МЕЖДУНАРОДНЫЙ ЦЕНТР НАУЧНОГО СОТРУДНИЧЕСТВА
«НАУКА И ПРОСВЕЩЕНИЕ»

НАУКА и ПРОСВЕЩЕНИЕ
МЕЖДУНАРОДНЫЙ ЦЕНТР НАУЧНОГО СОТРУДНИЧЕСТВА

АКТУАЛЬНЫЕ НАУЧНЫЕ ИССЛЕДОВАНИЯ

СБОРНИК СТАТЕЙ XXXII МЕЖДУНАРОДНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ,
СОСТОЯВШЕЙСЯ 15 ФЕВРАЛЯ 2026 Г. В Г. ПЕНЗА

ПЕНЗА
МЦНС «НАУКА И ПРОСВЕЩЕНИЕ»
2026

ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

УДК 004.021

АЛГОРИТМЫ ВЕРОЯТНОСТНОГО СОПОСТАВЛЕНИЯ ДАННЫХ СИСТЕМЫ ФОРМИРОВАНИЯ ОПТИМАЛЬНЫХ ПРЕДЛОЖЕНИЙ НА ВЫБОР ИЗДЕЛИЙ ЭЛЕКТРОННОЙ КОМПОНЕНТНОЙ БАЗЫ В ПРОЦЕССАХ РАЗРАБОТКИ РАДИОЭЛЕКТРОННОЙ АППАРАТУРЫ

РУБЦОВ ЮРИЙ ВАСИЛЬЕВИЧ
генеральный директор
Акционерное общество «Центральное конструкторское бюро «Дейтон»
(АО «ЦКБ «Дейтон»)

Аннотация: Разработка радиоэлектронных устройств требует знания о параметрах и показателях применяемых изделий электронной компонентной базы. Такие знания способна предоставить информационная справочная система. Информация в такую систему собирается из различных источников, обобщается, анализируется и предоставляется специалистам в виде проверенных и достоверных данных.
Качество предоставляемых данных обеспечивается методами сопоставления на этапе нормализации. Для их выполнения используются детерминированные алгоритмы, вероятностные и с применением технологий искусственного интеллекта.
В настоящей статье показаны результаты исследования, разработки и использования алгоритма сопоставления данных вероятностным методом для информационной справочной системы, разработанной АО «ЦКБ «Дейтон» и функционирующей для предприятий радиоэлектронной промышленности.
Ключевые слова: сопоставление данных, искусственный интеллект, машинное обучение, электронная компонентная база, радиоэлектронная аппаратура.

ALGORITHMS FOR PROBABILISTIC COMPARISON OF DATA OF THE SYSTEM FOR FORMING OPTIMAL PROPOSALS FOR THE SELECTION OF ELECTRONIC COMPONENTS IN THE DEVELOPMENT PROCESSES OF RADIO ELECTRONIC EQUIPMENT

Rubtsov Yuri Vasilievich

Abstract: The development of electronic devices requires knowledge of the parameters and performance indicators of the electronic components used. An information reference system can provide this knowledge. Information in such a system is collected from various sources, summarized, analyzed, and presented to specialists as verified and reliable data.
The quality of the provided data is ensured by comparison methods at the normalization stage. These

methods utilize deterministic, probabilistic, and artificial intelligence algorithms.
This article presents the results of the research, development, and implementation of a probabilistic data comparison algorithm for an information reference system developed by JSC “Central Design Bureau “Dayton” and operating for enterprises in the electronics industry.
Keywords: data comparison, artificial intelligence, machine learning, electronic component base, electronic equipment

Введение

В радиоэлектронной промышленности разработана и функционирует информационная справочная система (Система) для получения справочной информации об изделиях электронной компонентной базы (ЭКБ). Она обеспечивает специалистов данными о параметрах и показателей изделий ЭКБ. В том числе разработчиков радиоэлектронной аппаратуре (РЭА).

Информация в Систему поступает от различных источников: изготовителей, поставщиков и потребителей ЭКБ; исследовательских, измерительных и испытательных организаций; из конструкторской и технологической документации на ЭКБ, из протоколов результатов исследований, измерений, испытаний, применения, эксплуатации, и модернизации изделий ЭКБ. В качестве потребителя такой информации выступает Система, в том числе формирующая перечни ЭКБ для принятия решения о ее применении в РЭА. Для нормального функционирования Системы, важно иметь эффективные методы для работы с ошибками в данных. Оценка качества и исправления данных становится ключевым процессом в Системе. В такой ситуации необходимы быстрые, оптимизированные и точные методы анализа и очистки данных. Пропущенные элементы в наборах данных, несоответствия и дублированные являются дефектами, указывающие на низкое или недостаточное качество данных. Низкое качество данных связано с разнородными элементами, демонстрирующие различные форматы, структуры и типы (гетерогенность).

Характерная для данных низкого качества гетерогенность может быть либо структурной, либо семантической. Структурная гетерогенность возникает, когда данные не согласованы в рамках Системы, что на практике означает - разные источники могут передавать одни и те же данные, используя разные представления (например, разные наборы данных, разные типы, разные значения, ошибки формирования данных и т.д.). Семантическая гетерогенность возникает, когда одна и та же, или схожая структура используется для представления сущностей, которые по своей сути различаются друг от друга. Особенности одного и другого типа гетерогенности исследованы и учтены при построении алгоритма вероятностного сопоставления с помощью Information Classification, Marking and Handling - совокупности взаимосвязанных и взаимодействующих методов и инструментов, применяемых для решения задач сбора, обработки и анализа информации и получения достоверных данных [1].

В исследовании используются термины - прилагательные, по звучанию подобные, но имеющие различные значение. Вероятное – это возможное, или допустимое сопоставление. Вероятностное – относится к установлению вероятности сопоставления данных, основанное на предположениях или допущениях.

Под дискриминационной способностью в Системе понимается способность алгоритма различать и отделять различные наборы данных.

В настоящем исследовании используются термины, за которыми стоят определения: объединения множеств M и U– это ряд таких элементов, при котором каждый из них представляет собой элемент одного из первоначальных множеств, изображенные на рис. 1; пересечение множеств M и U - заключает в себе все элементы, общие для первоначальных множеств, изображенные на рис. 2.

Базовая теория вероятностного сопоставления [2] явилась основанием для разработки десятков алгоритмов. Основная цель состояла в том, чтобы исследовать как классические, так и более инновационные доступные методы и построить алгоритм вероятностного сопоставления для нормализации данных в Системе.

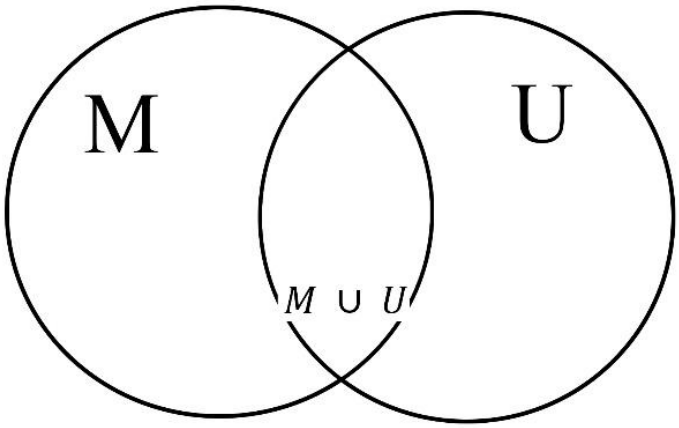

Рис. 1. Представление объединения множеств

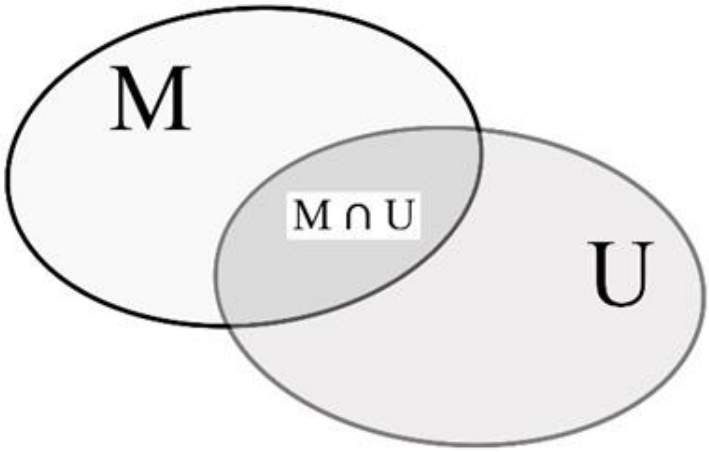

Рис. 2. Представление пересечения множеств

Использование при сопоставлении детерминированного (точного) метода недостаточно, особенно, если данные поступают из нескольких различных источников данных. Точное сопоставление работает только в том случае, если элементы идеальны и присутствуют во всех сопоставляемых наборах данных, например, если существуют уникальные идентификаторы.

Условия точного сопоставления могут не выполняться из-за ошибок, опечаток или невыполнения стандартов источниками информации. В этой ситуации используется вероятностное сопоставление которое является первым шагом к устранению проблем.

В алгоритме вероятностного сопоставления могут использоваться потенциальные идентификаторы. Вычисляются веса для каждого из них на основе предполагаемой способности правильно идентифицировать совпадения или отсутствие совпадения. Используются вычисленные веса для расчета вероятности того, что два элемента в различных наборах данных на самом деле соответствуют одному и тому же изделию ЭКБ.

Концепция разрабатываемого алгоритма вероятностного сопоставления данных предусматривает сопоставление элементов наборов данных только в том случае, если они совпадают с определенной вероятностью.

Помимо синтаксической информации, содержащейся в данных об ЭКБ, важную роль играет семантическая информация: формат и типы данных, допустимые значения, ссылочные ограничения и т. д. Данные неидентичные, если их значения совпадают неточно. Сопоставление неидентичных данных учитывает вероятность наличия общих черт на основе заранее определенных критериев или правил.

Сопоставляемые данные могут иметь семантическое соответствие обусловлено также термино-

логическими отношениями. Например, слова «резистор» и «сопротивление» в отношении изделий ЭКБ являются синонимами и, следовательно, соответствуют друг другу. В то же время «сопротивление» может быть номиналом резистора. Так же, как и его мощность рассеяния: например, 5 Ватт и 5 W. Это требует использования дополнительных источников, таких как словари.

