ФОРМАЛЬНАЯ КОНВЕЙЕРНАЯ АРХИТЕКТУРА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ГЕТЕРОГЕННЫХ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ С КОМПОЗИТНОЙ МЕРОЙ ГЕТЕРОГЕННОСТИ

Болтаев Ж.Э.[1], Абдирайимов Х.Б.[1], Насриддинов А.Н.[1],
Марахимов А.Р.[2], д.т.н., профессор, Худайбергенов К.К.[3], PhD
[1] Термезский государственный университет, г. Термез
[2] Ташкентский государственный технический университет им. И. Каримова, г. Ташкент
[3] Международный университет Кимё, г. Ташкент

Ключевые слова: гетерогенные данные, конвейерная архитектура, композитная мера, идемпотентная нормализация, информационный выигрыш.

Проблема. Системы интеллектуального мониторинга обрабатывают данные о субъектах из множества разнородных источников: банковская база хранит транзакции в смешанном тексте; налоговая — декларации в табличном формате; кадастровая система — геопривязанные записи о земле и имуществе; тендерная платформа — документы в формате PDF; суды — неструктурированные текстовые решения. Отсутствует единая формально верифицированная архитектура, обеспечивающая полный конвейер от гетерогенного входа до интерпретируемого результата с количественно измеримым снижением гетерогенности на каждом уровне обработки. Существующие подходы решают лишь фрагменты задачи и не предоставляют объективной меры качества канонизации, что препятствует сопоставимой оценке эффективности различных стратегий нормализации.

Предлагаемое решение. Разработана шестиуровневая конвейерная архитектура, представимая как композиция операторов:

$$F(s) = (E \circ A \circ G \circ M \circ N \circ P)(s),$$

где P — препроцессор (агрегация источников), N — оператор нормализации, M — экстрактор именованных сущностей, G — построитель графа знаний, A — детектор аномалий, E — модуль объяснимости. Каждый уровень оперирует данными со строго меньшим значением меры гетерогенности, что обеспечивает монотонность канонизации.

Композитная мера гетерогенности. Расширяя таксономию Sheth [1], формализованы десять видов гетерогенности (структурная, синтаксическая, семантическая, модальная, лингвистическая, масштабная, временна́я, качественная, кодировочная, онтологическая); каждому виду k сопоставлена частная мера $h_k \in [0, 1]$. Композитная мера определена как:

$$H(D) = \sum_{k=1}^{10} w_k \cdot h_k(D), \quad \sum_{k=1}^{10} w_k = 1,$$

где веса $w_k = IG(Y; h_k)$ определены методом информационного выигрыша. Мера удовлетворяет ограниченности $0 \le H(D) \le 1$ и субаддитивности $H(D_1 \cup D_2) \le H(D_1) + H(D_2)$, что позволяет применять её к корпусам, собираемым инкрементально из многих источников.

Оператор нормализации. Оператор N реализован как композиция 14 детерминированных преобразований $N = f_{14} \circ \dots \circ f_1$, объединённых в три группы: извлечение метаданных, канонизация аббревиатур и скриптовая унификация кириллицы и латиницы. Оператор обладает свойством идемпотентности N(N(s)) =

N(s), что гарантирует согласованность представлений между обучающей и инференсной фазами и исключает дрейф данных при повторной обработке.

Результаты. Валидация проведена на корпусе 76 642 записей из 25 источников финансового мониторинга с применением K-кратной перекрёстной проверки (K = 5), тестирования на unseen данных и ручной верификации 200 случайных записей. Таблица 1 подтверждает монотонное снижение H(D) по уровням конвейера.

Уровень обработки	**H(D)**	**Снижение, %**
Исходные данные	8,70	—
После нормализации N	3,21	63,1
После извлечения сущностей M	1,04	88,0
После построения графа G	0,37	95,7
После детектора аномалий A	**0,14**	**98,4**

Таблица 1. Динамика меры гетерогенности H(D) по уровням конвейера.

Совокупное снижение меры составляет 8,70 → 0,14, что соответствует относительному уменьшению на 98,4 %. Наибольший вклад вносит оператор нормализации N (63,1 % суммарного снижения), что подтверждает принципиальную роль детерминированной канонизации в обработке многоскриптовых текстов. Последующие уровни извлечения сущностей, построения графа и детектирования аномалий вносят дополнительное снижение до итогового значения 0,14. Доказанные свойства модели — ограниченность, субаддитивность композитной меры и идемпотентность оператора нормализации — обеспечивают формальную корректность архитектуры и её устойчивость при переносе на смежные предметные области, где сохраняется семантика категорий входных данных, но различаются конкретные распределения признаков.

Выводы. (1) Введена композитная мера гетерогенности H(D) с весами на основе информационного выигрыша, формализующая десять классов разнородности данных и удовлетворяющая свойствам ограниченности и субаддитивности. (2) Разработана шестиуровневая конвейерная архитектура $F = E \circ A \circ G \circ M \circ N \circ P$, реализующая операторную декомпозицию обработки гетерогенных данных с идемпотентным оператором нормализации. (3) Монотонное снижение композитной меры с 8,70 до 0,14 (на 98,4 %) экспериментально подтверждено на корпусе из 76 642 записей финансового мониторинга. (4) Предложенный формализм не является специфичным для финансовой сферы и допускает прямой перенос на смежные предметные области — интеллектуальный мониторинг образовательных процессов, медицинские информационные системы, логистические и таможенные платформы, — что исследуется в тезисе 2 на примере гибридного извлечения сущностей и ансамблевого детектирования аномалий.

Литература

1. Sheth A., Larson J. Federated database systems // ACM Comput. Surv. 1990. V. 22(3). P. 183–236.
2. Conneau A. et al. Unsupervised cross-lingual representation learning at scale // Proc. ACL. 2020. P. 8440–8451.
3. Dong X.L. et al. Knowledge Vault // Proc. KDD. 2014. P. 601–610.
4. McNemar Q. Sampling error // Psychometrika. 1947. V. 12(2). P. 153–157.
5. Devlin J. et al. BERT: Pre-training of deep bidirectional transformers // Proc. NAACL. 2019. P. 4171–4186.
6. Quinlan J.R. Induction of decision trees // Machine Learning. 1986. V. 1(1). P. 81–106.