УДК 004.8+004.942
DOI:10.25729/ESI.2026.42.2.008

Метрический анализ сходства композиционных данных для нечеткого поиска релевантных рецептур эластомерных смесей

Рыбанов Александр Александрович, Каблов Виктор Федорович

Волжский политехнический институт (филиал) Волгоградского государственного технического университета, Россия, Волжский, *rybanoff@yandex.ru*

Аннотация. В статье рассматривается актуальная задача разработки специализированных методов для поиска и ранжирования близких по составу рецептур резиновых смесей в базах данных. Целью исследования является разработка и сравнительный анализ метрик сходства, адаптированных для количественной оценки близости многокомпонентных композиционных данных, представленных в виде нормированных векторов весовых долей ингредиентов. Основное содержание работы включает формальную постановку задачи идентификации релевантных рецептур, требующую максимизации комплексной функции сходства, учитывающей как качественный состав (наличие ингредиентов), так и количественные пропорции. В качестве инструментария предложены и адаптированы четыре метрики: взвешенные коэффициенты Жаккара и Дайса, сходство Хеллингера и косинусное сходство. Теоретический анализ их свойств дополнен эмпирической валидацией на реальной промышленной базе данных, содержащей 6096 уникальных рецептур. Научная новизна исследования заключается в систематическом применении и адаптации аппарата метрического анализа к задаче поиска аналогов для композиционных данных материаловедения, а также в выявлении фундаментальной кластеризации рассматриваемых мер сходства. В отличие от существующих подходов, фокусирующихся на бинарном представлении состава или прогнозе свойств, представленная методология целенаправленно решает задачу точного поиска по составу и пропорциям. Полученные результаты выявили практически функциональную эквивалентность взвешенных коэффициентов Жаккара и Дайса (коэффициент корреляции r=0.991), образующих один кластер мер, чувствительных к полному набору компонентов. Сходство Хеллингера и косинусное сходство продемонстрировали сильную корреляцию (r=0.883), сформировав второй кластер мер, ориентированных на оценку структурного подобия пропорций, при этом метрика Хеллингера показала повышенную чувствительность к вариациям долей минорных ингредиентов. На основе этого сформулированы практические рекомендации по комбинированному использованию одной метрики из каждого кластера для создания эффективных поисковых систем. Разработанный метрический аппарат создает формальную основу для интеллектуального поиска аналогов, автоматизации подбора компонентов и сокращения времени на разработку новых рецептур в промышленности.

Ключевые слова: рецептуры резиновых смесей, поиск аналогов, метрики сходства, взвешенный коэффициент Жаккара, коэффициент Дайса, сходство Хеллингера, косинусное сходство, композиционные данные, база данных, материаловедение

Цитирование: Рыбанов А.А. Метрический анализ сходства композиционных данных для нечеткого поиска релевантных рецептур эластомерных смесей / А.А. Рыбанов, В.Ф. Каблов // Информационные и математические технологии в науке и управлении, 2026. – № 2(42). – С. 102-116. – DOI:10.25729/ESI.2026.42.2.008.

Введение. Современное материаловедение характеризуется возрастающей сложностью проектирования многокомпонентных систем, к которым относятся рецептуры резиновых смесей. Задача идентификации близких по составу рецептур в базах данных формулируется как поиск и ранжирование образцов, максимально приближенных по качественному составу и количественным пропорциям ингредиентов к заданной целевой рецептуре [1]. Актуальность решения обусловлена высокой комбинаторной сложностью составов, где даже незначительные вариации весовых долей компонентов могут детерминировать технологические и эксплуатационные характеристики конечного продукта [2]. Специфика объекта исследования требует разработки специализированных методов сравнения, учитывающих композиционную природу данных (сумма долей равна 100%), высокую

размерность признакового пространства и необходимость работы с количественными характеристиками [3].

Проведенный анализ публикаций, посвященных управлению рецептурами резиновых смесей, свидетельствует о существенном прогрессе в области цифровизации материаловедения, однако выявляет значительный пробел в решении задачи поиска близких аналогов. Исследования в области моделирования технологических процессов [4] демонстрируют важность точного контроля состава для управления свойствами материалов, но не предлагают методов для поисковых операций в базах данных. Разработки в области автоматизации расчета рецептур [5] подтверждают общую тенденцию к цифровизации, однако не содержат алгоритмов оценки схожести между рецептами. Применение алгоритмов поиска ассоциативных правил для выявления статистически значимых сочетаний компонентов [6] имеет фундаментальное ограничение – подход оперирует бинарными данными и не учитывает весовые доли, что делает его неприменимым для решения задачи точного поиска аналогов по составу и пропорциям. Эволюция информационных систем в сторону интеллектуальных хранилищ с интеграцией искусственного интеллекта [7] и создание автоматизированного банка данных нового поколения [8] формируют технологическую основу для внедрения сложных алгоритмов поиска, однако конкретные методы и метрики для идентификации близких по составу рецептур в указанных работах не представлены.

Для решения поставленной задачи необходим аппарат, позволяющий количественно оценивать близость между рецептурами, представленными в виде векторов весовых долей в n-мерном пространстве ингредиентов. Анализ литературных данных показывает перспективность использования улучшенной метрики sqrt-cosine similarity, демонстрирующей повышенную точность на разреженных данных [9]. Эффективность комбинирования различных метрик для повышения точности в сложных доменах [10] может быть продуктивно применена к многокритериальной оценке сходства рецептур. Систематизация мер подобия в контексте рекомендательных систем [11] предоставляет теоретическую базу для ранжирования рецептур-кандидатов, а универсальная технология оценки близости информационных объектов [12] предлагает методологический каркас для работы с рецептурами. Исследования по метрической классификации с использованием метода k-ближайших соседей [13] подтверждают применимость данного алгоритма для задач поиска аналогов, однако подчеркивают критическую зависимость его эффективности от адекватности выбранной метрики расстояния.