Алгоритм вероятностного сопоставления используется при работе с противоречивыми или неполными данными. Вероятностное (неточное) сопоставление позволяет сравнивать данные на основе предопределенных правил или критериев и помогает выявлять общие черты. Методы вероятностного сопоставления предназначены для обработки частичных совпадений, обеспечивая гибкость при работе с искажениями данных. Предоставляя оценку сходства, такую как веса, вероятностное сопоставление позволяет принимать более правильные решения и более высокую устойчивость к ошибкам в собранных данных.

Данные сопоставляются из различных источников. Результаты применения вероятностных методов показывают способность сопоставлять данные даже когда основные (ключевые) идентификаторы отсутствуют. Эта работа открывает путь для полноты нормализации данных и обеспечения целей для работы Системы.

Разработка РЭА является перспективной областью для применения методов искусственного интеллекта (ИИ) и, в частности, методов машинного обучения (МО). Большинство алгоритмов МО сильно полагаются на доступные данные - не только по объему и количеству образцов в наборе данных, но и по разнообразию, числу параметров или признаков в наборе данных. Интуитивно, наличие широкого разнообразия наборов данных облегчает поиск наиболее релевантных признаков. Это потому, что пространство поиска больше и, следовательно, с большей вероятностью сопоставление может быть релевантным, при условии, что объем достаточен, чтобы обеспечить доверие к ИИ. Однако, из-за независимости различных источников информации и отсутствия консолидации, данные, относящиеся к одному и тому же изделию ЭКБ, различаются.

Результаты МО, полученные из данных различных источников, не всегда будут соответствовать целям. Это потенциально ограничивает применимость технологий ИИ в разработанной Системе. В отличие от алгоритмов точного сопоставления, любое вероятностное сопоставление будет иметь ложные положительные и отрицательные результаты, что в конечном итоге влияет на модели МО, обученные на них. Полный анализ наборов данных в части сопоставления с применением алгоритмов детерминированного, вероятностного и с применением МО сопоставления позволит получить максимальный эффект.

Каскадное смешанное эвристическое сопоставление применяет точные, вероятностные алгоритмы и с применением МО в порядке от самого строгого к наименее строгому. Это позволяет алгоритмам определять совпадения на основе «каскада» критериев. Даже если нет совпадения между критериями наверху каскада, алгоритмы пытаются найти совпадение на основе критериев ниже в каскаде, которые имеют меньший вес при подтверждении совпадения. Каскадное эвристическое сопоставление уменьшает ложные положительные и ложные отрицательные результаты.

В большинстве случаев процесс деидентификации состоит в устранении ссылок, необходимых для точного сопоставления между наборами данных. В этой статье представлен вероятностный алгоритм для объединения наборов данных при условиях возможного отсутствии идентификаторов. Метод находит лучшее сопоставление между данными поступающими от различных источников, в различных форматах и в разное время.

Тип вероятностного сопоставления использует методы, которые предсказывают совпадения элементов среди нескольких похожих наборов данных. Помимо такой оценки, методы делают обоснованные предположения о вероятности того, что несколько элементов наборов данных относятся к одному и тому же изделию ЭКБ.

Хотя это может быть более рискованно, чем точное сопоставление, вероятностные методы сопоставления могут обнаруживать менее очевидные связи, поскольку они охватывают более широкий массив данных и делают допущения относительно неверных или отсутствующих данных.

Вероятностное сопоставление данных определяет совпадения, увеличивая толерантность к раз-

личиям между наборами данных. Методы поиска неправильно написанных слов или выражения, также используют этот тип сопоставления.

Представленные в этом введении результаты анализа методов сопоставления позволили разработать алгоритмы и компьютерные программы, применение которых расширили возможности Системы. Данные из различных источников собираются в Системе, обрабатываются и предоставляются конструктору – разработчику оптимально сформированные перечни ЭКБ для применения в РЭА.

Математические обоснования для алгоритма вероятностного сопоставления

Наиболее строгие математические обоснования для вероятностного сопоставления были предложены [2] как расширение ранней работы [3]. В проведенных исследованиях используются положения, определенные в [2,3], чтобы обосновать разработанный алгоритм для Системы.

Детерминированное сопоставление является самым простым методом сопоставления. Этот метод лучше всего работает, когда идентификаторы имеют равную важность. Детерминированное сопоставление хорошо работает, если идентификаторы не содержат ошибок и всегда присутствуют во всех наборах данных.

Вероятностное сопоставление позволяет использовать более широкий диапазон сопоставление, вычисляя веса для каждого набора данных на основе способности правильно идентифицировать совпадения или отсутствие совпадения элементов данных. Затем эти веса используются для расчета вероятности того, что наборы данных относятся к одному и тому же изделию. Точное сопоставление наборов данных требует серии потенциально сложных правил (например, SQL-запросов), которые необходимо запрограммировать заранее. Алгоритм вероятностного сопоставления наборов данных может функционировать с минимальным вмешательством человека.

Для определений и описаний алгоритма вероятностного сопоставления данных нами введены пространства вариантов сопоставления данных: Пл – истинных положительных (правильных) сопоставлений, Лп – ложных положительных сопоставлений, Ло – ложных отрицательных, От – истинных (правильных) отрицательных, Вс - вероятных совпадений, Вн - вероятных несовпадений. Разделения пространств вариантов сопоставления данных имеем условный вид, представленный на рис. 3.

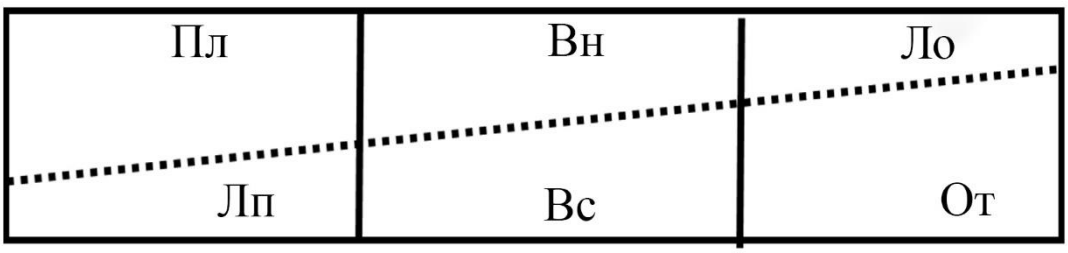

Рис. 3. Разделения пространств вариантов сопоставления данных

В методе вероятностного сопоставления данных используется два набора данных, обозначенных как A и B. Элементы, принадлежащие каждому набору данных обозначены строчными символами, a ∈ A и b ∈ B. Сопоставляемые элементы набора данных обозначаются как (a) и (b) соответственно, а сопоставляемые наборы A и B делятся на два множества, для сопоставляемых пар (1):

$$M = \{(a, b) \in A\ x\ B \mid a = b\} \quad (1)$$

для несопоставляемых пар (2):

$$U = \{(a, b) \in A\ x\ B \mid a \neq b\} \quad (2)$$

Для выражения (1) - обозначено прямое произведение множеств A и B, которое состоит из всех

возможных сопоставляемых пар (a, b), где элемент a принадлежит A, а элемент b принадлежит B. В контексте множеств, A x B представляет собой декартово произведение множеств A и B. Это множество всех упорядоченных пар (a, b), где a принадлежит множеству A, а b принадлежит множеству B. Количество элементов в декартовом произведении A x B равно количеству элементов в A, умноженному на количество элементов в B. Декартовым произведением A × B является множество, состоящее из всех возможных упорядоченных пар, в которых первый элемент принадлежит одному множеству, а второй - другому множеству.

Для выражение (2) - | a ≠ b: - условие, которое ограничивает множество U. И это означает, что для пар элементов (a, b), входящих в множество U, первый элемент (a) не должен быть равен второму элементу (b).

Данные могут сравниваться (3):

$$M \cup U = A \text{ x } B \quad (3)$$

В уравнении (3), M ∪ U обозначает объединение множеств M и U, а A x B обозначает декартово произведение множеств A и B. Выражение (3) используется в исследованиях [4] и представляет собой не совсем верное равенство, так как объединение и декартово произведение – это разные операции над множествами.

Объединение множества M ∪ U, содержит все элементы, которые либо находятся в M, либо находятся в U, либо в обоих множествах одновременно.

Пересечение множеств M ∩ U является пустым множеством, как показано в (4).

$$M \cap U = \emptyset \quad (4)$$

Для (4), у множества M и множества U нет положительных результатов при сопоставлении элементов.

Несмотря на то, что существуют различные подходы к проблеме вероятностно сопоставления данных, алгоритм для Системы может быть обобщен и детализирован: каждая пара элементов относится к M, либо к U, в соответствии со значением оценки и выбранными пороговыми уровнями. Этот базовый алгоритм подчеркивает основные принципы, которые имеют отношение к сравнению различных вероятностных методов сопоставляемых элементов.

Затем, два набора данных сопоставляются с помощью векторной функции пар элементов наборов данных (5):

$$(a, b) \in A \text{ x } B \quad (5)$$

С точки зрения вероятности, сопоставление, является событием, и, как результат, к вектору может быть прикреплена условная вероятность в виде (6):

$$m(\gamma) = P(\gamma|(a, b) \in M = P(\gamma|M), \quad (6)$$

и в виде (7):

$$u(\gamma) = P(\gamma|(a, b) \in U = P(\gamma|U), \quad (7)$$

которые соответственно представляют вероятность наступления события при совпадении и при несовпадении. Пары элементов наборов данных обозначаются как: A_1 для Пл; A_2 для Лп; A_3 для Ло; A_4 для От; A_5 для Вс и A_6 для Вн. Следовательно, правило сопоставления - это функция решения (8):

$$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma), P(A_4|\gamma), P(A_5|\gamma), P(A_6|\gamma)\} \quad (8)$$

Для нее выполняется равенство (9):

$$P(A_1|\gamma) + P(A_2|\gamma) + P(A_3|\gamma) + P(A_4|\gamma) + P(A_5|\gamma) + P(A_6|\gamma) = 1 \quad (9)$$

И правило принятия решения на сопоставление определяется как (10)

$$R = \frac{m(\gamma)}{u(\gamma)}, \quad (10)$$

Тогда, когда $R \geq t_m$ совпадение найдено, а $R \leq t_u$ когда несовпадение найдено, где t_m и t_u - пороги, которые нужно установить.

