ЛИНЕЙНОГО ОЦЕНИВАНИЯ

Рассмотрим свойства канонического информационного пространства $\mathfrak{I}$, определенного выше. Эти свойства не только представляют самостоятельный интерес, но и могут выступать в качестве примера общих свойств информационных пространств, возникающих в задачах обработки больших объемов распределенных данных.

3.1. Существование

Любой исходный набор данных должен допускать представление информации в каноническом виде. В частности, минимальный «атомарный» набор данных или даже «пустой» набор должны быть представимы в канонической форме.

В нашем примере это условие выполнено. Как мы видели, информации, содержащейся в данных (y,A,S) может быть недостаточно для построения результата оценивания. А именно, если ядро отображения A нетривиально, т.е., $\mathcal{N}(A) \neq \{0\}$, то оценка неизвестного вектора не может быть построена. Тем не менее, каноническая информация (v,T) может быть построена для любых исходных данных. Отметим, что даже полное отсутствие измерений (несущее нулевую информацию) может быть представлено в каноническом виде. Формально, любое измерение (y,A,S), в котором $A = 0 : \mathcal{D} \to \mathcal{R}$ - нулевое отображение, не несет никакой информации об измеряемом векторе. Любому такому измерению отвечает каноническая информация $\mathbf{0} = (0,0)$, т.е. $v = 0 \in \mathcal{D}$ и $T = 0 : \mathcal{D} \to \mathcal{D}$.

3.2. Полнота

Каноническая форма должна содержать всю информацию, содержащуюся в исходных данных, а имен-

но, она должна приводить к тому же результату, что и исходные данные, из которых она получена. Формально в рассмотренном примере это означает что $\boldsymbol{P}(y,A,S)=\boldsymbol{P}_2(\boldsymbol{P}_1(y,A,S))$ для всех данных (y,A,S) из области определения преобразования $\boldsymbol{P}$. Это свойство напоминает понятие достаточности в математической статистике.

3.3. Операция композиции на информационном пространстве

На каноническом информационном пространстве $\mathfrak{I}$ определена операция композиции $\oplus$, описывающая сложение фрагментов информации, отвечающих данным. При этом $(\mathfrak{I},\oplus,\mathbf{0})$ является коммутативным моноидом, т.е., выполнены следующие свойства для любых $\mathbf{a},\mathbf{b},\mathbf{c}\in\mathfrak{I}$:

1. $\mathbf{a}\oplus\mathbf{b}=\mathbf{b}\oplus\mathbf{a}$ — коммутативность,
2. $(\mathbf{a}\oplus\mathbf{b})\oplus\mathbf{c}=\mathbf{a}\oplus(\mathbf{b}\oplus\mathbf{c})$ — ассоциативность,
3. $\mathbf{a}\oplus\mathbf{0}=\mathbf{a}$ — наличие нейтрального элемента.

Отметим, что моноид $(\mathfrak{I},\oplus,\mathbf{0})$ обладает также свойством сокращения:

4. $\mathbf{a}\oplus\mathbf{b}=\mathbf{a}\oplus\mathbf{c}\Rightarrow\mathbf{b}=\mathbf{c}$,

но не имеет обратимых элементов отличных от $\mathbf{0}$, т.е. не существует «отрицательной» информации.

Коммутативность и ассоциативность позволяют складывать фрагменты канонической информации произвольным образом, а наличие сократимости позволяет в любой момент «вычесть» из накопленной информации любую включенную ранее информацию, если впоследствии выяснится, что по тем или иным причинам соответствующее измерение было недостоверно.

3.4. Качество информации

На каноническом информационном пространстве $\mathfrak{I}$ определено отношение предпорядка $\succcurlyeq$, отражающее понятие точности информации, и обладающее следующими свойствами:

5. $\mathbf{a}\succcurlyeq\mathbf{a}$ — рефлексивность,
6. $\mathbf{a}\succcurlyeq\mathbf{b}\ \&\ b\succcurlyeq\mathbf{c}\ \Rightarrow\mathbf{a}\succcurlyeq\mathbf{c}$ — транзитивность.

Заметим, что отношение $\succcurlyeq$ не обладает свойством антисимметричности, т.е. не является частичным порядком. Действительно, из $(v,T)\succcurlyeq\left(\tilde{v},\tilde{T}\right)$ и $\left(\tilde{v},\tilde{T}\right)\succcurlyeq(v,T)$ следует лишь, что $T=\tilde{T}$, но не обязательно $v=\tilde{v}$. Однако, на классах эквивалентной точности это отношение антисимметрично, $\mathbf{a}\succcurlyeq\mathbf{b}\ \&\ b\succcurlyeq\mathbf{a}\Rightarrow\mathbf{a}\approx\mathbf{b}$ и, следовательно, является отношением частичного порядка.