Таким образом, проведенный анализ позволяет идентифицировать существенный пробел на стыке материаловедения и компьютерных наук. Существующие системы управления рецептурами не предлагают оптимизированных решений для точного поиска близких по составу и пропорциям рецептур, в то время как арсенал прикладной математики содержит разнообразные метрики и алгоритмы, чья эффективность для работы с композиционными, многокомпонентными и разреженными данными рецептур требует тщательной верификации и адаптации. Задачами настоящего исследования являются разработка и сравнительный анализ специализированных методов и метрик для оценки сходства рецептур резиновых смесей, направленных на эффективную идентификацию ближайших аналогов в информационных базах данных с учетом как качественного состава, так и количественных пропорций ингредиентов.

1. Постановка задачи идентификации релевантных рецептур резиновых смесей. Рассмотрим задачу поиска и ранжирования рецептур резиновых смесей в базе данных на основе количественного измерения их комплексного сходства с запросом (целевой

рецептурой), учитывающего как качественный состав (набор ингредиентов), так и количественные характеристики рецептуры (весовые доли компонентов).

Входными данными для задачи идентификации близких рецептур резиновых смесей являются:

1. Целевая рецептура (запрос), представленная как:

$$R_{target} = \{(c_i, w_{target.i}) | i = 1, 2, \dots, m\},$$

где i – индекс ингредиента, последовательно принимающий значения от 1 до m (m – количество ингредиентов в R_{target}); c_i – уникальный идентификатор где i-го ингредиента резиновой смеси R_{target}; $w_{target.i}$ – весовая доля ингредиента c_i в рецептуре R_{target}, $w_{target.i} > 0$, $\sum_{i=1}^{m} w_{target.i} = 1$.

2. База данных рецептур резиновых смесей:

$$DB = \{R_k | k = 1, 2, .., N\},$$

где N – количество рецептур в базе данных DB; R_k – рецептура, представленная аналогично запросу (целевой рецептуре в R_{target}), как $R_k = \{(c_j, w_{k.j}) | j = 1, 2, \dots, m_k\}$.

Для формализации поиска введём упорядоченное множество всех возможных ингредиентов:

$$C = \{c_j | j = 1, 2, \dots, n\},$$

где n – общее количество уникальных ингредиентов во всей системе (DB и R_{target}); $C = \bigcup_{k=1}^{N} \text{comp}(R_k) \cup \text{comp}(R_{target})$, $comp(R)$ – оператор извлечения идентификаторов рецептуры R;

Каждая рецептура $R_k \in DB \cup R_{target}$ преобразуется в векторы фиксированной размерности, соответствующей универсальному множеству всех уникальных ингредиентов, встречающихся в рецептурах резиновых смесей.

$$V(R_k) = \{v_j | j = 1, 2, \dots, n\} \in \mathbb{R}^n,$$

где координата v_j соответствует ингредиенту $c_j \in C$:

$$v_j(R_k) = \begin{cases} w_{kj}, \text{если } c_j \in R_k \text{ как } (c_j, w_{k.j}) \\ 0, \text{если } c_j \notin R_k \end{cases}.$$

Координаты вектора соответствуют весовым долям конкретных компонентов; отсутствие компонента в рецептуре кодируется нулевым значением.

Ключевым элементом постановки задачи является определение комплексной функции сходства $Sim\left(V(R_{target}), V(R_k)\right)$, количественно оценивающей близость вектора целевой рецептуры $V(R_{target})$ к вектору рецептуры из базы данных $V(R_k)$. Эта функция должна удовлетворять двум фундаментальным требованиям: во-первых, учитывать факт наличия или отсутствия каждого ингредиента в сравниваемых рецептурах; во-вторых, интегрировать различия в весовых долях присутствующих ингредиентов. Функция $Sim\left(V(R_{target}), V(R_k)\right)$, должна удовлетворять следующим условиям:

- $Sim\left(V(R_{target}), V(R_k)\right) \in [0,1]$
- $Sim\left(V(R_{target}), V(R_{target})\right) = 1$ (максимальное сходство вектора целевой рецептуры с самим собой).
- чем больше значение $Sim\left(V(R_{target}), V(R_k)\right)$, тем ближе рецептура R_k к R_{target}.

Постановку задачи идентификации релевантных рецептур резиновых смесей сформулируем следующим образом: разработать алгоритм идентификации и ранжирования рецептур резиновых смесей из базы данных DB, максимизирующих меру сходства с целевой рецептурой R_{target} в соответствии с заданной функцией сходства Sim.

Пусть заданы:

– База данных $DB = \{R_k | k = 1, 2, .., N\}$ из N рецептур;
– Целевая рецептура R_{target};
– Функция сходства $Sim\left(V(R_{target}), V(R_k)\right)$.

Требуется найти ранжированное подмножество $S^* \subseteq DB$ мощностью $K \leq N$:

$$S^* = \{R_{s_1}, R_{s_1}, \dots, R_{s_K}\},$$

удовлетворяющее двум условиям:

1) условие оптимальности (доминирования по сходству):

$$\forall R_{s_i} \in S^*, \forall R_j \in DB \backslash S^*: Sim\left(V(R_{target}), V(R_{s_i})\right) > Sim\left(V(R_{target}), V(R_j)\right);$$

2) последовательность $\{R_{s_1}, R_{s_2}, \dots, R_{s_K}\}$ упорядочена по убыванию меры сходства (условие монотонного убывания):

$$Sim\left(V(R_{target}), V(R_{s_1})\right) \geq Sim\left(V(R_{target}), V(R_{s_2})\right) \geq \cdots \geq Sim\left(V(R_{target}), V(R_{s_K})\right),$$

Предлагаемая формализация задачи представляет существенную практическую ценность для материаловедения, непосредственно способствуя решению ключевых прикладных задач при проектировании резиновых смесей: идентификации технологически эквивалентных составов; ранжированной визуализации альтернативных рецептур; автоматизации подбора компонентной базы; минимизации временных ресурсов, затрачиваемых на разработку новых композиций.