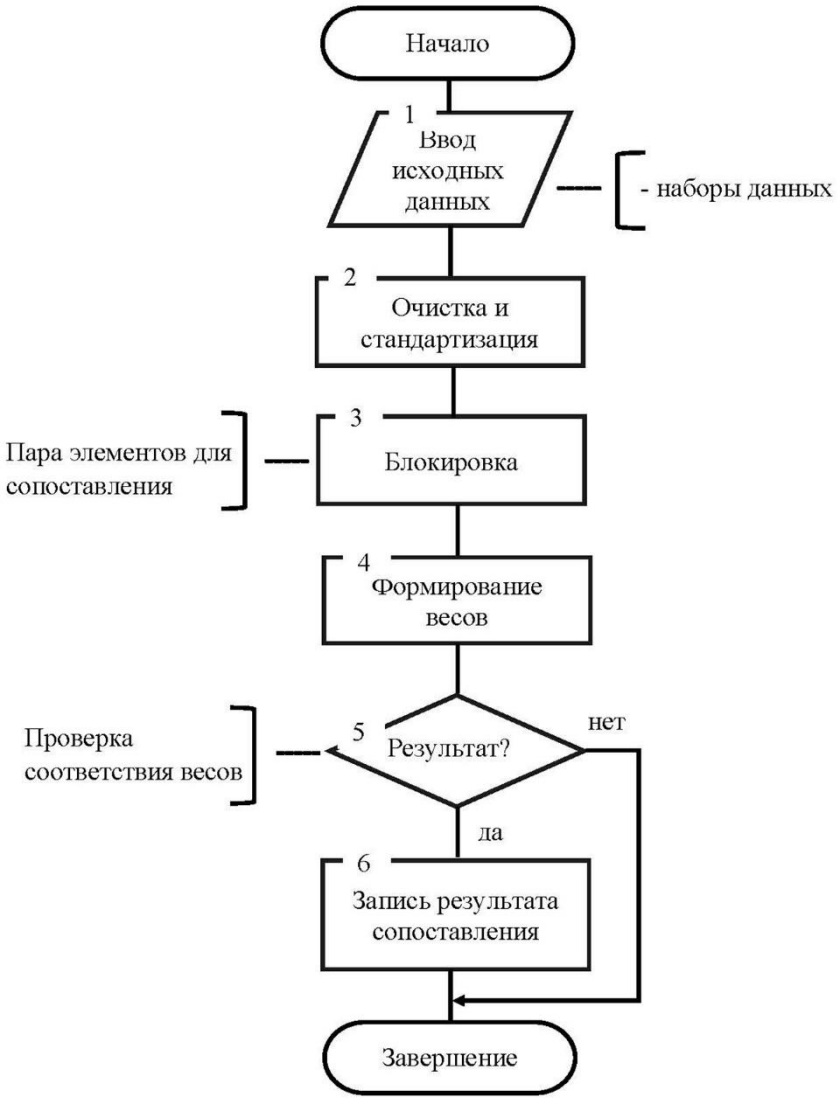

Рис. 4. Блок схема алгоритма вероятностного сопоставления

Методология вероятностного сопоставления подразумевает применение алгоритма максимизации ожидания (EM - состоит из двух шагов: «шаг ожидания» (E-шаг) и «шаг максимизации» (M-шаг), которые повторяются до получения результатов) для оценки параметров [5]. При этом используются приближенные сравнения элементов для расчета весов частичного их равенства, когда предполагается, что они содержат значения на основе набора данных. В последние годы исследователи начали изучать подходы, основанные на МО, такие как контролируемое обучение на основе обучающего набора дан-

ных с известной связью. Хотя подходы на основе МО многообещающие, в этой статье они не будут рассматриваться, в основном из-за перспективы описания алгоритма сопоставления с применением МО в Системе в следующей статье.

Вероятностное сопоставление ограничено размерностью данных поскольку подходы, использующие сравнения всех со всеми, не применениями в этой ситуации. Методы блокировки позволяют ограничить сравнения, уменьшая количество наборов данных, используя один или несколько различающих идентификаторов. Поскольку стратегии блокировки могут влиять на успешность сопоставления, необходим алгоритм описания конкретных шагов. В то время как простые стратегии блокировки сравнивают все пары, которые сравниваются с одним и тем же значением, существуют более продвинутые методы, которые включают неконтролируемую кластеризацию, например, формируют кластеры вокруг определенных ключевых значений [5]. Альтернативой стандартной блокировке и кластеризации является подход сортированного соседства, где элементы сортируются по блокирующей переменной до того, как скользящее окно перемещается по сортированному набору, что позволяет проводить сравнения между элементами в пределах окна и существенно сократить затраты на сопоставление.

Вероятностная связь оказывается более успешной для сопоставления строковых идентификаторов, таких как наименование изделий, функциональное назначение, область применения и др. Блок схема алгоритма вероятностного сопоставления представлена на рис. 4.

На рис. 4, блок 2 - очистка и стандартизация данных**.** Основной целью таких действий является преобразование необработанной входной информации в четкие и определенные данных, устранение всех возможных несоответствий в способе их представления или кодирования

В алгоритме вероятностного сопоставления использованы и показаны на рис. 5, в качестве частного алгоритма, следующие процедуры очистки данных:

а) изменение регистра в строчных выражениях;
б) удаление знаков препинания;
в) удаление последовательных пробелов;
г) удаление конечных или начальных пробелов;
д) удаление префиксов и суффиксов;
е) игнорирование повторений;
ж) поиск и изменение транспозиций;
з) унификация форматов дат;
и) использование контрольных сумм (т. е. метода, который проверяет идентификатор с помощью математического алгоритма).

Описание роли блока 3 алгоритма, представленного на рис. 4 – блокировка. Без использования метода блокировки, количество сопоставлений равно произведению количества элементов в наборах данных. Цель методологии блокировки - обойти как можно больше пар наборов данных из множества несовпадений U, как показано в (2), которые являются очевидными несовпадениями, без обхода пар наборов данных из множества совпадений M как показано в (1).

В процессе исследований возник вопрос, может ли алгоритм вероятностного сопоставления динамически выбирать разные методы блокировки с параметрами и настройками для различных наборов данных с минимальным вмешательством человека? Исследования показывают, что для изделий ЭКБ динамика изменения данных не справляется с разнообразными сценариями использования этих данных. Чтобы получить результаты исследований более гибкими и применимыми, предполагается, что адаптивность сыграет критическую роль в блокировке данных. Предварительно утверждается, что можно достичь значительного улучшения в сопоставлении данных за счет адаптивного и динамического изменения параметров алгоритма вероятностного сопоставления.

Исследования методов блокировки показали целесообразность применения в них сортированного соседства [5]. Сортировка выполняется таким образом, чтобы похожие данные были близки друг к другу, что позволяет сравнивать их в пределах заданного окна.

Метод состоит из трех шагов. Сначала создается идентификатор для сопоставляемого набора данных. Затем наборы данных сортируются. Окно фиксированного размера перемещается по последо-

вательному списку наборов данных, чтобы ограничить объем сопоставления.

Эффективность метода сортированного соседства заключается в том, чтобы избежать сравнения всех пар наборов данных. Благодаря сортировке и использованию скользящего окна значительно сокращается количество сравнений, особенно при большом количестве данных. Тем не менее, выбор ключа и размера окна может повлиять на эффективность алгоритма, поскольку слишком маленькое окно может пропустить совпадения, а слишком большое окно приведет к ненужным сравнениям.

Представленные в открытом доступе методы сортированного соседства не совсем обеспечивают заданные параметры точности сопоставления, определенные в [11]. Данные поступают от различных источников и их необходимо также сортировать по времени поступления.

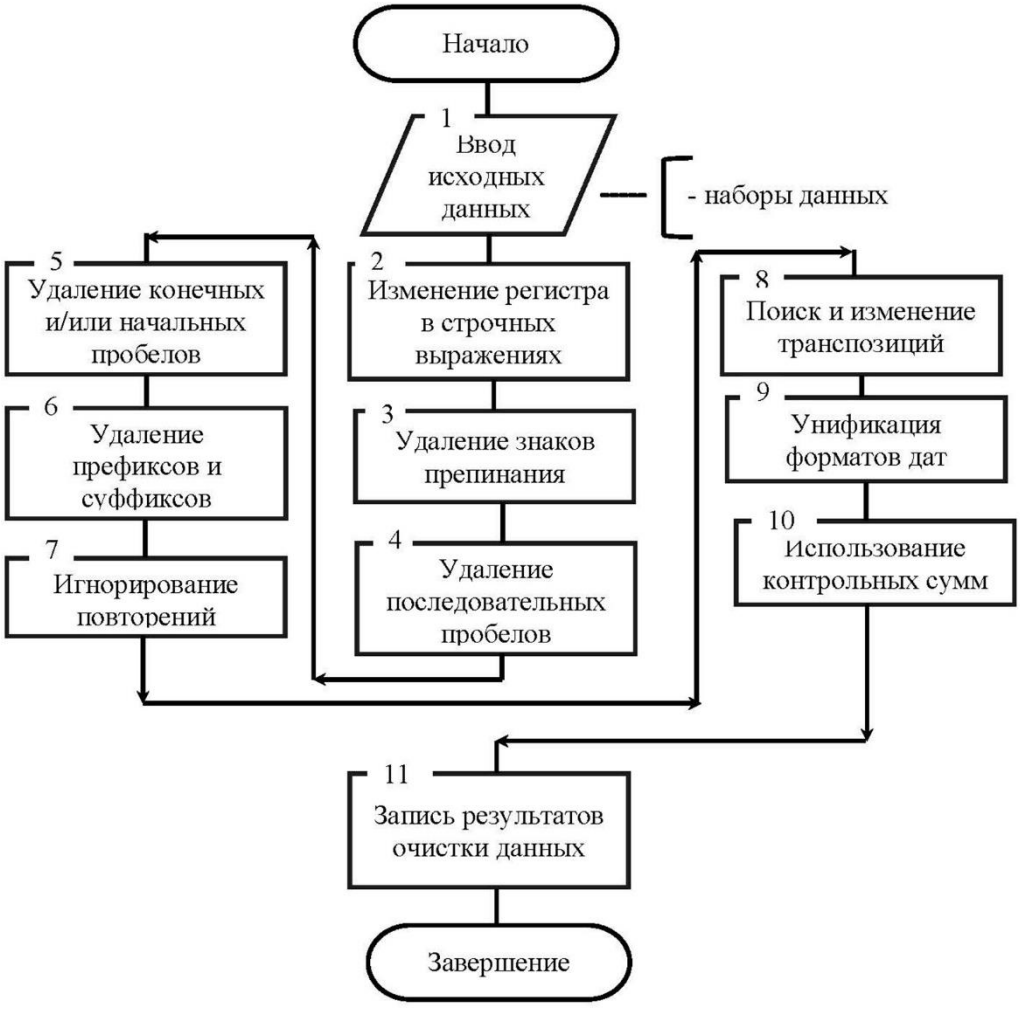

Рис. 5. Блок схема алгоритма очистки данных

Фиксированные исходные данные для сортированного соседства не могут справиться с динамикой изменения и поступления данных об ЭКБ. Например, для разрабатываемой специалистами АО «НИИЭТ» (г.Воронеж) микросхемы К1921ВГ1Т. Это двухъядерный 32-разрядный микроконтроллер. Изделие находится в состоянии разработки, но опытно применяется в различных уздах радиоэлектронной аппаратуры. Содержит энергонезависимую память объемом 4096 Кбайт, широкий набор универсальных и специализированных устройств и периферийных интерфейсов. В процессе разработки, измерений, результатов применения и испытаний, на изделие в Систему поступают уточняющие и новые данные. На момент написание статьи, это 224 наборов данных, содержащих от одного до 112 элементов. С помощью программных средств построено соотношение элементов к набору данных и представлено на рис.6.

