

Оценка связанности представлений данных при проектировании мультимодальных аналитических процессов обработки больших данных

В.В. Чаругин, В.В. Чаругин, А.Н. Чесалин

Аннотация – В работе исследуется проблема проектирования мультимодальных аналитических процессов обработки больших данных с различной связанностью представлений данных. В зависимости от структуры процесса данные разных типов могут обрабатываться независимо, преобразовываться между представлениями или совместно использоваться при формировании результата. Для учета этих особенностей предлагается система показателей связанности представлений данных и процедура оценки, основанная на логическом правиле предварительного отбора и аддитивной оценке.

Проводится оценка связанности представлений данных на примерах аналитических процессов, описанных с использованием контрактно-графового подхода: первичной обработки данных промышленного мониторинга, предварительной проверки выполненного учебного задания и формирования медицинского отчета на основе результатов лабораторных исследований.

Предложенная процедура оценки применяется для определения необходимости согласования данных, операций и промежуточных результатов и может быть использована аналитиками данных, разработчиками аналитических систем и специалистами предметных областей при проектировании мультимодальных аналитических процессов.

Ключевые слова – большие данные, мультимодальность, аналитический процесс, представления данных, связанность представлений данных, показатели связанности, процедура оценки, согласование данных, машинное обучение, глубокое обучение.

I. ВВЕДЕНИЕ

Современные информационные системы работают с постоянно растущими объемами разнородных данных, поступающих из множества источников. Такие данные отличаются по структуре, семантике и способам представления. Для их анализа применяются мультимодальные аналитические процессы, позволяющие организовать последовательное и параллельное выполнение операций обработки, извлечения признаков, преобразования и совместного использования данных.

При проектировании таких процессов возникает необходимость учитывать связи между различными представлениями данных. В одних случаях данные

разных типов обрабатываются независимо, и достаточно задать последовательность операций. В других случаях данные преобразуются между представлениями, используются совместно на промежуточных этапах и влияют на формирование итогового результата. Для таких процессов требуется учитывать модальности, допустимые преобразования, условия передачи данных и взаимодействие операций.

В данной работе рассматривается оценка связанности представлений данных при проектировании мультимодальных аналитических процессов обработки больших данных. Вводятся показатели связанности, предлагается процедура их оценки и демонстрируется ее использование на аналитических процессах различной структуры. Результаты оценки позволяют определить необходимость согласования данных, операций, промежуточных результатов и условий их использования. В качестве примера используется подход, основанный на контрактах модальностей и графовом описании процесса.

II. ПРОБЛЕМА ПРОЕКТИРОВАНИЯ АНАЛИТИЧЕСКИХ ПРОЦЕССОВ С РАЗЛИЧНОЙ СВЯЗАННОСТЬЮ ПРЕДСТАВЛЕНИЙ ДАННЫХ

Обработка больших данных в информационных системах предполагает работу с информацией, поступающей из разных источников и отличающейся форматом, структурой, семантикой и способом представления [1]. В рамках одного аналитического процесса такие данные могут проходить несколько этапов преобразования, использоваться совместно и участвовать в формировании общего результата. Сложность процесса определяется не только объемом входной информации, но и характером связей между ее представлениями.

При построении аналитических процессов требуется согласование входных и выходных данных операций. В простейшем случае оно сводится к проверке формата, типа или структуры передаваемых данных. Такая проверка позволяет определить, может ли результат одной операции быть передан на вход другой операции без нарушения требований к представлению данных. Для простых последовательных процессов или независимых ветвей анализа такой проверки, как правило, достаточно.

В мультимодальных аналитических процессах данные могут переходить из одной формы представления в другую, объединяться с данными других модальностей, последовательно передаваться между несколькими моделями и участвовать в формировании общего результата анализа [2], [3].

Требуется проверять не только совместимость форматов и структур данных, но и корректность преобразований, пригодность промежуточных результатов для последующих этапов и соответствие условиям решаемой аналитической задачи. Согласование данных в таких процессах должно учитывать несколько аспектов их использования. К ним относятся модальность, форма представления данных, допустимые операции преобразования, ограничения на использование промежуточных результатов, условия готовности операций к выполнению, а также возможность совместного использования нескольких представлений. Такое описание связей между представлениями данных позволяет рассматривать аналитический процесс как единую структуру обработки, в которой операции связаны не только порядком выполнения, но и преобразованиями данных между ними.