2. Метрики сходства рецептур резиновых смесей. В контексте задачи идентификации релевантных рецептур резиновых смесей предлагается к рассмотрению класс метрик сходства, основанных на операциях с непрерывными весовыми долями компонентов. Эти метрики предполагают интерпретацию рецептур, как векторов в многомерном пространстве ингредиентов. Они обеспечивают градуированную количественную оценку близости составов резиновых смесей, учитывающую как наличие совпадающих компонентов, так и вариации их весовых соотношений. Каждая метрика характеризуется уникальным набором формально выраженных свойств, детерминирующих ее чувствительность к специфическим типам расхождений: различиям в пропорциях общих ингредиентов, наличию компонентов, уникальных для одного из составов, либо отклонениям в концентрациях минорных составляющих.

В рамках настоящего исследования рассматриваются четыре ключевые метрики, адаптированные для анализа нормированных весовых долей ингредиентов:

Взвешенный коэффициент Жаккара. Для целевой рецептуры R_{target} и рецептуры R_k из базы данных, функция сходства на основе взвешенного коэффициента Жаккара определяется как:

$$Sim_J\left(V(R_{target}), V(R_k)\right) = \frac{\sum_{j=1}^{n} min\left(v_j(R_{target}), v_j(R_k)\right)}{\sum_{j=1}^{n} max\left(v_j(R_{target}), v_j(R_k)\right)} = \frac{\sum_{j=1}^{n} min\left(w_{target.j}, w_{k.j}\right)}{\sum_{j=1}^{n} max\left(w_{target.j}, w_{k.j}\right)},$$

где n – мощность универсального множества ингредиентов C; $w_{target.j}$ – весовая доля -го ингредиента в R_{target} (0 при отсутствии); $w_{k.j}$ – весовая доля j-го ингредиента в R_k (0 при отсутствии);

В отличие от бинарного аналога, оперирующего дискретными индикаторами наличия ингредиентов, взвешенный коэффициент Жаккара использует непрерывные весовые доли компонентов, что обеспечивает количественную оценку сходства рецептур. Ключевая семантика метрики определяется операциями над множествами: числитель $\sum_{j=1}^{n} min\left(w_{target.j}, w_{k.j}\right)$ интерпретируется, как пересечение рецептур, отражающее

суммарный вклад ингредиентов, присутствующих в обоих составах, взвешенный по минимальным долям; знаменатель $\sum_{j=1}^{n} max\left(w_{target.j}, w_{k.j}\right)$ характеризует их объединение, учитывающее все уникальные компоненты через максимальные доли. Метрика демонстрирует высокую чувствительность к различиям: расхождение в долях общих ингредиентов ($\min(\mathrm{w_j}, \mathrm{w_j'}) \ll \max(\mathrm{w_j}, \mathrm{w_j'})$) снижает сходство, а отсутствие компонента в одной из рецептур ($\min(\mathrm{w_j}, 0) = 0, \max(\mathrm{w_j}, 0) = \mathrm{w_j} > 0$) уменьшает итоговое значение. Коэффициент нормирован на интервал [0,1], где единица достигается исключительно при полном совпадении весовых долей всех компонентов ($\forall \mathrm{j}: \mathrm{w_{target.j}} = \mathrm{w_{k.j}}$), а нулевое значение соответствует отсутствию общих ингредиентов.

Взвешенный коэффициент Дайса (Weighted Dice Coefficient). Для целевой рецептуры R_{target} и рецептуры R_k из базы данных функция сходства на основе взвешенного коэффициента Дайса определяется как:

$$Sim_{Dice}\left(V\left(R_{target}\right), V(R_k)\right) = \frac{2 \cdot \sum_{j=1}^{n} min\left(v_j\left(R_{target}\right), v_j(R_k)\right)}{\sum_{j=1}^{n} v_j\left(R_{target}\right) + \sum_{j=1}^{n} v_j(R_k)} =$$

$$= \frac{2 \cdot \sum_{j=1}^{n} min\left(w_{target.j}, w_{k.j}\right)}{\sum_{j=1}^{n} w_{target.j} + \sum_{j=1}^{n} w_{k.j}}.$$

Данная метрика фокусируется на совпадении компонентного состава рецептур, где числитель $2 \cdot \sum_{j=1}^{n} min\left(w_{target.j}, w_{k.j}\right)$ отражает удвоенную сумму перекрытия весовых долей общих ингредиентов, а знаменатель $\sum_{j=1}^{n} w_{target.j} + \sum_{j=1}^{n} w_{k.j}$ (равный 2 для нормированных рецептур) представляет суммарную массу обоих составов. В отличие от коэффициента Жаккара, явно учитывающего уникальные компоненты через операцию максимума, метрика Дайса концентрируется на совпадающих элементах, игнорируя несовпадения. Уникальные ингредиенты ($w_{target.j} > 0, w_{k.j} = 0$ или обратно) уменьшают сходство, внося нулевой вклад в числитель, но полный вклад в знаменатель, однако их влияние менее выражено, чем у Жаккара, где знаменатель дополнительно увеличивается через $max\left(w_j, 0\right)$. Коэффициент строго нормирован на интервал [0,1], достигая единицы исключительно при полном совпадении всех весовых долей всех компонентов ($\forall j: w_{target.j} = w_{k.j}$) и нуля – при отсутствии общих ингредиентов ($\sum_{j=1}^{n} min\left(w_j, w_j'\right) = 0$). Для любых рецептур выполняется $Sim_{Dice} \geq Sim_J$ с равенством только при идентичности составов или отсутствии уникальных ингредиентов.