Каждый набор данных содержит разное количество элементов, как показано на рис. 6. Таким образом, выбор правильного размера окна, требует решения автоматического изменения размера окна.

Контролируя, насколько близки или далеки два сопоставимые набора данных и их элементы, необходимо адаптивно увеличить или уменьшить скользящее окно. Для этого в наших исследованиях апробировано использование различной эвристики (линейные, экспоненциальные и простые числа).

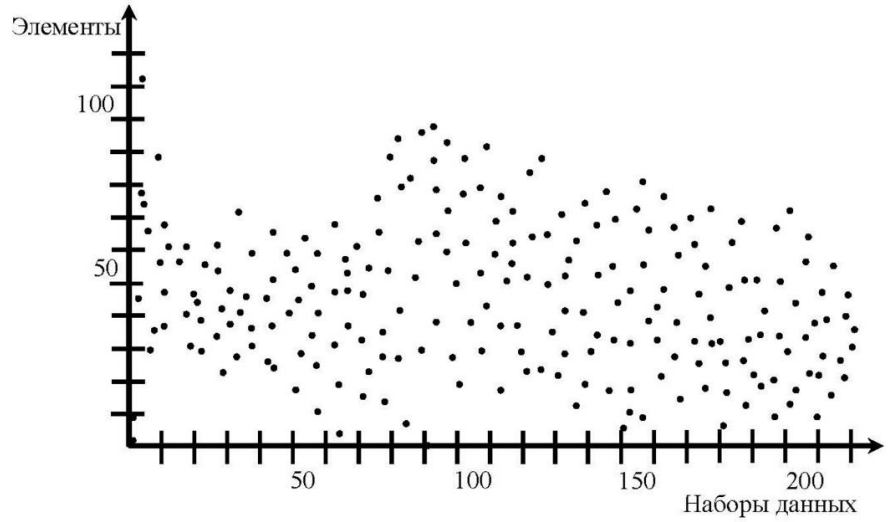

Рис. 6. Наборы данных и число элементов для микросхемы К1921ВГ1Т

Исследована проблема адаптивности алгоритмов сопоставления данных. Размер окна равен числу элементов в наборе данных, который, в свою очередь, связан с блокировкой. То есть, используя характеристики наборов данных, рассмотрена возможность адаптивно контролировать блокировки. Результаты исследований в следующем:

1) определена важность адаптивности в проблеме сопоставления и исследование в деталях использования классического метода сортированного соседства;
2) разработаны и апробированы методы адаптивности сортированного соседства, показывающие значительные улучшения сопоставления;
3) определены методы адаптивности сортированного соседства.

Поскольку предлагаемое является адаптивной структурой, то же самое применяется к другим алгоритмам сопоставления наборов данных в Системе.

Для неадаптивного метода сортированного соседства используется окно фиксированного размера w. На рис.7 представлено такое перемещение окна, где: w=w1=w2. Окна перекрывают наборы данных за счет фиксированного размера.

В адаптивном методе сгенерированные окна будут иметь разные (то есть адаптивные) размеры.

Рис. 7. Окна при неадаптивном случае сопоставления данных

Предположим, что два элемента, E1n и Ein, являются первым и последним в одном и том же наборе данных n, состоящего из in элементов. В это время элемент $E1n_1$ находится в следующем наборе данных n_1, состоящим из in_1 элементов. Тогда может выполняться условие: расстояние(E1n, Ein) ≤ φ < расстояние(E1n, $E1n_1$), где φ - это минимальный порог расстояний. То есть, если блокировка выполнена, тогда позиция элемента вне набора данных должна отличаться от элементов внутри набора данных. Чтобы определить размер окна, необходимо найти и определить пограничные элементы (пограничные пары) между соседними наборами данных.

Чтобы найти пограничные пары, надо сравнить расстояние для каждого соседнего набора данных. Если расстояние между пограничными элементами больше порогового значения, это и есть пограничная пара. После нахождения всех пограничных пар весь отсортированный список наборов данных можно разделить на неперекрывающиеся наборы данных различных размеров.

Например, на рис. 8 элемент $E1n_{-1}$ является первым для набора данных n_{-1}, а элемент E1n - первым для набора данных n. Расстояние ($E1n_{-1}$, E1n) > φ. Необходимо найти только те наборы данных, расстояние в которых до следующей соседнего набор данных больше порога φ. Для предыдущего набора данных n_{-1}, минимальный порог расстояний: $φ_{-1}$= расстояние($E1n_{-1}$, E1n). Для текущего набора данных n, φ= расстояние(E1n, $E1n_1$).

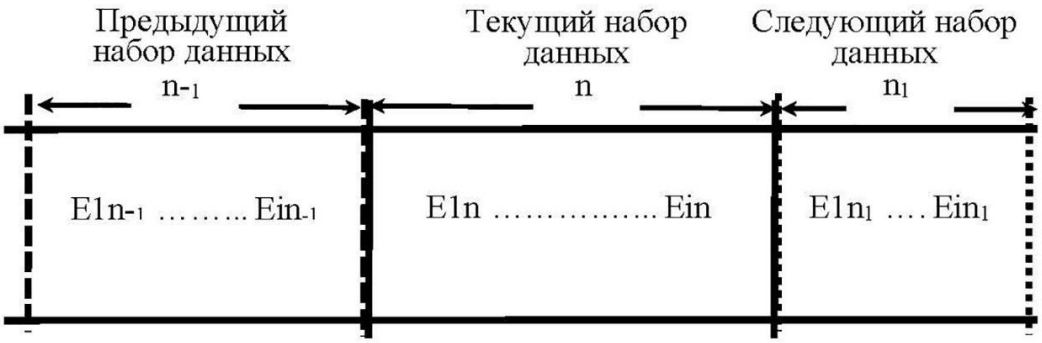

Рис. 8. Пример пограничных элементов

Чтобы сократить количество сравнений используется адаптивная настройка окон для нахождения границ пар со значительно меньшим количеством приблизительных сравнений. Это выполняется методом увеличения размеров окна для сбора как можно большего числа потенциально схожих наборов данных по количеству элементов в них и уменьшения размеров окна при сопоставлении элементов для отсечения несопоставимых наборов данных.

Для реализации метода необходимо провести измерение расстояний между элементами наборов данных, и провести увеличения/уменьшения размера окна.

Рассмотрим случай для текущего окна шириной w, предыдущего – w_{-1} и следующего w_1, показанный на рис. 9.

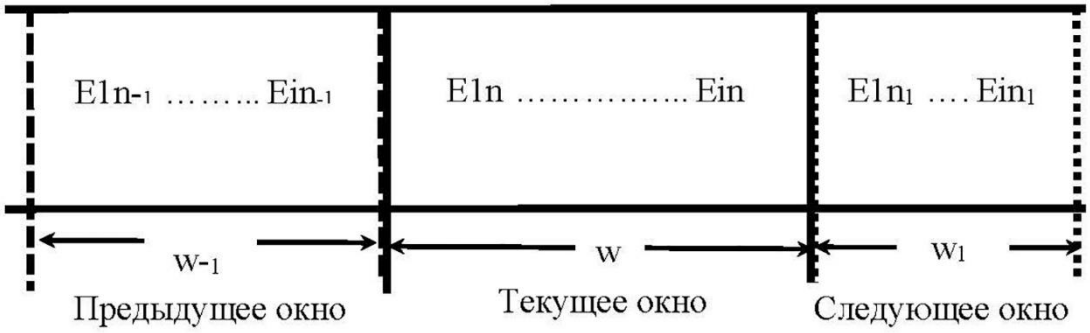

Рис. 9. Пример метода адаптивного сортированного соседства

Расстояние между первым элементом предыдущего окна и первым элементом текущего окна удовлетворяет условию: расстояние(E1n-1, E1n) < φ, где φ - это порог расстояния. Это расстояние указывает на то, крайние элементы в текущем окне близки друг к другу, поэтому еще есть возможность увеличить размер окна для поиска большего количества сопоставляемых элементов. В противном случае окно должно быть уменьшено. Теперь необходимо измерить величину изменения окна. Расстояние($E1n_{-1}$, E1n)/w_{-1} представляется как среднее расстояние между элементами внутри окна w_{-1}. Выбирается следующая ширина окна, основываясь на следующее: расстояние(E1n, $E1n_1$) * w/ w_1 = φ. Таким образом, ширина окна на следующем шаге определяется из ширины текущего окна. Используя вышеуказанную формулу как инвариант, можно адаптивно изменять ширину окна для оптимизации сопоставления данных (при риске увеличения времени выполнения). Регулировка ширины окна завершается, на крайнем наборе данных. На каждом шаге регулируется только ширина окна, и стартовая точка окна не меняется. Поэтому используются инкрементально регулируемые окна, чтобы приблизиться к оптимальной их ширине.