Кроме того, алгебраическая структура информационного пространства согласована со структурой порядка, а именно:

7. $\mathbf{a}\succcurlyeq\mathbf{0}$. — Любая информация точнее, чем отсутствие информации.
8. $\mathbf{a}\oplus\mathbf{b}\succcurlyeq\mathbf{a}$, $\mathbf{b}$. — Композиция двух фрагментов информации точнее, чем каждый из них по отдельности.
9. $\mathbf{a}\succcurlyeq\mathbf{b}\ \&\ c\succcurlyeq\mathbf{e}\Rightarrow\mathbf{a}\oplus\mathbf{c}\succcurlyeq\mathbf{b}\oplus\mathbf{e}$. — Чем точнее фрагменты информации, тем точнее результат композиции.

3.5. Единственность

Любые исходные данные должны быть представимы в канонической форме единственным образом. Фактически, это свойство означает отсутствие избыточности в канонической информации. Отсюда, в частности, следует, что каноническая информация не должна зависеть от порядка данных в исходном наборе.

Несложно убедиться, что одной и той же канонической информацией (v,T) могут описываться различные данные (y,A,S). При этом, они будут приводить к одному и тому же результату оценивания. Однако, для каждого элемента исходных данных (y,A,S) существует единственное представление элементом пространства $\mathfrak{I}$, согласованное с операцией композиции и обеспечивающего соответствующий результат оценивания.

Отметим, наконец, два «практических» свойства данного способа представления промежуточной информации. Они носят скорее технический характер, связанный с особенностями реализации соответствующих алгоритмов.

3.6. Компактность

Информация, представленная в канонической форме, должна занимать небольшой (желательно минимальный) объем, по возможности, не зависящий от объема представленных данных.

В рассмотренном примере линейного оценивания, информация в канонической форме, занимает фиксированный объем $\frac{m(m+3)}{2}$ чисел и не зависящий от количества исходных измерений и их размерностей.

3.7. Эффективность

Представление промежуточной информации в канонической форме должно обеспечивать эффективное выполнение всех стадий обработки данных:

1. Извлечение канонической информации из исходных данных — преобразование $\boldsymbol{P}_1$. В рассмотренном примере требуется несколько матричных умножений для матриц, определяемых отдельными фрагментами данных. При этом извлечение канонической информации из отдельных фрагментов может производиться параллельно.

2. Комбинирование и накопление канонической информации — операция $\oplus$ в информационном пространстве. Сводится к сложению векторов и матриц фиксированной размерности и требует незначительных вычислительных ресурсов.

3. Вычисление результата на основании накопленной канонической информации - преобразование $\boldsymbol{P}_2$. В рассмотренной задаче эта операция требует решения системы линейных уравнений фиксированного размера $m \times m$ (или обращения соответствующей матрицы. Даже при постоянном поступлении новых данных обновление оценки может осуществляться лишь время от времени по мере необходимости.

ЗАКЛЮЧЕНИЕ

Отметим, что чисто техническая попытка «распараллелить алгоритм», фактически привела нас к необходимости нахождения специального вида представления информации, обладающему удобными алгебраическими свойствами. В некотором смысле, такое представление отражает саму суть информации, содержащейся в данных. Можно сказать, что сама потребность эффективно манипулировать огромными распределенными массивами данных выдвигает новые требования к осмыслению и формализации понятия информации.

В рассмотренном выше примере выбор канонической формы информации довольно очевиден. В общем случае выбор компактной промежуточной информации может быть не очевиден или даже не возможен. В связи с этим, представляется важным выявление класса задач, в которых возможно выделение достаточно компактной промежуточной информации и нахождение эффективных методов построения подходящих информационных пространств.

Несмотря на то, что задача линейного оценивания имеет самостоятельную ценность и часто встречается в приложениях, мы использовали ее, в первую очередь, в качестве иллюстрации. Мы показали, что, как и в [12], проблема адаптации алгоритма к работе в системах больших данных приводит к построению специального вида представления информации, обладающему естественными алгебраическими свойствами.

Во многих практических задачах преобразование обработки $\boldsymbol{P}$, трансформирующее исходные данные в окончательный результат обработки, имеет специфическое «происхождение», а именно, является решением некоторой оптимизационной задачи. В нашем случае рассматривалась задача построения оценки минимальной погрешности. Оптимизационная постановка исходной задачи, фактически, привела к тому, что понятие качества решения (точности оценки) индуцировало на информационном пространстве упорядочение, отражающее «качество» информации. Как показано в [23, 25–28], подобные естественное упорядочение и алгебраическая структура всегда возникают при исследовании информативности различных классов источников информации, включая, например, многозначные [29] и нечеткие [30–33]. Можно ожидать, что подобное упорядочение, согласованное с алгебраической структурой информационного пространства, всегда будет возникать в контексте задач оптимального принятия решений в распределенных системах.