Сходство Хеллингера (Hellinger Similarity). Сходство Хеллингера измеряет близость между целевой рецептурой R_{target} и рецептурой R_k из базы данных, интерпретируя их, как дискретные вероятностные распределения:

$$Sim_H\left(V\left(R_{target}\right), V(R_k)\right) = 1 - \sqrt{\frac{1}{2}\sum_{j=1}^{n}\left(\sqrt{v_j\left(R_{target}\right)} - \sqrt{v_j(R_k)}\right)^2} =$$

$$= 1 - \sqrt{\frac{1}{2}\sum_{j=1}^{n}\left(\sqrt{w_{target.j}} - \sqrt{w_{k.j}}\right)^2}.$$

Значения метрики всегда лежат в диапазоне [0,1]. Ключевой особенностью является повышенная чувствительность к различиям в пропорциях минорных ингредиентов: операция взятия квадратного корня $\sqrt{w_j}$ усиливает относительные отклонения малых долей (например,

разница между 0.01 и 0.02 существеннее, чем между 0.1 и 0.11). При этом метрика корректно обрабатывает нулевые компоненты, что позволяет сравнивать рецептуры с несовпадающим составом. Благодаря этим свойствам сходство Хеллингера эффективно для задач, требующих точного учёта пропорций всех ингредиентов, включая второстепенные.

Косинусное сходство (Cosine Similarity). Косинусное сходство измеряет близость между целевой рецептурой R_{target} и рецептурой R_k из базы данных, представляя их, как векторы весовых долей в n-мерном пространстве компонентов. В отличие от метрик, требующих нормировки на сумму (например, Хеллингера), косинусное сходство фокусируется на угловой близости векторов, вычисляемой по формуле:

$$Sim_{cos}\left(V(R_{target}), V(R_k)\right) = \frac{\sum_{j=1}^{n} v_j(R_{target}) \cdot v_j(R_k)}{\sqrt{\sum_{j=1}^{n}\left(v_j(R_{target})\right)^2} \cdot \sqrt{\sum_{j=1}^{n}\left(v_j(R_k)\right)^2}} =$$

$$= \frac{\sum_{j=1}^{n} w_{target.j} \cdot w_{k.j}}{\sqrt{\sum_{j=1}^{n}\left(w_{target.j}\right)^2} \cdot \sqrt{\sum_{j=1}^{n}\left(w_{k.j}\right)^2}}.$$

В отличие от метрик, чувствительных к нормировке суммы долей (например, Хеллингера), косинусное сходство фокусируется на схожести относительных пропорций компонентов, инвариантно к абсолютным масштабам векторов (общей массе рецептуры). Метрика принимает значения в диапазоне $[0,1]$ и оптимальна для поиска рецептур со схожей структурой компонентов, когда критичны именно относительные соотношения, а не абсолютные количества или полный набор компонентов. Чувствительность к уникальным компонентам (нулевым значениям в одном из векторов) зависит от размерности пространства.

Выбор оптимальной метрики сходства осуществляется на основании специфики решаемой аналитической задачи, заданных требований к чувствительности к различным аспектам различий составов, а также соображений вычислительной эффективности.

3. Результаты исследований. В ходе исследования использовалась база данных промышленных рецептур резиновых смесей, содержащая информацию о компонентном составе, ингредиентах и характеристиках резиновых смесей с учетом технических синонимов [8, 14]. Общий объем анализируемого корпуса данных составил 6096 уникальных рецептур (рис. 1). На основе указанного массива данных была реализована процедура нечеткого поискового запроса, направленная на идентификацию релевантных рецептур резиновых смесей.

Анализ иерархического распределения видов каучука в рецептурах резиновых смесей БД, представленный на рис. 2, показывает, что хлоропреновый каучук лидирует с 1162 рецептурами, за ним следует натуральный каучук с 935 рецептурами, что свидетельствует об их универсальности и широком применении в различных отраслях. В то же время, такие виды, как бутиловый и стирол-бутадиеновый каучук, имеют значительно меньше рецептур, что может указывать на узкую область применения.

Фрагмент логической схемы базы данных, содержащей данные о составе рецептов резиновых смесей, на основе которых была сформирована выборка исходных данных для процедуры нечеткого поиска, представлен на рис. 2.

Представленная схема отображает ключевые сущности («Рецептура», «Ингредиент») и связи между ними типа «входит в состав». Данная модель обеспечивает однозначную идентификацию компонентного состава каждой рецептуры и ассоциированных с ней физико-механических свойств, что формирует основу для последующего метрического

анализа. Особенностью базы данных является наличие справочников технических синонимов ингредиентов и свойств резиновых смесей, которые, упрощают поиск информации и позволяют применять методы машинного обучения для анализа данных