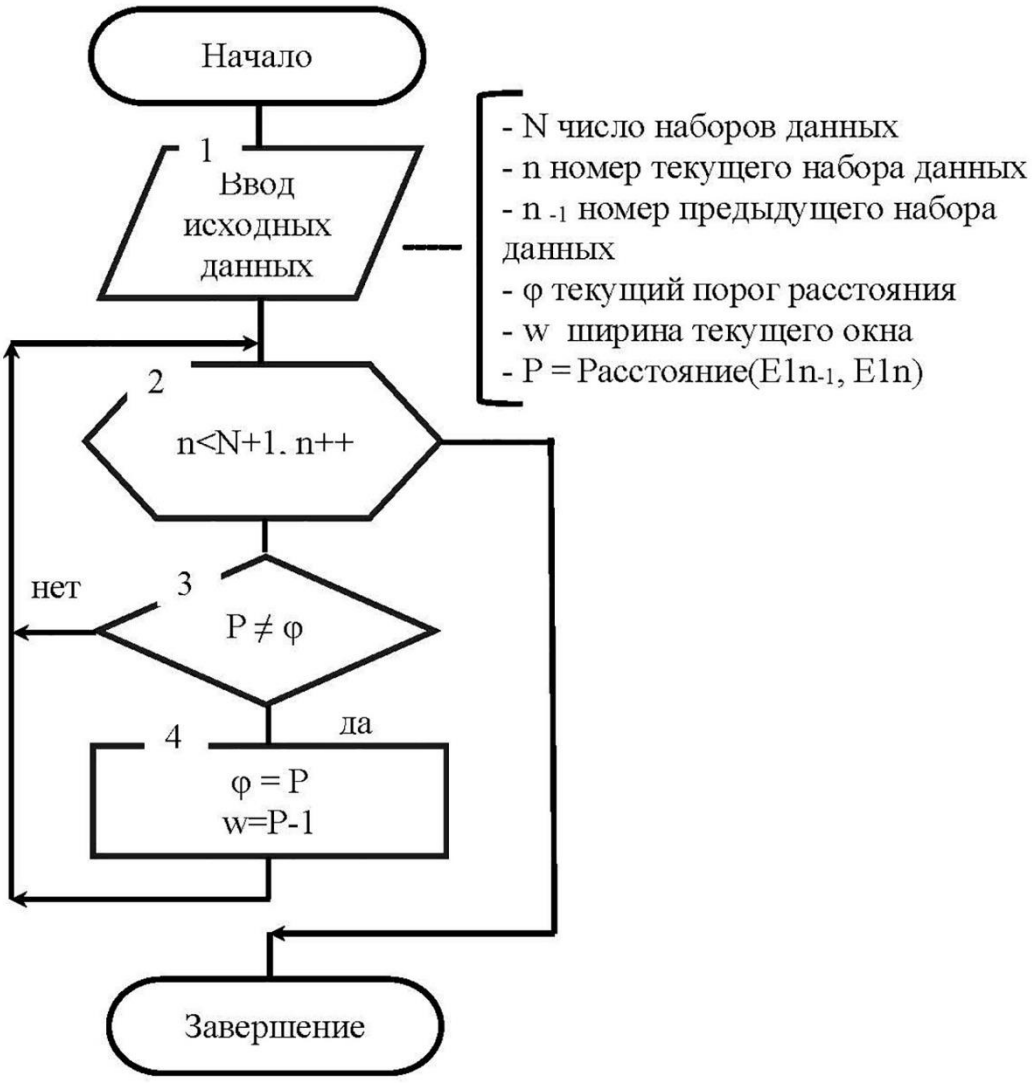

Рис.10. Алгоритм метода адаптивного сортированного соседства

Для применения вышеуказанных результатов исследований, адаптивная настройка окна для реальной задачи блокировки начинается с минимальной ширины и далее продолжается процесс ее регулировки (увеличение и уменьшение). В зависимости от выбора ключа и рассчитанного расстояния, после первого уменьшения может произойти колебание в регулировке ширины окна. Чтобы предотвра-

тить потенциальное колебание в регулировке ширины окна, которое может произойти в этом методе, после первого уменьшения необходимо привести бинарное сравнение, чтобы найти пограничную пару. Другими словами, на этапе бинарного сравнения, уменьшается ширина окна вдвое каждый раз при регулировке ширины окна. Алгоритм метода в соответствие с выше представленным описанием показан на рис. 10.

При этом, в алгоритме, представленном на рис. 10, блок 4, корректировка ширины окна экспериментально была проведена тремя способами:

1) линейно: путем добавления или вычитания постоянного значения α к предыдущей ширине окна;

2) геометрически: путем умножения или деления постоянного значения α на предыдущую ширину окна;

3) объемно: для текущего шага определяется с использованием ширины нескольких предыдущих окон.

Наиболее эффективным явился третий способ. Метод, лежащий в основе объемного способа, заключается в том, что на этапе изменения ширины окна на текущем шаге используется усредненная ширина окон, уже определенных до него. Например, для рис.9, $w_1=(w+w_{-1})/2$.

Для проверки качества блокировки, для адаптивного метода сортированного соседства, разработаны две метрики, необходимые для количественной оценки эффективности и качества метода:

1) коэффициент сокращения, определяется как (11):

$$К_{\text{сокр}} = 1 - \frac{N_b}{A \times B}, \quad (11)$$

где $N_b \leq A \times B$ - количество пар элементов, не удаленных блокировкой. Это метрика относительного сокращения пространства сопоставления, но без учета качества сокращения.

2) полнота пар сопоставления, измеряется как (12):

$$П_{\text{пар}} = \frac{\text{Пл}}{\text{Пл+Лп+Вс}}, \quad (12)$$

где Пл - общее количество правильно классифицированных истинно сопоставленных пар данных в пределах окна, а Пл+Лп+Вс - общее количество совпадений.

Существует компромисс между этими двумя метриками. Желательно достичь наивысшего коэффициента сокращения при наивысшей полноте пар элементов данных. Чтобы проверить влияние этого компромисса в методологии блокировки, проверена работа алгоритма сопоставления для разных временных интервалов поступления данных от различных источников. Результат показал, что единственным ограничением ширины окна является требуемое время вычислений для алгоритма.

Роль блока 4 на рис. 4 – формирование весов. При сопоставлении двух наборов данных, обозначается шаблон согласия для j-го сравнения как γ_j. Бинарный индикатор сопоставления для i-го поля связи j-го сравнения обозначается как γ_{ij}, кодируется как 1 для сопоставленных и ноль для несопоставленных данных. Например, индикатор сопоставления для наименования изделия (i=1) и кода классификатора (i=2) дает профиль сопоставления (13):

$$\gamma_j = [\gamma_{1j}, \gamma_{1j}] \quad (13)$$

Предполагая, что истинный статус соответствия всех наборов данных известен, условная вероятность того, что пара элементов соответствует шаблону сопоставления - γ_j, при условии, что это истинное сопоставление, обозначается как (14):

$$m_j = P(\gamma_j = 1 | M_j = 1) \equiv \text{P}(\gamma_j | M_j) \quad (14)$$

Аналогично, условная вероятность того, что пара элементов имеет шаблон сопоставления γ_j, при условии, что они являются Ло данными, обозначается как (15)

$$u_j = P(\gamma_j = 1 \mid U_j = 1) \equiv \text{P}(\gamma_j | U_j) \quad (15)$$

Отношение m_j и u_j (m_j/u_j) является отношением правдоподобия и составляет основу веса сопоставления.

Вероятности m_j и u_j называют m- и u- вероятностями, они могут быть условно независимы, и представлены как (16):

$$m_j = P(\gamma_{1j}|M_{1j})\mathrm{P}(\gamma_{2j}|M_{2j}) \quad (16)$$

и как (17)

$$u_j = P|\mathrm{P}(\gamma_{2j}|U_{2j}) \quad (17)$$

Условная независимость вероятностями m_j и u_j, , является ключевым теоретическим утверждением. Например, если обозначения изделия совпадают, код классификации изделия с большей вероятностью будет совпадать. Поэтому, веса совладения считаются точными.

M-вероятность концептуализируется как индикатор качества данных. Например, частота ошибок данных в наборах, которые действительно совпадали, была известна. Элемент сопоставления бинарный (например, логические значения 0 и 1 для входа и выхода в 2 165 логических микросхемах. Эта частота ошибок данных составляла 20% для пары элементов набора данных. В этом случае можно было бы ожидать, что 64% (0,8 × 0,8 = 0,64) совпадений будут правильно соответствовать, а 4% совпадения будут неправильно соответствовать (0,2 × 0,2 = 0,04), что приводит к m-вероятности 68% для логического значения входа и выхода в логической микросхеме. Если совпадающие элементы не являются бинарными, то вероятность того, что они будут неправильно соответствовать, вероятно, ближе к нулю, чем 4%. Расхождения в оставшихся 32% пар элементов, т.е. ложноотрицательные результаты, могут быть вызваны ошибками сбора данных, или их отсутствием.

Случайное совпадение между двумя элементами набора данных определяется u-вероятностью. При чем, они на самом деле не имеют сопоставления. Это может быть концептуализировано и упрощено как случайное согласие, используя следующую логику. Предположим, что два набора данных содержат по сто элементов каждый. Тогда полное сравнение приведет к десяти тысячам потенциальных сравнений, из которых сто совпадениями сравнений могут быть Лп. Таким образом, девять тысяч девятисот сравнений не совпадают. Поскольку несовпадающие пары составляют большинство сравнений, часто предполагается, что все сравнения являются частью несопоставленного множества. Предполагая, что могут быть некоторое повторение, становится вполне естественным исследовать частоты в каждом совпадающем, а также вероятность того, что пара элементов будет совпадать только случайно. Как m-, так и u-вероятности могут быть скорректированы в зависимости от уникальности (частоты) совпадающих элементов.

Формально, если частоту повторения элементов в первом наборе определить как f_1, f_2, …, f_K и частоту повторения элементов во втором наборе определить как g_1, g_2, …, g_K, количество повторяющихся элементов в первом наборе данных равно (18):

$$N_1 = \sum_{k=1}^{K} f_k \quad (18)$$

во втором наборе данных равно (19):

$$N_2 = \sum_{k=1}^{K} g_k \quad (19)$$

В наборах данных одинакового размера и сопоставления можно определить истинную частоту совпадающих пар h_1, h_2, …, h_K, где количество элементов в истинном наборе совпадений S равно N_s (20):

$$N_s = \sum_{k=1}^{K} h_k \quad (20)$$

Следовательно, m- вероятность с поправкой на частоту равна (21):

$$m_f = \frac{h_k}{N_s} \quad (21)$$

А u-вероятность с поправкой на частоту равна (22):

$$u_f = \frac{(f_k + g_k - h_k)}{N_1 + N_2 - N_s} \quad (22)$$

Предполагается, что u-вероятности являются безусловной вероятностью случайного совпадения, такого что (23):

$$u_f = \frac{(f_k + g_k)}{N_1 + N_2}, \quad (23)$$

независимо от статуса сопоставления.

Несмотря на кажущуюся простоту расчета весов сопоставления, скорректированных или нескорректированных для их относительных частот, требуется знание истинного статуса сопоставления. Истинный статус сопоставления двух наборов данных редко известен, и поэтому m- и u-вероятности оцениваются с использованием предыдущего опыта и предполагаемого набора данных.