Возникает проблема определения условий, при которых согласования входных и выходных данных только по формату, типу или структуре оказывается недостаточно. Для ее решения необходимо выделить показатели, позволяющие оценить связанность представлений данных и определить, когда аналитический процесс требует учета зависимостей между данными, операциями, преобразованиями и ограничениями их применения.

III. ПОКАЗАТЕЛИ СВЯЗАННОСТИ ПРЕДСТАВЛЕНИЙ ДАННЫХ

Для оценки связанности представлений данных в аналитическом процессе вводится система показателей, характеризующих структуру процесса, характер взаимодействия данных и условия выполнения операций. Объектом оценки является аналитический процесс, включающий входные данные, операции обработки, промежуточные результаты, связи между операциями и итоговый результат [4].

Показатели связанности используются для описания признаков аналитического процесса, влияющих на необходимость согласования данных не только по формату, типу и структуре, но и с учетом их роли в последующей обработке. Они делятся на две группы: основные и дополнительные.

Основные показатели отражают признаки, определяющие наличие связанности представлений данных на уровне структуры аналитического процесса [5]. Они представлены в таблице 1.

Таблица 1 – Основные показатели связанности представлений данных

Обозначение	Наименование	Назначение
O_1	Показатель разнородности данных	Характеризует наличие в процессе данных, отличающихся способом представления
O_2	Показатель преобразования представлений	Характеризует наличие операций, изменяющих форму представления данных
O_3	Показатель	Характеризует совместное

	совместного использования данных	использование нескольких представлений или промежуточных результатов
--	----------------------------------	--

Дополнительные показатели используются для уточнения оценки с учетом требований к проверке корректности, применения моделей искусственного интеллекта и зависимости операций от промежуточных результатов [6]. Они представлены в таблице 2.

Таблица 2 – Дополнительные показатели связанности представлений данных

Обозначение	Наименование	Назначение
D_1	Показатель предварительной проверки корректности	Характеризует необходимость проверки связей между операциями до выполнения процесса
D_2	Показатель применения моделей искусственного интеллекта	Характеризует наличие в аналитическом процессе моделей машинного и глубокого обучения
D_3	Показатель условий готовности операций	Характеризует наличие операций, выполнение которых зависит от поступления нескольких входных данных или выполнения нескольких условий
D_4	Показатель сложности выполнения операций	Характеризует необходимость учета состояний операций в ходе выполнения процесса

Выделенные показатели задают основу для последующей оценки аналитического процесса и позволяют перейти от общего описания его структуры к обоснованию необходимости согласования данных, операций и промежуточных результатов при проектировании. Основные показатели используются для предварительного отбора процесса, а дополнительные – для уточнения условий применения такого согласования.

IV. ПРОЦЕДУРА ОЦЕНКИ СВЯЗАННОСТИ ПРЕДСТАВЛЕНИЙ ДАННЫХ

Для использования введенных показателей задается процедура оценки, позволяющая определить уровень связанности представлений данных и обосновать необходимость согласования данных, операций и промежуточных результатов для рассматриваемого аналитического процесса. Процедура основана на логическом правиле предварительного отбора по основным показателям и аддитивной оценке дополнительных показателей.

Каждому показателю присваивается бинарное значение:

$$O_i \in \{0, 1\}, i = 1, \dots, 3, D_j \in \{0, 1\}, j = 1, \dots, 4$$

где O_i – основные показатели, D_j – дополнительные показатели. Значение 1 соответствует выполнению показателя, значение 0 – его невыполнению.

На первом этапе осуществляется проверка необходимых условий, позволяющих выявить наличие

связанности представлений данных и определить, требуется ли дальнейшая оценка условий согласования данных при проектировании аналитического процесса. Показатель O_1 рассматривается как предусловие, задающее наличие нескольких представлений данных. Показатели O_2 и O_3 характеризуют наличие операций, связанных с преобразованием или совместным использованием представлений данных, и учитываются только при выполнении данного предусловия [7]–[9]. С учетом указанной логической зависимости правило предварительного отбора задается следующим выражением:

$$Q_0 = O_1 \wedge (O_2 \vee O_3)$$

Для наглядного представления порядка проверки условий логическое правило Q_0 может быть представлено в виде дерева решений, представленного на рисунке 1.



Рисунок 1 – Дерево предварительного отбора по основным показателям

При $Q_0 = 1$ аналитический процесс допускается ко второму этапу. При $Q_0 = 0$ признаки связанности представлений данных на уровне структуры аналитического процесса не выявляются. В данном случае достаточно проверки совместимости входных и выходных данных операций по формату, типу и структуре, поэтому дальнейшая оценка по дополнительным показателям не выполняется.