[1] *Bekkerman R., Bilenko M., Langford J.* Scaling up machine learning: Parallel and distributed approaches. Cambridge University Press, 2012.
[2] *Fan J., Han F., Liu H.* National Science Review. 2013. **1**, N 2. P. 293.
[3] *Mayer-Schцnberger V., Cukier K.* Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York, Houghton Mifflin Harcourt, 2013.
[4] *Shannon C. E., Weaver W.* The Mathematical Theory of Communication. Univ of Illinois Press. 1949.
[5] *Яглом А. М, Яглом И. М.* Вероятность и информация. М.: Наука, 1973.
[6] *Стратонович Р. Л.* Теория информации. М.: Сов. радио, 1975.
[7] *Барра Ж.-Р.* Основные понятия математической статистики. М.: Мир, 1974.
[8] *Боровков А. А.* Математическая статистика. Оценка параметров, проверка гипотез. М.: Наука, 1984.
[9] *Golubtsov P. V.* Pattern Recognition and Image Analysis. 1991. **1**, N 1. P. 77.
[10] *Голубцов П. В.* Пробл. передачи информ. 1999. **35**, № 3. С. 109.
[11] *Golubtsov P. V.* Information Processes. 2002. **2**, N 1. P. 62.
[12] *Голубцов П. В.* НТИ Сер. 2. Информационные процессы и системы. 2018. № 1. С. 31.
[13] *White T.* Hadoop: The Definitive Guide. O'Reilly, 2015.
[14] *Dean J., Ghemawat S.* Communications of the ACM. 2008. **51**, N 1. P. 107.
[15] *Palit I., Reddy C. K.* IEEE Transactions on Knowledge and Data Engineering. 2012. **24**, N 10. P. 1904.
[16] *Ekanayake J., Pallickara S., Fox G.* MapReduce for Data Intensive Scientific Analyses // Fourth IEEE International Conference on eScience. 2008. P. 277.
[17] *Ryza S. Laserson U. Owen S. Wills J.* Advanced Analytics with Spark: Patterns for Learning from Data at Scale.

O'Reilly, 2015.
[18] *Пытьев Ю. П.* Мат. сб. 1982. **118**, № 5. С. 19.
[19] *Пытьев Ю. П.* Математические методы интерпретации эксперимента. М.: Высшая школа, 1989.
[20] *Голубцов П. В.* НТИ Сер. 2. Информационные процессы и системы. 2018. № 3. С. 23.
[21] *Алберт А.* Регрессия, псевдоинверсия и рекуррентное оценивание — М.: Наука, 1977.
[22] *Пытьев Ю. П.* Мат. сб. 1983. **120**, № 2. С. 240.
[23] *Голубцов П. В.* Пробл. передачи информ. 1992. **28**, № 2. С. 30.
[24] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. М.: Физматлит, 2012.
[25] *Голубцов П. В.* Пробл. передачи информ. 1998. **34**, № 3. С. 60.
[26] *Голубцов П. В.* Пробл. передачи информ. 1999. **35**, № 3. С. 109.
[27] *Golubtsov P. V.* Information Processes. 2002. **2**, N 1. P. 62.
[28] *Golubtsov P. V.* Hadronic Journal Supplement. 2004. **19**, N 4. P. 375.
[29] *Голубцов П. В., Филатова С. А.* Мат. моделирование. 1992. **4**, № 7. С. 79.
[30] *Pyt'ev Y. P.* Pattern Recognition and Image Analysis. 1993. **3**, N 2. P. 150.
[31] *Голубцов П. В.* Пробл. передачи информ. 1994. **30**, № 3. С. 47.
[32] *Пытьев Ю. П.* Интеллектуальные системы. 2004. **8**, № 1–4. С. 147.
[33] *Пытьев Ю. П.* Возможность как альтернатива вероятности. М. : Физматлит, 2007.

Parallel distributed data processing and information spaces

P.V. Golubtsov

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

The data in modern studies often have a huge volume, are distributed among numerous sites and are constantly replenished. In such cases, collecting all the research-related data on one computer is usually impossible and impractical, since one computer will not be able to process them in a reasonable time. A suitable algorithm for data analysis must, working in parallel on many computers, extract from each set of source data some intermediate compact "information gradually combine it and finally use the accumulated information to obtain the result. As new data arrive, it should be able to add them to the accumulated information and, if necessary, update the result. The paper considers the features of such a well-organized intermediate form of information and its natural algebraic properties. As an example, the problem of transforming the optimal linear estimation procedure has been investigated so that individual fragments of the original data can be processed independently and in parallel. A canonical form of information is proposed that allows the algorithm to extract such information in parallel from each set of source data, combine it and use it to obtain the result. It is shown that on the constructed information space, in addition to the algebraic structure, the compatible ordering, which reflects the concept of information quality, is also induced.

PACS: 07.05.Kf, 89.70.-a.
Keywords:big data, canonical information, distributed data collection and processing systems, linear estimation, information algebra, information space.
Received 25 June 2018.

Сведения об авторе
Голубцов Петр Викторович — доктор физ.-мат. наук, доцент, профессор; тел.: (495) 939-10-33, e-mail: golubtsov@physics.msu.ru.