Для оценки m- и u-вероятностей сопоставления для i-го набора данных и для j-го элемента можно построить общий вес сопоставления, обозначаемый $R(\gamma j)$, который определяется с использованием отношения m- и u-вероятностей (24):

$$R(\gamma_f) = \frac{P(\gamma_f|M_f)}{P(\gamma_f|U_f)}, \quad (24)$$

когда данные сопоставляются, и (25):

$$R(\gamma_f) = \frac{1-P(\gamma_f|M_f)}{1-P(\gamma_f|U_f)}, \quad (25)$$

когда данные не сопоставляются. Эти отношения являются положительными и отрицательными отношениями сопоставления.

Предполагая, что сопоставляемые i-е наборы данных условно независимы, вес сопоставления можно выразить как отношение произведений m- и u-вероятностей для j-го элемента (26):

$$R(\gamma_f) = \frac{P(\gamma_f|M_f)}{P(\gamma_f|U_f)} = \frac{\prod_i P(\gamma_f|M_f)}{\prod_i P(\gamma_f|U_f)}, \quad (26)$$

Использование логарифма по основанию 2 отношения (26), упрощает расчет и облегчает интерпретацию весов сопоставления (27):

$$\log_2 R(\gamma_f) = \log_2 \left(\frac{P(\gamma_f|M_f)}{P(\gamma_f|U_f)}\right) \quad (27)$$

Кроме того, веса сопоставления можно корректировать для сложных шаблонов $R(\gamma' j)$. Описанный выше индикатор сопоставления (γij) кодировался как ноль (несопоставление) или единица (сопоставление), тогда как сложные индикаторы сопоставления $(\gamma' ij)$ могут принимать любое значение от ноля (полное несопоставление) до единицы (полное сопоставление), где значения больше 0 и меньше 1 указывают на частичное сопоставление. Веса сопоставления, основанные на сложных шаблонах, рассчитываются путем вычитания разницы между весами, когда данные сопоставляются и не сопоставляются (28).

$$\log_2 R(\gamma'_f) = \log_2 \left(\frac{P(\gamma_f|M_f)}{P(\gamma_f|U_f)} - \frac{1-P(\gamma_f|M_f)}{1-P(\gamma_f|U_f)}\right) \quad (28)$$

Завершающая операция заключается в определении двух порогов, которые классифицируют сопоставления по трем категориям: сопоставление есть, сопоставления нет и потенциальное сопоставление. Можно сгенерировать два разных порога, используя распределение весов сопоставления, $log_2 R(\gamma j)$ или $log_2 R(\gamma j)$, но их сложно интерпретировать. Поэтому предпочтительнее определять статус по шкале вероятностей. Используя теорему Байеса, формируется $R(\gamma j)$ или $R(\gamma' j)$ и групповые априорные вероятности совпадения $P(M)'/(1- P(M)')$, которые преобразуются согласно выражения (29):

$$P(M') = \frac{N_e}{N_1 N_2}, \quad (29)$$

где Ne - количество ожидаемых совпадений между элементами наборов данных с N1 и N2, определенных в (18) и (19). Апостериорное отношение вероятностей определяется в выражении (30):

$$\frac{P(M_f|\gamma_f)}{P(U_f|\gamma_f)} = \frac{P(\gamma_f|M_f)}{P(\gamma_f|U_f)} \times \frac{P(M')}{P(U')} \quad (30)$$

Таким образом: вероятность того, что данные совпадают, можно рассчитать следующим образом (31):

$$P(M'') = \frac{\frac{P(M_f|\gamma_f)}{P(U_f|\gamma_f)}}{1+\frac{P(M_f|\gamma_f)}{P(U_f|\gamma_f)}} \quad (31)$$

Применение выражения (31) к сложным весам сопоставления, приводит к апостериорным вероятностям сопоставления.

Представленное выражение для сопоставления дает вероятность близкую к единице, это оправдывает использование совпадений и несовпадений для качества исходных данных.

Точное размещение порогов, используемых для определения статуса связи, может быть вопросом проб и ошибок. Необходимость максимизировать чувствительность обнаружения совпадений, несомненно, потребует более рутинную обработку данных по сравнению с порогом, которая оптимизируется за счет настройки параметров.

Роль блока 3 на рис. 4 – блокировка и стратификация. Размер сопоставляемых наборов данных в процессах сбора быстро увеличивается. Поэтому используется блокировка или стратификация. Этот процесс включает в себя разбиение данных на части, что изначально представлено как ограничение сопоставлений подпространством. Например, если выбираемые изделия ЭКБ – полупроводниковые приборы - ограничители напряжения, проводится сопоставление только по ним. Это означает, что Система ищет только совпадающие данные в пределах полупроводниковых приборов - ограничителей напряжения. Разделение наборов данных значительно сокращает пространство сопоставления; например, попытка сопоставить наборы данных, каждый из которых содержит 10 000 элементов, поступающих от 10 источников, приведет к сопоставлениям миллиона элементов, в отличие от незаблокированного сопоставления, которое приведет к десяти миллионам сопоставлений. Тем не менее, при блокировке существует явный компромисс между размером наборов данных и возможностью полностью исследовать набор данных в поисках потенциальных совпадений, с явным предположением, что данные, не входящие в набор, не будут совпадениями.

При выполнении блока 5 представленного на рис. 4 алгоритма реализуется поиск ошибок сопоставления. После создания сопоставленного набора данных важно учитывать качество сопоставления и как это может повлиять на результаты нормализации данных, т.е. сколько неправильных ошибочных сопоставлений выполнено и сколько Пл пропущено и к каким последствиям это может привести. Выполнение таких действий может считаться нелогичным, потому что если бы априори знать истинное состояние сопоставлений, не было бы необходимости проводить вероятностное сопоставление.

Существует ряд различных подходов, которые могут быть использованы для количественной оценки коэффициента ошибок сопоставления, включая:

1) сравнение с требованиями стандартов;
2) анализ чувствительности;
3) опоставление совпадающих и несовпадающих данных;
4) выявление неправдоподобных совпадений.

Сравнение с требованиями стандартов является наиболее интуитивно понятным методом установления ошибок сопоставления. Сравнение вероятностно сопоставленного набора данных с требованиями стандартов приведет к таблице ошибок сопоставления. После создания такой таблицы можно рассчитать и составить таблицу простых статистических данных, таких как чувствительность, специфичность и положительные или отрицательные прогностические значения.

Структурированный анализ чувствительности может быть использован для того, чтобы увидеть, как изменения m- и u-вероятностей влияют на количество потенциально Л. Сравнение данных полезно для определения того, легче ли сопоставлять одни наборы данных по сравнению с другими. Например, предполагая, что существуют некоторые общие наборы данных, не используемые в качестве идентификаторов сопоставления, такие как электрические параметры, можно сравнить коэффициенты сопоставления. Аналогичным образом, изучение того, как коэффициенты сопоставления изменяются во времени, может быть полезным индикатором смещений, зависящих от времени.

Выявление неправдоподобных совпадений возможно только в определенных сценариях. Предположим, например, что сопоставление используется для анализа информации об отказах изделий, собираемых в рамках рекламаций. Есть отказы изделий, имеющие свойство исчезать и больше не проявляться. Есть свойство изделий к самовосстановлению. По таким причинам может быть неприемлемая для выбора к применению модель надежности изделий: параметры соответствуют заявленным, отказ, работоспособность нормальная и т.д. По этим причинам могут возникнуть вопросы относительно качества сопоставления или достоверности собранных данных.

Результаты практической реализация алгоритма вероятностного сопоставления

Два основных шага в сопоставлении пары элементов набора данных, относящихся к конкретному изделию ЭКБ, включают: во-первых, операцию поиска, в ходе которой потенциально сопоставляемые элементы набора данных рассматриваются для тщательного изучения, за которым, во – вторых, следует детальное сопоставление, чтобы решить, относится ли данные к одному и тому же изделию ЭКБ, и являются ли они обновлением конкретного набора данных об изделии.

Цель первого шага состоит в том, чтобы сократить до приемлемого уровня количество неудач в сопоставлении потенциально сопоставляемых пар элементов, при этом до минимума сокращая число ошибок. На втором шаге два вида ошибок: принятие ложных сопоставлений и отклонение потенциальных подлинных, должны быть настолько редкими, насколько это практически осуществимо. А соотношение между ними должно быть скорректировано до получения наилучшего компромисса.

Дальнейшее упорядочение того, что решение о принятии или отклонении сопоставления основывается на количественной оценке шансов того, что пара элементов данных относится или не относится к одному и тому же изделию, делает возможным предварительно принять небольшую часть сомнительных отношений, которые впоследствии могут быть выделены для визуального изучения. Алгоритм визуализации таких элементов данных реализован в компьютерной программе, которая используется в Системе и получила свидетельство о регистрации Роспатента [6].

При разработке вариантов оптимизация как шагов поиска, так и шагов сопоставления данных, алгоритмы лучше всего основываются на количественных исследованиях эффективности альтернативных процедур, в вышеуказанных терминах, и дискриминационных возможностях, присущих данным, общим для сопоставляемых элементов.

Когда поиск сопоставляемых элементов основан на алгоритме детерминированного сопоставлении, любая ошибка в написании может привести к невозможности найти необходимый элемент данных. Поэтому при вероятностном сопоставлении область поиска расширяется, включая более распространенные формы представления наборов данных, временно откладывая в сторону те части информации, которые, наиболее подвержены ошибкам. Например, дробной части значений параметров изделий.

Для реализации такой функции используется показатель частоты обновления данных. С высокой частотой обновления являются наборы с большим количеством значений, например, параметры изделий ЭКБ при различных условиях эксплуатации. Примерами наборов данных с небольшой частотой обновления являются данные о наработке до отказа изделий ЭКБ. Данными с нулевой частотой обновления являются обозначения изделия, классификационные коды и функциональное назначения. В алгоритме сопоставление учитывается такая особенность и применение показателя частоты обновления сокращает затраты на сопоставление.

Дальнейшее подразделение в последовательности элементов может быть основано на дополнительных условиях, при этом, выбираются только самые надежные части наборов данных.

Для проведения исследований по информации об аналогах изделий ЭКБ [7], элементы были расположены в порядке пар. Обнаружено, что данные имеют до 14 процентов расхождений в вариантах указания аналогов. Потери потенциального сходства были сокращены с 14 до 4 процентов от общего числа сопоставлений (т. е. с 640 до 183 в исследовании для 4576 типономиналов микросхем).

Степень сопоставления для областей применения изделий ЭКБ, достигнутая парным сопоставлением элементов, была такова, что в наборах для 7845 типономиналов полупроводниковых приборов, 3 процента имели уникальные названия областей применения, отличающиеся от данных для этих же изделий в других наборах. Это составляет 235 типономиналов полупроводниковых приборов из 7845.