На втором этапе выполняется уточняющая оценка по дополнительным показателям, позволяющая определить уровень связанности представлений данных для рассматриваемого аналитического процесса [10, 11]. Для учета дополнительных показателей используется аддитивная оценка с равными весами:

$$Q_D = \sum_{j=1}^4 D_j$$

где D_j – значение j -го дополнительного показателя.

На основе значения Q_0 и Q_D определяется уровень связанности представлений данных. Классификация представлена в таблице 3.

Таблица 3 – Классификация уровней связанности представлений данных

Q_0	Q_D	Уровень связанности
0	–	Отсутствует
1	0	Низкий
1	1-2	Средний

1	3-4	Высокий
---	-----	---------

Предложенная процедура позволяет определить уровень связанности представлений данных на этапе проектирования аналитического процесса. Результат такой оценки может использоваться при выборе способа описания процесса и определении необходимости учета связей между данными, операциями и условиями их использования.

V. КОНТРАКТНО-ГРАФОВЫЙ ПОДХОД К ПРОЕКТИРОВАНИЮ МУЛЬТИМОДАЛЬНЫХ АНАЛИТИЧЕСКИХ ПРОЦЕССОВ

В качестве примера подхода к проектированию мультимодальных аналитических процессов со связанными представлениями данных рассматривается контрактно-графовый подход. Данный подход ориентирован на описание не только последовательности выполнения операций, но и состава используемых данных, допустимых преобразований и условий передачи промежуточных результатов.

Аналитический процесс представляется как совокупность взаимосвязанных операций обработки данных различных модальностей. Для каждой модальности задается контракт, определяющий допустимые форматы и типы данных, операции обработки и преобразования, а также условия и ограничения их использования. Он позволяет до выполнения процесса проверить, могут ли выходные данные одной операции быть использованы на входе другой с учетом формата, типа, допустимых преобразований и условий применения [12].

Формализованное описание аналитического процесса строится на основе синтеза трех взаимодополняющих моделей: структурной модели процесса, модели потоков данных и модели поведения элементов обработки. Структурная модель позволяет определить состав операций и порядок их выполнения. Модель потоков данных задает передачу промежуточных результатов между операциями и позволяет учитывать типизированные потоки данных. Модель поведения элементов обработки используется для описания состояний операций, включая ожидание входных данных, выполнение, успешное завершение, ошибку или повторный запуск.

Обобщенная схема контрактно-графового подхода представлена на рисунке 2.

Проектирование аналитического процесса выполняется поэтапно. Сначала определяются используемые модальности и ограничения их применения. Затем задается набор допустимых операций обработки и преобразования данных. После этого формируется структура процесса, описываются потоки данных между операциями и проверяется корректность их согласования.

Для наглядного представления контрактно-графового подхода на рисунке 2 приведена обобщенная схема проектирования мультимодального аналитического процесса, включающая модальности,

контракты модальностей, модели процесса и этапы его построения.

Особенность рассматриваемого подхода состоит в возможности описания данных, операций, связей между ними и условий их совместного использования. При переходе к другой предметной области могут изменяться используемые модальности, состав операций и способы объединения данных. Общий принцип проектирования при этом сохраняется и предполагает представление аналитического процесса как системы согласованных преобразований данных.

Практическая реализация контрактно-графового подхода может быть выполнена с использованием конструктора мультимодальной обработки данных, обеспечивающего построение, настройку и повторное использование аналитических процессов. Сформированные в конструкторе схемы позволяют наглядно представить структуру процесса, состав операций и связи между ними и служат основой для последующей оценки связанности представлений данных [13].