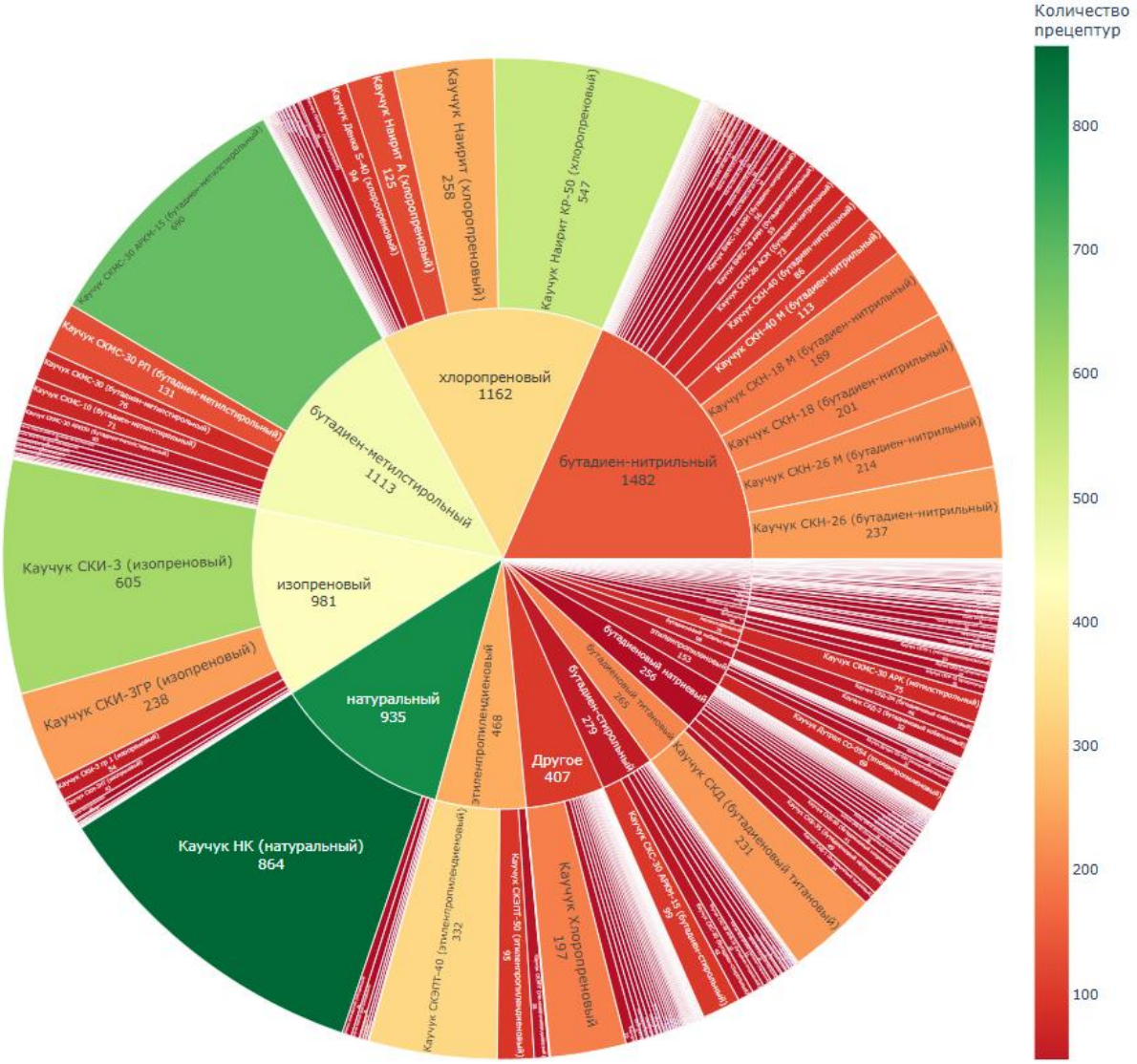

Рис. 1. Иерархическое распределение видов каучука в базе данных

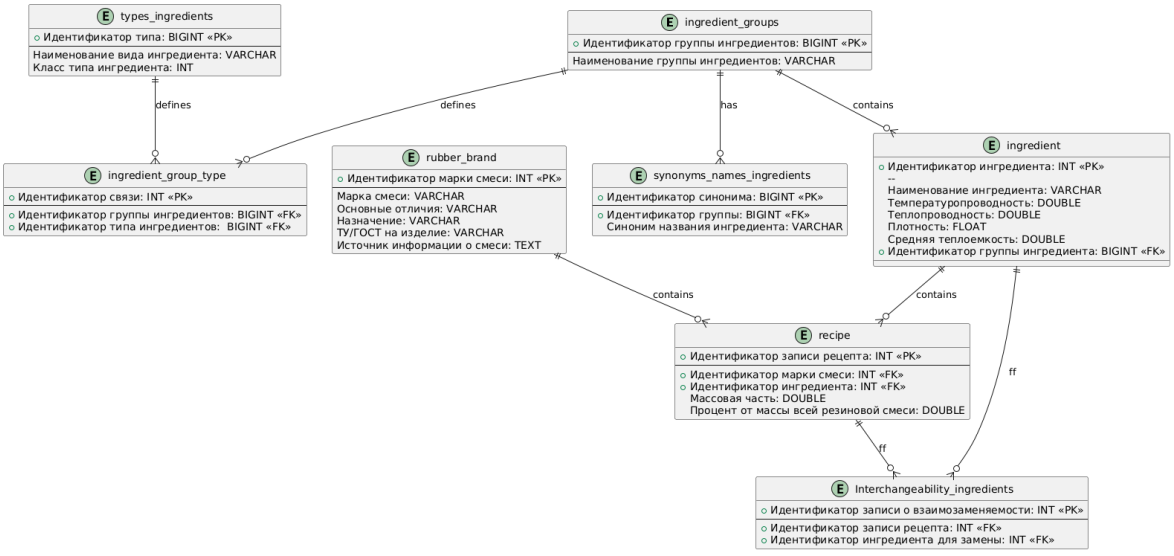

Рис. 2. Фрагмент логической схемы БД с исходными данными по составу рецептов резиновых смесей

На рис. 3 представлен компонентный состав резиновой смеси марки 1831-2, для которой с помощью процедуры нечёткого поиска были идентифицированы релевантные рецептуры-аналоги.