Значительная часть машинного времени в Системе при проведении нормализации посвящена детальным сравнениям пар элементов данных. При этом, основным показателем общей эффективности Системы с точки зрения стоимости для заданной точности определено соотношение истинных положительных (правильных) сопоставлений Пл и ложных положительных сопоставлений Лп (32).

$$Э_{\text{общая системы}} = \frac{Пл}{Лп} \quad (32)$$

Относительная эффективность альтернативных видов идентификационной информации, как

средств разбиения наборов данных на более мелкие единицы или элементы, исследована для разработанного алгоритма. Схожая степень разбиения достигается путем замены функционального назначения изделия на классификационный код Классификатора изделий и конструкторских документов ОК 012 - Классификатора ЕСКД, Одно или другое из обозначений функционального назначения и классификационного кода будет отличаться на 7 процентов потенциально сопоставляемых пар элементов данных для 4576 типономиналов микросхем, что составило 320 пар элементов. Это приведет к относительно большому увеличению неудач в достижении сопоставления без компенсирующего сокращения доли непродуктивных сопоставлений.

В алгоритме вероятностного сопоставления применены альтернативные методы кодирования. Так, для сокращения числа неудачных сопоставлений потенциально сопоставляемых пар элементов, потери значительно уменьшены путем сортировки наборов данных в более чем одну последовательность. Каждая из которых основана на различных идентифицирующих деталях. Таким образом, временно игнорируются элементы, которые могут быть ошибочными. Альтернативы кодирования основной последовательности, были основаны на пропуске сначала одной, а затем другой части набора данных с присвоением им локальных кодов.

Сопоставление в алгоритме оптимизировано. Эффективность достигнута посредством комбинированного использования одной или нескольких альтернатив основной последовательности, и описана путем сравнения сокращения числа потерь с затратами в терминах возросшего числа нерезультативных сопоставлений данных. В проведённых исследованиях потерянные связи были сокращены таким образом на порядок величины, с 14 до двух процентов, при двукратном увеличении общего числа сравнений элементов.

В ручном режиме решение о том, что два набора данных, которые сопоставляются, относятся или не относятся к одному и тому же изделию, обычно принимается субъективно. Однако правила, по которым производится такая человеческая оценка, в целом более просты, чем можно было бы предположить на первый взгляд. Также существует дополнительная возможность, в случае компьютерного сопоставления, сделать суждение количественным, так что в случае сложных решений степень уверенности может быть выражена численно. При упрощенном подходе можно просто отметить в паре набора данных, сколько из различных элементов сопоставляются, а сколько не сопоставляются, таким образом, придавая равные веса каждому элементу.

Такая процедура при проведении исследований оказалась достаточно затратной, и значительно усовершенствована. Например, сопоставление начальной буквы К (для условного обозначения микросхем) в паре данных увеличит уверенность в том, что в обеих наборах упоминается изделие категории качества «1» (или «ОТК»), но не в такой степени, как согласие начальной буквы Б (для бескорпусных микросхем), что встречается гораздо реже. Предположив, что первые буквы К и Б имеют частоты 1/16 и 1/1024 соответственно, можно сделать вывод, что шансы случайного совпадения будут 16:1 и 1024:1 соответственно. Чтобы преобразовать эти отношения в весовые коэффициенты, необходимо выразить их в виде логарифмов; и логарифмы по основанию 2 оказались удобными для этой цели. Таким образом, совпадения первых букв К и Б в приведенном выше примере можно было бы рассматривать как несущие вес +4 и + 10 (представляющие шансы 2^4:1 и 2^{10}:1). Таблицы таких значений, известных как бинарные (двоичные) веса, были составлены для всех условных обозначений микросхем, и процедура была расширена для включения других элементов идентификации. Аналогичные разногласия в данных будут иметь отрицательные веса, и можно использовать тот же принцип. Элемент, который редко расходится в сопоставленных парах элементов наборов данных, будет, когда он не совпадает, иметь гораздо больший отрицательный вес, чем тот, который сомнительный. Например, несовпадение первых букв, если оно произошло в 1/64 из сопоставляемых пар элементов, будет иметь вес минус шесть (указывая на шансы 1 : 2^6 против подлинного положительного совпадения). Для получения весов использованы отношения (27) и (28). А вероятность – (31), для наборов данных и элементов в парах, собранных вместе, конкретного сопоставления или несопоставления. Например, применительно к выводу одного веса, который должен быть прикреплен к соглашениям об условных обозначениях изделий, начинающихся с цифры, где первая буква отсутствует, и к выводу таблицы весов, специфичных для каждой

буквы алфавита и цифры для условного обозначения изделий ЭКБ. Этот подход может быть использован для информации даже самого низкого качества.

Использование весов показано на примере в таблицах 1 и 2 пары элементов сопоставление параметров микросхемы триггера К555ТМ2, относящихся к последовательным событиям.

Таблица 1

Сопоставление параметров микросхемы триггера К555ТМ2

Номер элемента в паре	Напряжение питания	Ток потребления	Потребляемая мощность	Выходное напряжение низкого уровня
1	5 В ± 5%	≤ 8 мА	42 мВт	≤ 0,5 В
2	5 V ± 5%	<6 mA	40 mW	100 mV

В таблице 1 показаны, результаты измерений выходного контроля у изготовителя и результаты измерения входного контроля у потребителя. Изготовитель изделия и потребитель являются источниками информации для Системы. Информация поступает об одном и том же изделии - схемы цифрового устройства – два D -триггера с динамическим управлением, микросхемы К555ТМ2. Независимо от давности срока разработки, микросхема более 30 лет успешно применяется в РЭА. В процессе изготовления совершенствуются технологии, применяемые материалы. Поэтому информации о микросхеме может незначительно изменяться. При этом, такое незначительное изменение может иметь значение для принятия решения на ее применение в РЭА. Условное графическое обозначение микросхемы К555ТМ2, для представления информации о параметрах, показано на рис. 11.

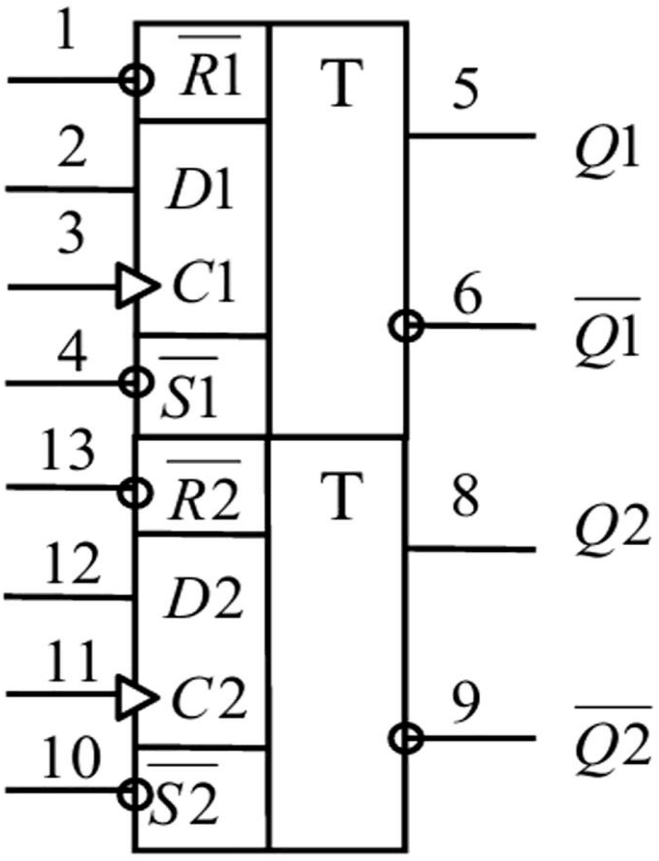

Рис.11. Условное графическое обозначение микросхемы К555ТМ2

Согласно рис. 4 у микросхемы К555ТМ2 два независимых D-триггера, срабатывающих по фронту тактового сигнала на входе С. Низкий уровень напряжения (логический ноль) на входах установки (S) или сброса (R) устанавливает выходы триггера в соответствующее состояние вне зависимости от состояния других входов (С и D). По входным и выходным уровням сигналов, микросхема совместима с другими микросхемами стандартной ТТЛ логики. Имеет импортный аналог: SN74LS74 фирмы Texas Instruments, Inc. И может использоваться для замещения.

Таблица 2

Расчет бинарных весов для пар элементов данных представленных в таблице 1

Виды сопоставлений	Напряжение питания	Ток потребления	Потребляемая мощность	Выходное напряжение низкого уровня
Пл	+	-	-	-
Лп	-	-	+	-
Ло	-	-	-	-
От	-	-	-	+
Вс	-	+	-	-
Вн	-	-	-	-

Сумма бинарных весов в таблице 2 указывает на то, что шансы в пользу подлинных сопоставлений находятся в районе четырех к одному. Если бы элементы данных не относились к одному и тому же изделию, многие не совпадали бы, и общий вес был бы резко отрицательным. Добавление весов предполагает отсутствие корреляции между расхождениями в различных элементах идентификационной информации. Некоторая корреляция была, по сути, обнаружена в собранных данных, но она была слишком мала, чтобы привести к неверным решениям, за исключением редких случаев.

Однако исправления можно провести автоматизировано, что позволило бы сократить случайные ошибки источника. На практике было обнаружено, что только очень небольшая часть связей имеет бинарных веса, которые можно было бы считать пограничными. Этот класс представляет особый технический интерес, поскольку он указывает на необходимость дополнительных уточнений. Бинарные веса допускают их применение к перекрестным сопоставлениям и к сравнениям цифровых значений параметров во время двух разных событий, допуская округление. Информация, которая особенно подвержена изменениям, часто ошибочно игнорируется как малополезная для сопоставлений. Когда такая информация не сопоставляется от двух и более источников, этот факт обычно имеет небольшой вес, но когда есть сопоставление, вес становится значимым. Таким образом, в наборах данных, используемых в настоящем исследовании, разница в местах двух последовательных событий, относительно не важна, поскольку в измерение параметров изделий могла быть ошибка. Кроме того, когда события происходят в разных местах и данные поступают от разных источников, сопоставление имеет небольшой вес, поскольку треть всех таких ошибок может допускаться.

Значимым примером использования параметров, которые обычно не считаются идентифицирующей информацией, является сравнение наработки до отказа. Если обнаруживается что изделие отказало до указанного параметра – это показывает снижение его надежности. И сдвигает назад его место в выборке на применение в РЭА.