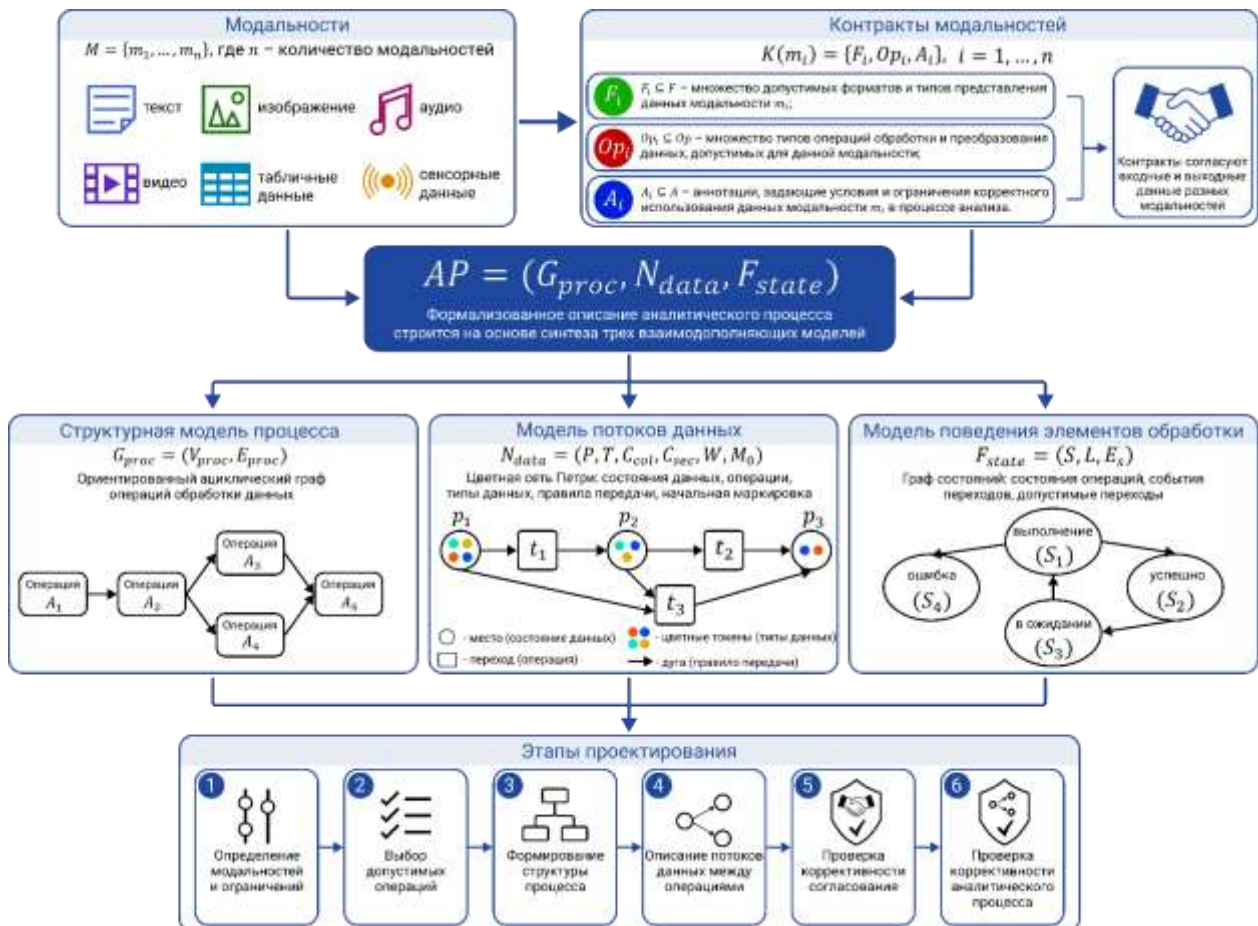


Рисунок 2 – Контрактно-графовый подход к проектированию мультимодального аналитического процесса

VI. ОЦЕНКА НЕОБХОДИМОСТИ СОГЛАСОВАНИЯ ДАННЫХ НА ПРИМЕРАХ ПРОЦЕССОВ, ПОСТРОЕННЫХ С ИСПОЛЬЗОВАНИЕМ КОНТРАКТНО-ГРАФОВОГО ПОДХОДА

Для демонстрации предложенной процедуры рассмотрены задачи обработки больших данных, на основе которых формируются аналитические процессы, отличающиеся структурой обработки, характером взаимодействия данных и требованиями к согласованию промежуточных результатов [14]. Рассматриваемые задачи включают данные разных представлений, степень связанности которых в процессе обработки различается.

В качестве примеров рассмотрены следующие задачи:

- первичная обработка данных промышленного мониторинга;

- предварительная проверка выполненного учебного задания;

- формирование медицинского отчета на основе результатов лабораторных исследований.

Для наглядного представления структуры рассматриваемых процессов используются схемы, сформированные в конструкторе мультимодальной обработки данных. Объектом оценки является не сама схема, а аналитический процесс, его данные, операции и связи между этапами обработки.

Первым примером является задача первичной обработки данных промышленного мониторинга, в которой на вход поступает единый пакет данных, сформированный за заданный интервал наблюдения и включающий временные ряды показаний датчиков и фотографии оборудования. Целью является подготовка данных разных типов к последующему использованию, а именно удаление выбросов во временных рядах и отбор изображений по заданным параметрам качества. Результаты обработки сохраняются отдельно и могут быть использованы далее в других системах.