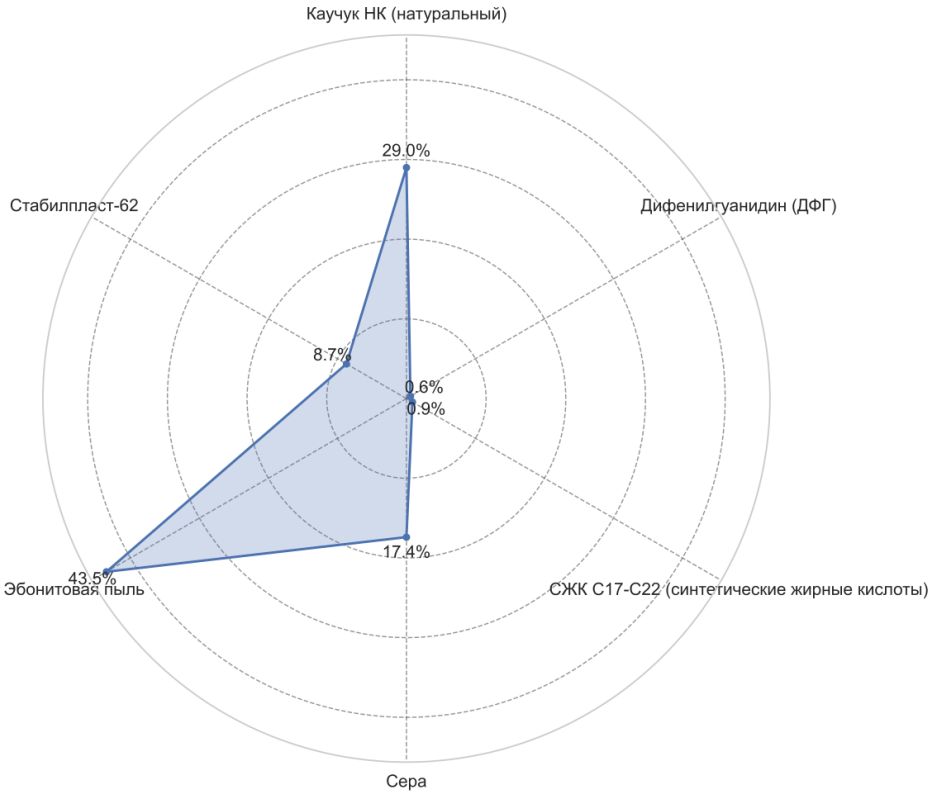

Рис. 3. Состав резиновой смеси марки 1831-2

Для каждой из четырёх метрик композиционного подобия (взвешенный коэффициент Жаккара, взвешенный коэффициент Дайса, сходство Хеллингера, косинусное сходство) выполнено ранжирование рецептур и селекция 10 максимально близких к эталону 1831-2. В результате объединения полученных множеств сформирована результирующая выборка из 15 уникальных промышленных марок (табл. 2), что математически подтверждает значимое пересечение отобранных рецептур для различных метрик (коэффициент перекрытия $k=0.625$).

На основе результатов корреляционного анализа данных, представленных в таблице 2, выявлены следующие закономерности. Обнаружена практически функциональная эквивалентность метрик Жаккара (J) и Дайса (D), о чем свидетельствует их предельно сильная корреляция ($r=0.991$, $p<0.0001$), что предполагает их взаимозаменяемость при анализе сходства рецептур. Аналогично, установлена сильная корреляция между метрикой Хеллингера (H) и косинусным сходством (C) ($r=0.883$, $p<0.0001$), подтверждающая их общую природу, основанную на L2-нормировании и чувствительности к пропорциональному распределению компонентов. Все междуметрические корреляции между разными парами (*J-H: r=0.932; J-C: r=0.923; D-H: r=0.899; D-C: r=0.925*) оказались статистически высоко значимыми ($p<10^{-5}$), демонстрируя фундаментальную согласованность всех исследуемых метрик в оценке композиционного сходства. Иерархия силы корреляционных связей (J-D > J-H ≈ J-C > D-C > D-H > H-C) указывает на максимальную близость J и D и относительно меньшую сопряженность H и C. Данные подтверждают гипотезу о двух кластерах: J-D (чувствительные к уникальным компонентам) и H-C (чувствительные к пропорциям). Практически это позволяет оптимизировать вычисления, используя по одной репрезентативной метрике из каждого кластера (например, J и C), и интерпретировать

расхождения между кластерами как указание на различия в уникальности компонентов или пропорциях минорных ингредиентов. Высокая согласованность всех метрик подтверждает надежность оценки сходства рецептур и обосновывает их комбинированное применение для многомерного анализа.

Таблица 2. Результаты поиска рецептур резиновых смесей, композиционно близких к эталонной рецептуре марки 1831-2, на основе применения метрик сходства

№	ID рецептуры	Марка рецептуры	Взвешенный коэффициент Жаккара	Взвешенный коэффициент Дайса	Сходство Хеллингера	Косинусное сходство
1	799	1831-1	0.8400	0.9130	0.7051	0.9757
2	761	1791	0.8254	0.9043	0.7170	0.9821
3	3553	Л-1788	0.6102	0.7579	0.5942	0.9177
4	522	10998-1	0.5066	0.6725	0.4531	0.7366
5	521	10998	0.4435	0.6145	0.3916	0.7209
6	2130	6348-3	0.4356	0.6069	0.4381	0.7840
7	2129	6348-19	0.4356	0.6069	0.4381	0.7840
8	4184	Э-230	0.3767	0.5472	0.3499	0.6988
9	2072	6279-4	0.3678	0.5378	0.3518	0.5578
10	2071	6279-3	0.3678	0.5378	0.3518	0.5578
11	2174	640-13	0.3637	0.5334	0.4367	0.5567
12	2128	6345-8	0.3184	0.4831	0.3251	0.6055
13	2127	6345-7	0.3184	0.4831	0.3251	0.6055
14	963	24Р	0.3019	0.4638	0.4336	0.6014
15	4726	4508-3	0.1949	0.3262	0.3619	0.5312

С позиции прикладного материаловедения, полученные результаты (рис. 4) позволяют сформулировать конкретные методические рекомендации по выбору рецептур-аналогов.