В нашем исследовании использовалось около 250 различных весовых коэффициентов, а операции сопоставления проводились в Системе с более чем 10 000 наборов данных. Лп и Ло, возникающие из-за отсутствия дискриминационной способности, были редки. Используя коды классификации, обозначение, категорию качества, Лп и Ло составили 0,12 и 0,60 процентов соответственно на тысячу пар сопоставления. Когда также использовались сопоставления функционального назначения, эти цифры были уменьшены примерно в пять раз до 0,025 и 0,10 процентов, а включение области применения изделий в сопоставление привело к дальнейшему существенному снижению.

Скорость комбинированных операций поиска и сопоставления в Системе составляла около 3000

в минуту, 10-кратное увеличение скорости до примерно 30000 комбинированных операций поиска и сопоставления в Системе в минуту кажется разумной перспективой для дальнейшего применения более современных вычислительных средств.

Заключение

В настоящей статье объясняются и показаны результаты применения метод вероятностного сопоставления данных об изделиях ЭКБ, собираемых от нескольких источников и алгоритмы их реализации. Представлена методология, которая использует сходство данных в качестве весов. Представлена методология блокировки данных, основанная на скользящем, с изменяемой шириной, окне, которая обеспечивает управляемую вычислительную сложность в условиях больших данных. Методология, разработанный на ее основе алгоритм и программные средства применяется в Системе и дают ожидаемый эффект.

Метод вероятностного сопоставления является многообещающим при работе с параметрами и показателями изделий и его применимость выходит за рамки точного сопоставления. Он позволяет сопоставлять нестандартизированные гетерогенные наборы данных. С применением метода получено более 80% Пл.

В то же время как детерминированное сопоставление улучшает качество базы данных (БД), вероятностное сопоставление увеличивает размер БД. Алгоритмы вероятностного сопоставления менее точны, чем детерминированные, поскольку они указывают связи между различными наборами данных от различных источников. Эксперименты в этой части были проведены с помощью разработанных и зарегистрированных в Роспатенте БД и программ по ее формированию и управления данными [8-10].

Также в статье показаны результаты исследования процессов, лежащий в основе детерминированного и вероятностного сопоставление. Несмотря на кажущуюся сложность вероятностного сопоставления, процессы можно разбить на относительно небольшое число простых операций по манипулированию данными. Кроме того, статистические принципы, лежащие в основе весовых расчетов, используемых для определения сопоставлений, выведены из теоремы Байеса (29), которая применена в настоящих исследованиях.

Используя простой пример, проиллюстрированы критические шаги и предположения, лежащие в основе вероятностного сопоставления. К ним относятся:

1) проблемы при сравнении наборов данных различной длины;
2) допущение условной независимости (17) при расчете весов сопоставления;
3) выбор размера набора данных, который влияет на вычислительную нагрузку при сопоставлении;
4) выбор пороговых значений сопоставлений перед принятием или отклонением пар элементов, требующих проверки;
5) процессы очистки данных перед сопоставлением, которые выполняются в целях увеличения поиска совпадающих данных.

Преимуществами вероятностного сопоставления данных являются уменьшение количества пропущенных данных; улучшенная классификация с использованием сопоставленных данных. Однако, дальнейшие исследования в области вероятностного сопоставления данных требует решение вопросов анализа эффективности сопоставления наборов данных и эффективности автоматизированного выбора сопоставленных и несопоставленных наборов данных в том числе с применением технологий ИИ.

С помощью разработанного алгоритма проведено надежное вероятностное сопоставление данных, и показатели результатов достаточны для функционирования Системы [11]. Вероятностное сопоставление данных максимизирует ценность регулярно собираемой информации за счет улучшения связи между данными, представляющими интерес, что, в свою очередь, увеличивает объем недостающих данных, тем самым усиливает результаты сопоставления.

Сопоставление данных, будь то вероятностное или детерминированное, с примыканием технологий ИИ [12], будет приобретать все большее значение по мере быстрого расширения объема регулярно собираемых данных об изделиях ЭКБ. Перспективы усовершенствования алгоритма вероят-

ностного сопоставления - в непрерывном анализе данных в связи с увеличением объёма информации об изделии ЭКБ. Чтобы избежать невозвратных потерь, разумно предварительно принять часть сопоставлений, связанных с общими весами, которые являются пограничными и редактировать их позже, если они, не совсем вписываются в группу сопоставляемых наборов данных. Например, разночтение в связанных параметрах наработки до отказа и интенсивности отказа для одного и того же изделия дает дополнительное доказательство того, что рассматриваемый набор данных не может принадлежать к одному и тому-же изделию. Однако, эти данные можно принять, а при следующем сопоставлении проверить. Так же как если один из параметров пустой, это указывает на то, что сопоставление может иметь статус Лп.

В статье показаны результаты исследования правил, которые могут использоваться в автоматической процедуре сопоставления. Проработаны способы усовершенствования таких правил:

1) самопроверяющиеся процедуры, которые будут непрерывно изменять таблицы весов по мере изменения характера сопоставлений;

2) специальные правила и весовые коэффициенты, разработанные для применения только в тех случаях, когда сомнения являются существенными;

3) методы обнаружения редких потенциально сопоставляемых пар элементов, которые при точном сопоставления давали отрицательные результаты;

4) обобщенные методы сопоставления, которые могут использоваться с таблицами весовых коэффициентов для сопоставления практически любого вида наборов данных;

5) выводы из результатов сопоставлений свойств изделий ЭКБ, которые могут быть присущи определенным типономиналам. Например, одно - и двухполярное напряжение питания.

Оптимизация алгоритма вероятностного сопоставления, требует дальнейшее изучение и совершенствование его функционирования на реальных данных для выведения правил, на основе которых функционирует алгоритм в Системе. Такие исследования потребовали значительные затраты. Повышение эффективности функционирования алгоритма вероятностного сопоставления достигнуто посредством последовательности уточнений, каждое из которых основывается на данных из реальных правил применения изделий и подвергалось детальным исследованиям, прежде чем сформированные правила были включены в методы, в совершенствованный алгоритм и программные средства Системы.

Список источников

1. Дормидошина Д.А., Савин М.Л., Рубцов Ю.В., Применение ICMH в процессах сбора, обработки и анализа информации о надежности изделий микроэлектроники. «Нано- и микросистемная техника», Том 22, №9, г. Москва 2020 - URL: https://www.deyton.ru/doc/ISSN18138586.pdf (дата обращения: 26.01.2026).

2. I.P. Fellegi and A.B. Sunter. A theory for record linkage. Journal of the American Statistical Association, 64(328):11831210, 1969 - URL: https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf (дата обращения: 26.01.2026).

3. Howard B. Newcombe and James M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. Commun. ACM, 5(11):563566, 1962 - URL: https://dl.acm.org/doi/pdf/10.1145/368996.369026 (дата обращения: 26.01.2026).

4. Federico Maggi. A Survey of Probabilistic Record Matching Models, Techniques and Tools. Cycle XXII Scienti c Report TR-2008-22 - URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ce67e95bc9e8f6c848b3ff576d9810929778321e (дата обращения: 26.01.2026).

5. McCallum, Nigam, Ungar, Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, 2000 - URL: https://www.researchgate.net/publication/221653893_Efficient_clustering_of_high-dimensional_data_sets_with_application_to_reference_matching (дата обращения: 26.01.2026).

6. Свидетельство о государственной регистрации программы для ЭВМ №2023614155 Российская Федерация, Программа визуализации данных об изделиях электронной техники: заявка

№2023612746, дата поступления 14.02.2023: дата государственной регистрации в Реестре программ для ЭВМ 27.02.2023/ Рубцов Ю.В., Дормидошина Д.А., Криницкий А.В., Курилов А.В., Окунев К.Е.; заявитель АО «ЦКБ «Дейтон».

7. Рубцов Ю.В., Оптимизация процессов подбора аналогов изделий электронной компонентной базы // Электронная техника. Серия 3: Микроэлектроника. 2015 - URL: https://www.niime.ru /upload/zhurnal-mikroelektronika/%D0%92%D1%8B%D0%BF%D1%83%D1%81%D0%BA%203%20(159) %202015.pdf (дата обращения: 26.01.2026).

8. Свидетельство о государственной регистрации базы данных №2015621293 Российская Федерация, Корпуса ЭКБ: заявка №2015620803, дата поступления 26.06.2015: дата государственной регистрации в Реестре баз данных 26.08.2015/ Грязнова Т.В., Довгань И.Д., Рубцов Ю.В. заявитель АО «ЦКБ «Дейтон».

9. Свидетельство о государственной регистрации программы для ЭВМ №2022683437 Российская Федерация, Программа формирования базы данных об изделиях электронной техники: заявка №2022682511, дата поступления 22.11.2022: дата государственной регистрации в Реестре программ для ЭВМ 05.12.2022/ Рубцов Ю.В., Дормидошина Д.А., Криницкий А.В., Шишкова Ю.М., Окунев К.Е.; заявитель АО «ЦКБ «Дейтон».

10. Свидетельство о государственной регистрации программы для ЭВМ №2022668891 Российская Федерация, Программа управления данными об изделиях электронной техники: заявка №2022667557, дата поступления 27.09.2022: дата государственной регистрации в Реестре программ для ЭВМ 13.10.2022/ Рубцов Ю.В., Дормидошина Д.А., Криницкий А.В., Владимиров А.И., Окунев К.Е.; заявитель АО «ЦКБ «Дейтон».

11. Рубцов Ю.В. Оценка метода машинного обучения для системы автоматизированного выбора компонентной базы радиоэлектронной аппаратуры. Автоматизация и измерения в машиноприборостроении: научный журнал, №3, 2025 - URL: https://www.deyton.ru/doc/stat25.06.2025.pdf (дата обращения: 26.01.2026).

12. Рубцов Ю.В. Алгоритм сопоставления для нормализации данных системы формирования оптимальных предложений на выбор изделий электронной компонентной базы в процессах разработки радиоэлектронной аппаратуры // Автоматизация в промышленности. Алгоритмическое и программное обеспечение. №7 2025 - URL: https://www.deyton.ru/doc/rub09.07.2025.pdf (дата обращения: 26.01.2026).

НАУЧНОЕ ИЗДАНИЕ

АКТУАЛЬНЫЕ НАУЧНЫЕ ИССЛЕДОВАНИЯ

Сборник статей
Международной научно-практической конференции
г. Пенза, 15 февраля 2026 г.
Под общей редакцией
кандидата экономических наук Г.Ю. Гуляева

Подписано в печать 16.02.2026.
Формат 60×84 1/16. Усл. печ. л. 17,8

МЦНС «Наука и Просвещение»
440062, г. Пенза, Проспект Строителей д. 88, оф. 10
www.naukaip.ru