Для данной задачи выделяются следующие



Рисунок 3 – Структура аналитического процесса первичной обработки данных промышленного мониторинга

Для оценки связанности представлений данных в рассматриваемом процессе проанализированы основные показатели. Результаты оценки представлены в таблице 4.

Таблица 4 – Оценки процесса первичной обработки данных промышленного мониторинга по основным показателям

Показатель	Значение	Обоснование
O_1	1	Используются данные разных представлений: временные ряды и фотографии



модальности:

Рисунок 4 – Структура аналитического процесса предварительной проверки выполненного учебного задания

$$M = \{m_1, m_2\}$$

где m_1 – временные ряды сенсоров, m_2 – изображение оборудования.

В рамках рассматриваемого процесса контракты модальностей задаются следующим образом:

- для m_1 : $K(m_1) = \{F_1, Op_1, A_1\}$, где $F_1 = \{\text{временные ряды}\}$, $Op_1 = \{\text{импорт, фильтрация выбросов, экспорт}\}$, $A_1 = \{\text{наличие пропусков, источник данных}\}$.

- для m_2 : $K(m_2) = \{F_2, Op_2, A_2\}$, где $F_2 = \{\text{изображение}\}$, $Op_2 = \{\text{импорт, фильтрация по качеству, экспорт}\}$, $A_2 = \{\text{разрешение, формат изображения}\}$.

В процессе используются операции импорта пакета данных, фильтрации выбросов во временных рядах, отбора изображений оборудования по параметрам качества, а также отдельные операции экспорта временных рядов и изображений.

Структура рассматриваемого аналитического процесса представлена на рисунке 3.

O_2	0	Данные не преобразуются в общее представление
O_3	0	Результаты обработки сохраняются и используются отдельно

Рассчитаем Q_0 :

$$Q_0 = O_1 \wedge (O_2 \vee O_3) = 0$$

Значение $Q_0 = 0$ означает, что для рассматриваемого процесса достаточно проверки совместимости данных по формату, типу и структуре, а согласование с использованием контрактно-графового подхода является избыточным.

В качестве второго примера рассматривается задача предварительной проверки выполненного учебного задания. На вход поступают текст задания и изображение рукописного решения. Цель состоит в сопоставлении условия задания с представленным решением. Результат проверки формируется как вспомогательный вывод и может быть использован преподавателем или другой системой для последующего анализа.

Для данной задачи выделяются следующие модальности:

$$M = \{m_1, m_2, m_3\}$$

где m_1 – текст, m_2 – изображение, m_3 – векторное представление.

В рамках рассматриваемого процесса контракты модальностей задаются следующим образом:

- для m_1 : $K(m_1) = \{F_1, Op_1, A_1\}$, где $F_1 = \{\text{текст}\}$, $Op_1 = \{\text{импорт, преобразование в эмбединг, распознавание текста, семантическое сопоставление, формирование отчета}\}$, $A_1 = \{\text{язык, кодировка, длина текста}\}$.

- для m_2 : $K(m_2) = \{F_2, Op_2, A_2\}$, где $F_2 = \{\text{изображение}\}$, $Op_2 = \{\text{импорт, распознавание текста}\}$, $A_2 = \{\text{разрешение, формат изображения}\}$.

- для m_3 : $K(m_3) = \{F_3, Op_3, A_3\}$, где $F_3 = \{\text{тензор}\}$, $Op_3 = \{\text{преобразование в эмбединг, объединение эмбедингов, семантическое сопоставление}\}$, $A_3 = \{\text{размерность, способ формирования}\}$.

В процессе используются операции импорта текста задания и изображения рукописного решения, распознавания текста, преобразования текстовых данных в эмбединги, объединения эмбедингов, оценки корректности выполнения задания и формирования итогового отчета.

Структура аналитического процесса представлена на рисунке 4.

Для оценки связанности представлений данных в процессе проанализированы основные показатели. Результаты оценки представлены в таблице 5.

Таблица 5 – Оценки процесса предварительной проверки выполненного учебного задания по основным показателям

Показатель	Значение	Обоснование
O_1	1	Используются данные разных представлений: текст, изображение и векторное представление
O_2	1	Присутствуют операции преобразования изображения в текст, текста – в векторное представление и формирования текстового вывода на основе векторных представлений
O_3	0	Объединение выполняется в пределах одной модальности – векторного представления

Рассчитаем Q_0 :

$$Q_0 = O_1 \wedge (O_2 \vee O_3) = 1$$

Значение $Q_0 = 1$ означает, что рассматриваемый процесс проходит предварительный отбор и допускается к оценке по дополнительным показателям. Результаты оценки по дополнительным показателям представлены в таблице 6.