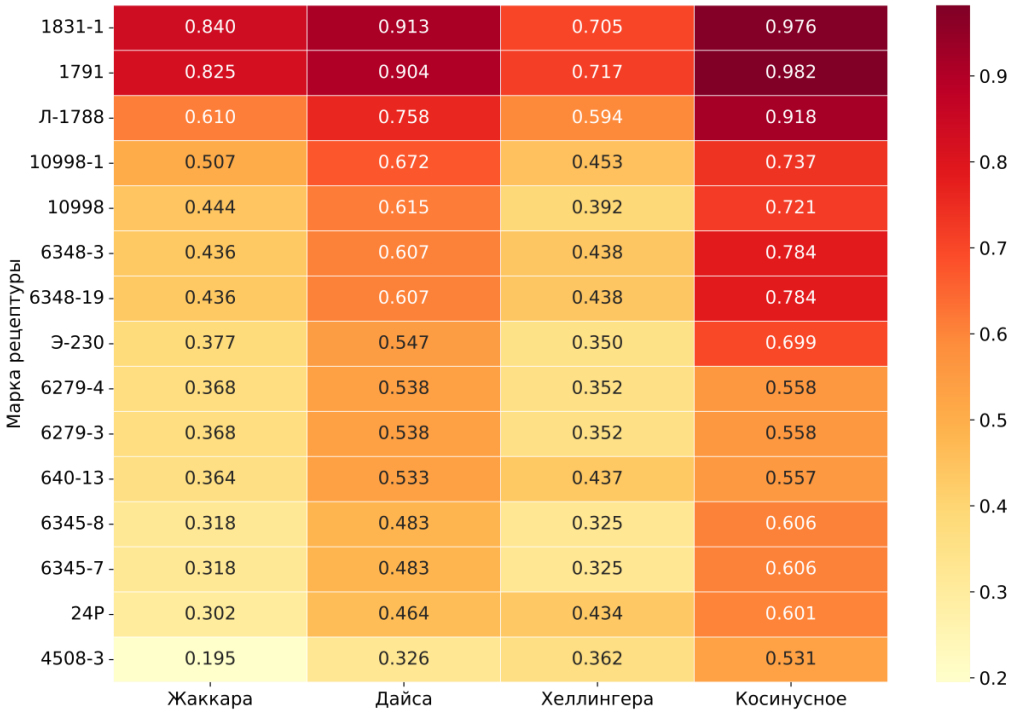

Рис. 4. Матрица сходства рецептур по различным метрикам

Визуальный анализ матрицы позволяет выделить два ярко выраженных кластера рецептур: первый включает марки 1831-1, 1791 и Л-1788, демонстрирующие высокие значения сходства по всем метрикам, что указывает на их максимальную близость к эталону, как по компонентному составу, так и по пропорциям. Второй кластер образован рецептурами с умеренными и низкими показателями сходства (например, 10998-1, 6348-3, 6279-4), при этом внутри этого кластера наблюдается тесная группировка рецептур-пар (6348-3/6348-19, 6279-3/6279-4, 6345-7/6345-8), что свидетельствует о их внутренней схожести и, вероятно, общих технологических модификациях. Данная кластеризация наглядно демонстрирует практическую применимость метрик для категоризации рецептур по степени их взаимного сходства.

Важным аспектом методологии является интеграция экспертного знания в области рецептуростроения на всех стадиях работы – от формализации задачи до интерпретации выходных данных алгоритмов. В рамках проведенного исследования основной акцент сделан на объективной количественной валидации предлагаемых математических методов по их способности идентифицировать близкие аналоги в реальных промышленных данных. При этом детальная технологическая и производственная апробация каждого выявленного аналога, требующая натурного эксперимента, представляет собой отдельный, последующий этап внедрения, выходящий за рамки настоящей научной публикации, посвященной разработке и сравнению алгоритмических подходов.

4. Обсуждение результатов. Настоящее исследование было целенаправленно сфокусировано на фундаментальной задаче метрического сравнения композиционного состава рецептур, как на критически важном первом шаге информационного поиска в материаловедении. Разработанный и верифицированный метрический аппарат (раздел 2) создает формальную основу для решения этой задачи. Наша рабочая гипотеза, основанная на экспертизе в области проектирования эластомерных материалов, заключалась в том, что сходство состава является необходимым (хотя и не всегда достаточным) базисом для потенциального сходства свойств. Таким образом, разработанные методы идентификации близких по составу аналогов формируют естественную основу для последующей интеграции с прогнозными моделями «состав-свойство», что является логичным предметом следующих исследовательских задач.

Проведенный корреляционный анализ метрик сходства рецептур резиновых смесей выявил две принципиально различные группы мер близости с различными свойствами и областями применения. Исследование взаимосвязи между взвешенными коэффициентами Жаккара (J) и Дайса (D) (рис. 5) показало наличие статистически значимой линейной зависимости ($r = 0{,}9911$; $p < 0{,}0001$) с точным соответствием экспериментальных данных теоретической кривой, описываемой соотношением $D = (2J)/(1 + J)$. Данный результат свидетельствует о функциональной эквивалентности рассматриваемых метрик, что подтверждается наблюдаемым во всем диапазоне значений неравенством $D \geq J$, являющимся следствием математической природы данных мер.

В противоположность этому, корреляционный анализ между сходством Хеллингера (H) и косинусным сходством (C) (рис. 5) показал статистически значимую, но менее выраженную связь ($r = 0{,}8826$; $p < 0{,}0001$) с существенным разбросом экспериментальных точек. Данное обстоятельство указывает на комплементарный характер этих метрик, отражающих различные аспекты оценки сходства рецептур. Метрика Хеллингера демонстрирует повышенную чувствительность к вариациям концентраций минорных компонентов благодаря операции взятия квадратного корня, в то же время как косинусное

сходство более эффективно характеризует общую структурную схожесть пропорциональных соотношений ингредиентов.