Таблица 6 – Оценки процесса предварительной проверки выполненного учебного задания по дополнительным показателям

Показатель	Значение	Обоснование
D_1	0	Результат является вспомогательным и может быть дополнительно проверен преподавателем
D_2	1	Используются OCR и модель искусственного интеллекта
D_3	1	Объединение эмбедингов возможно только после получения двух промежуточных представлений

D_4	0	Результат проверяется специалистом, поэтому расширенный жизненный цикл операций не требуется.
-------	---	---

Рассчитаем Q_D :

$$Q_D = \sum_{j=1}^4 D_j = 2$$

Полученные значения $Q_0 = 1$ и $Q_D = 2$ соответствуют среднему уровню связанности представлений данных. Полученный результат связан с тем, что рассматриваемый процесс предполагает более сложную обработку, чем простая последовательная схема, однако не требует уровня контроля, характерного для критических сценариев с жесткими требованиями к корректности, состояниям выполнения и обработке ошибок.

В качестве третьего примера рассматривается задача формирования медицинского отчета по результатам лабораторного анализа крови. На вход поступают структурированные показатели капиллярной и венозной крови. Цель состоит в совместной обработке результатов анализов, выявлении отклонений относительно заданных значений и формировании итогового отчета. Результат процесса используется как вспомогательный аналитический материал для последующей медицинской интерпретации специалистом.

Для данной задачи выделяются следующие модальности:

$$M = \{m_1, m_2, m_3\}$$

где m_1 – таблица, m_2 – векторное представление, m_3 – текст.

В рамках рассматриваемого процесса контракты модальностей задаются следующим образом:

- для m_1 : $K(m_1) = \{F_1, Op_1, A_1\}$, где $F_1 = \{\text{таблица}\}$, $Op_1 = \{\text{импорт, преобразование в эмбединг}\}$, $A_1 = \{\text{тип анализа, наличие пропусков}\}$.

- для m_2 : $K(m_2) = \{F_2, Op_2, A_2\}$, где $F_2 = \{\text{тензор}\}$, $Op_2 = \{\text{преобразование в эмбединг, генерация отчета}\}$, $A_2 = \{\text{размерность, способ формирования}\}$.

- для m_3 : $K(m_3) = \{F_3, Op_3, A_3\}$, где $F_3 = \{\text{текст}\}$, $Op_3 = \{\text{генерация отчета, формирование отчета}\}$, $A_3 = \{\text{язык, кодировка, длина текста}\}$.

В процессе используются операции импорта результатов капиллярной и венозной крови, преобразования табличных данных в эмбединги, генерации промежуточных выводов по каждому виду анализа и формирования итогового отчета.

Структура рассматриваемого аналитического процесса представлена на рисунке 5.



Рисунок 5 – Структура аналитического процесса формирования медицинского отчета по результатам лабораторного анализа крови

Для оценки связанности представлений данных в рассматриваемом процессе проанализированы основные показатели. Результаты оценки представлены в таблице 7.

Таблица 7 – Оценки процесса формирования медицинского отчета по результатам лабораторного анализа крови по основным показателям

Показатель	Значение	Обоснование
O_1	1	Используются данные разных представлений: таблица, векторное представление и текст
O_2	1	Присутствуют операции преобразования таблицы в векторное представление и формирования текстового вывода на основе векторных представлений
O_3	0	Объединение выполняется в пределах одной модальности – текста

Рассчитаем Q_O :

$$Q_O = C_1 \wedge (C_2 \vee C_3) = 1$$

Значение $Q_O = 1$ означает, что рассматриваемый процесс проходит предварительный отбор и допускается к оценке по дополнительным показателям. Результаты оценки по дополнительным показателям представлены в таблице 8.

Таблица 8 – Оценки процесса формирования медицинского отчета по результатам лабораторного анализа крови по дополнительным показателям

Показатель	Значение	Обоснование
D_1	1	Ошибки входных или промежуточных данных могут повлиять на содержание отчета
D_2	1	Используется модель искусственного интеллекта для анализа результатов
D_3	1	Формирование итогового отчета возможно только после получения нескольких результатов анализа
D_4	1	При отсутствии или ошибке одного из результатов требуется учитывать состояния ожидания и повторной обработки

Рассчитаем Q_D :

$$Q_D = \sum_{j=1}^4 D_j = 4$$

Полученные значения $Q_O = 1$ и $Q_D = 4$ соответствуют высокому уровню связанности представлений данных. Полученный результат связан с тем, что рассматриваемый процесс относится к задачам с высокой ценой ошибки. Некорректные входные данные, ошибочная передача промежуточных результатов или преждевременное формирование отчета могут повлиять на последующую медицинскую интерпретацию и, соответственно, на принимаемые решения о состоянии пациента.

Проведенная оценка показывает, что необходимость согласования данных, операций и промежуточных результатов определяется не только

наличием данных разных представлений, но и характером их участия в аналитическом процессе. Если данные обрабатываются отдельно и не участвуют в формировании общего результата, достаточно проверки совместимости по формату, типу и структуре, а применение контрактно-графового подхода является избыточным. При наличии преобразований между представлениями, зависимостей между этапами и необходимости контроля промежуточных результатов требуется согласование данных, операций и условий их использования, которое может быть реализовано с использованием контрактно-графового подхода.

VII. ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача оценки связанности представлений данных при проектировании мультимодальных аналитических процессов обработки больших данных. Предложены основные и дополнительные показатели, позволяющие учитывать разнородность данных, наличие преобразований между представлениями, совместное использование результатов, применение моделей искусственного интеллекта и условия выполнения операций.

Разработана процедура оценки, основанная на логическом правиле предварительного отбора и аддитивной оценке. На примерах задач промышленного мониторинга, проверки учебного задания и формирования медицинского отчета продемонстрировано применение процедуры для определения необходимости согласования данных, операций и промежуточных результатов с учетом характера участия данных в аналитическом процессе. Для процессов со средней и высокой связанностью такое согласование может быть реализовано с использованием контрактно-графового подхода.

Предложенные показатели и процедура оценки могут использоваться аналитиками данных, разработчиками аналитических систем и специалистами предметных областей при проектировании аналитических процессов. Дальнейшие исследования будут направлены на автоматическое построение мультимодальных аналитических процессов с учетом оценки связанности представлений данных.

БИБЛИОГРАФИЯ

- [1] Rozony F.Z., Aktar M.N.A., Ashrafuzzaman M. et al, "A systematic review of big data integration challenges and solutions for heterogeneous data sources," Academic Journal on Business Administration, Innovation & Sustainability, pp. 1–18, 2024.
- [2] Abdurrahman A., Zeigler B.P., "Multimodal semantic fusion of heterogeneous data silos," Systems 2025, 13, no. 11, 2025.
- [3] Zhao F., Zhang C., Geng B., "Deep multimodal data fusion," ACM Computing Surveys 56, pp. 1–36, 2024.
- [4] Peng Y., Heyn H., Horkoff J., "Data challenges in ai systems and their solutions: a requirements and ai

engineering systematic literature review and comparison,” 2025 — 10.21203/rs.3.rs-8077403/v1.

[5] Yoo S., Lee J., Yoon C. et al, “Data transformation strategies to remove heterogeneity,” 2025 — arXiv:2507.12677.

[6] Oakes B.J., Famelis M., Sahraoui H. “Building domain-specific machine learning workflows: a conceptual framework for the state of the practice,” ACM Transactions on Software Engineering and Methodology, vol. 33, issue 4, no. 91, pp. 1–50, 2024.

[7] Li S., Tang H. “Multimodal alignment and fusion: a survey,” 2025 — arXiv:2411.17040.

[8] Takemura K., Murakami H., Tamari Y. “A probabilistic modeling approach to decision strategies: predicting expected information search and decision time in multi-attribute choice tasks with varying numbers of attributes and alternatives,” Mathematics 2026, no. 1, 2026.

[9] Steffen V., Oliveira M.S., Trojan, F., “A novel approach for systematic literature reviews using multi-criteria decision analysis,” Journal of Intelligent Management Decision, vol. 3, issue 2, pp. 116–138, 2024.

[10] Trung D.D., Ozcalici M., Dua T.V., “Comparative analysis of the simple additive weighting method and the deviation-based pairwise assessment ratio technique for material selection,” Tribology and Materials ISSN 2812-9717, 2026.

[11] Liborio M.P., Ekel P.I., Fusco E., et al, “Fuzzy inference system for measuring composite indicators' overall quality,” Social indicators research vol. 179, pp. 1587–1613, 2025.

[12] Чаругин В.В., Чаругин В.В., “Ставцев А.В., Чесалин А.Н. Архитектура конструктора мультимодальной обработки данных: графовые модели и их применение в задачах информационной безопасности,” 2026 — 10.33693/2313-223X-2026-13-1-55-67.

[13] Чаругин В.В., Чаругин В.В., Чесалин А.Н. “Конструктор мультимодальной обработки данных,” Инновационные, информационные и коммуникационные технологии. Сборник трудов XXI Международной научно-практической конференции, с. 146-150, 2024.

[14] Zheng Y., “Fusing cross-domain knowledge from multimodal data to solve problems in the physical world,” ACM Transactions on Intelligent Systems and Technology, vol. 17, issue 1, pp. 1–32, 2025.

Assessment of the connectedness of data representations in the design of multimodal analytical processes for big data processing

V.V. Charugin, V.V. Charugin, A.N. Chesalin

Abstract – The paper investigates the problem of designing multimodal analytical processes for big data processing with varying connectedness of data representations. Depending on the process structure, data of different types may be processed independently, transformed between representations, or jointly used in forming the result. To account for these features, a system of indicators of the connectedness of data representations and an assessment procedure based on a logical rule of preliminary selection and additive evaluation are proposed.

The connectedness of data representations is assessed using examples of analytical processes described with the contract-graph approach: primary processing of industrial monitoring data, preliminary verification of a completed educational assignment, and formation of a medical report based on laboratory test results.

The proposed assessment procedure is applied to determine the need for aligning data, operations, and intermediate results and can be used by data analysts, developers of analytical systems, and domain specialists in the design of multimodal analytical processes.

Keywords – big data, multimodality, analytical process, data representations, connectedness of data representations, connectedness indicators, assessment procedure, data alignment, machine learning, deep learning.

REFERENCES

- [1] Rozony F.Z., Aktar M.N.A., Ashrafuzzaman M. et al, "A systematic review of big data integration challenges and solutions for heterogeneous data sources," Academic Journal on Business Administration, Innovation & Sustainability, pp. 1–18, 2024.
- [2] Abdurrahman A., Zeigler B.P., "Multimodal semantic fusion of heterogeneous data silos," Systems 2025, 13, no. 11, 2025.
- [3] Zhao F., Zhang C., Geng B., "Deep multimodal data fusion," ACM Computing Surveys 56, pp. 1–36, 2024.
- [4] Peng Y., Heyn H., Horkoff J., "Data challenges in ai systems and their solutions: a requirements and ai engineering systematic literature review and comparison," 2025 — 10.21203/rs.3.rs-8077403/v1.
- [5] Yoo S., Lee J., Yoon C. et al, "Data transformation strategies to remove heterogeneity," 2025 — arXiv:2507.12677.
- [6] Oakes B.J., Famelis M., Sahraoui H. "Building domain-specific machine learning workflows: a conceptual framework for the state of the practice," ACM Transactions on Software Engineering and Methodology, vol. 33, issue 4, no. 91, pp. 1–50, 2024.
- [7] Li S., Tang H. "Multimodal alignment and fusion: a survey," 2025 — arXiv:2411.17040.
- [8] Takemura K., Murakami H., Tamari Y. "A probabilistic modeling approach to decision strategies: predicting expected information search and decision time in multi-attribute choice tasks with varying numbers of attributes and alternatives," Mathematics 2026, no. 1, 2026.
- [9] Steffen V., Oliveira M.S., Trojan, F., "A novel approach for systematic literature reviews using multi-criteria decision analysis," Journal of Intelligent Management Decision, vol. 3, issue 2, pp. 116–138, 2024.
- [10] Trung D.D., Ozcalici M., Dua T.V., "Comparative analysis of the simple additive weighting method and the deviation-based pairwise assessment ratio technique for material selection," Tribology and Materials ISSN 2812-9717, 2026.
- [11] Liborio M.P., Ekel P.I., Fusco E., et al, "Fuzzy inference system for measuring composite indicators' overall quality," Social indicators research vol. 179, pp. 1587–1613, 2025.
- [12] Charugin V.V., Charugin V.V., Stavtsev A.V., Chesalin A.N. "Architecture of a constructor for multimodal data processing: graph models and their application in information security tasks," 2026 — 10.33693/2313-223X-2026-13-1-55-67.
- [13] Charugin V.V., Charugin V.V., Chesalin A.N. "Constructor for multimodal data processing," Innovative, information and communication technologies. proceedings of the xxi international scientific and practical conference, pp. 146–150, 2024.
- [14] Zheng Y., "Fusing cross-domain knowledge from multimodal data to solve problems in the physical world," ACM Transactions on Intelligent Systems and Technology, vol. 17, issue 1, pp. 1–32, 2025.