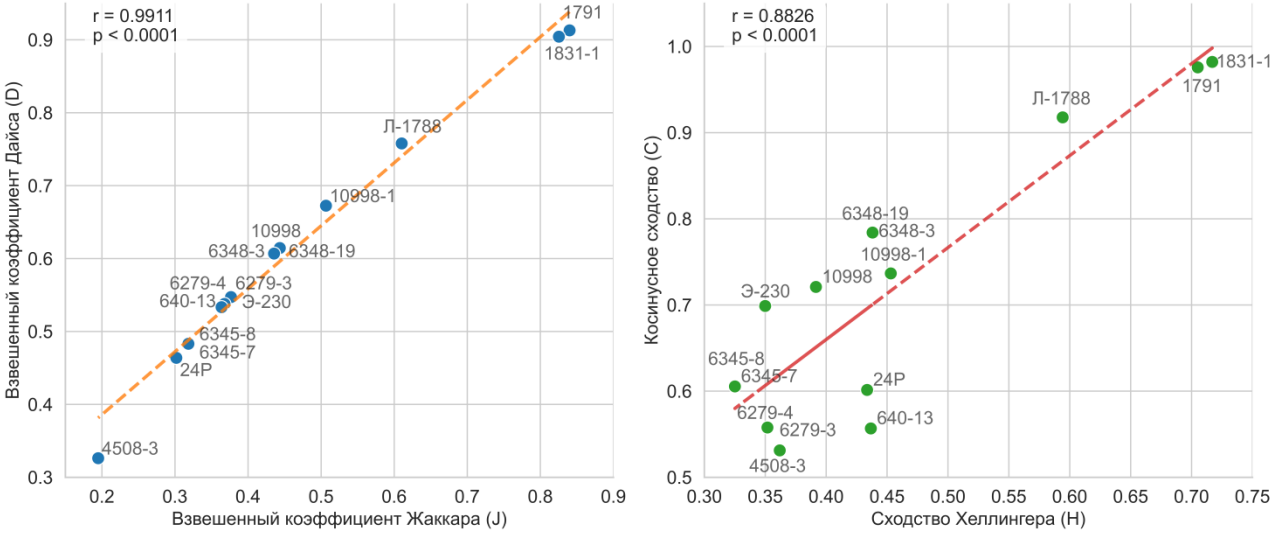

Рис. 5. Корреляционная зависимость между взвешенными коэффициентами

Выбранные четыре метрики репрезентируют два фундаментальных и взаимодополняющих математических подхода к оценке близости композиционных данных: подход, основанный на операциях над множествами (взвешенные коэффициенты Жаккара и Дайса), и подход, основанный на измерении расстояний в векторном пространстве (сходство Хеллингера и косинусное сходство). Продемонстрированная функциональная эквивалентность внутри выявленных кластеров (*J–D* и *H–C*) является ключевым результатом, подтверждающим адекватность и достаточность данного компактного набора мер для решения задачи многокритериальной классификации и ранжирования рецептур по степени сходства. Таким образом, использованная в исследовании база данных промышленных рецептур, репрезентующая значительный сегмент известных композиций, выступила адекватным полигоном для валидации предложенной методологии. Ключевым итогом является не получение результатов, специфичных для данной выборки, а выявление универсального принципа: целесообразность комбинирования метрик, репрезентирующих два выявленных кластера (основанных на операциях над множествами и на измерении расстояний). Данный принцип инвариантен к предметной области и применим для анализа других баз композиционных данных.

С практической точки зрения, установленная взаимозаменяемость метрик Жаккара и Дайса позволяет оптимизировать вычислительные алгоритмы путем использования только одной из них без ущерба для точности анализа. В то же время, комбинированное применение метрик Хеллингера и косинусного сходства обеспечивает многомерную оценку сходства рецептур, где расхождения в их показаниях содержат дополнительную информацию о характере различий между составами.

Таким образом, рациональная организация процедуры поиска аналогов в базах данных резиновых смесей может быть достигнута путем совместного использования одной метрики из кластера функционально эквивалентных мер (J или D) и одной метрики из кластера комплементарных мер (H или C). Данный подход обеспечивает оптимальный баланс между вычислительной эффективностью и полнотой анализа как качественного состава, так и количественных пропорций ингредиентов.

Предложенные и верифицированные метрики сходства служат базовыми компонентами для построения сложных поисковых и рекомендательных систем в

материаловедении. Детальное рассмотрение архитектур на основе искусственного интеллекта, таких, как системы, использующие нейросетевые эмбеддинги или глубокое обучение для прогнозирования свойств на основе состава, представляет собой отдельную масштабную исследовательскую задачу. Развитие таких систем в будущем будет опираться на полученный в данной работе фундамент – формализованный и проверенный на реальных данных метрический аппарат для оценки композиционного сходства.

Заключение. Проведенное исследование демонстрирует эффективность формализованного подхода к задаче идентификации близких по составу рецептур резиновых смесей. Разработанная постановка задачи поиска и ранжирования, основанная на максимизации комплексной функции сходства, представляет существенную практическую ценность для автоматизации проектирования новых композиций и поиска технологически эквивалентных замен.

Сравнительный анализ четырех метрик сходства позволил выявить их фундаментальные свойства и области рационального применения. Установлена функциональная эквивалентность взвешенных коэффициентов Жаккара и Дайса, что подтверждается предельно сильной корреляцией (r=0,991) и их взаимозаменяемостью для оценки полного компонентного состава. С другой стороны, сходство Хеллингера и косинусное сходство, демонстрируя сильную, но менее выраженную связь (r=0,883), образуют кластер комплементарных мер, ориентированных на анализ пропорциональных соотношений ингредиентов, где сходство Хеллингера проявляет повышенную чувствительность к минорным компонентам.

Полученные результаты позволяют сформулировать конкретные методические рекомендации. Для решения прикладных задач в материаловедении наиболее эффективной является стратегия комбинированного использования одной метрики из кластера «Жаккар-Дайс» (для контроля за полным набором ингредиентов) и одной из кластера «Хеллингер-Косинус» (для анализа структурного подобия пропорций). Такой подход обеспечивает оптимальный баланс между вычислительной эффективностью и полнотой анализа, позволяя интерпретировать расхождения между показаниями метрик из разных кластеров как указание на специфический характер различий между рецептурами.

Таким образом, работа вносит вклад как в развитие методов компьютерного анализа данных в материаловедении, так и в решение практических задач промышленного проектирования резиновых смесей, предлагая надежный и интерпретируемый инструмент для интеллектуального поиска в специализированных базах данных.

Список источников

1. Каблов В.Ф. Эволюция информационных систем управления рецептурами резиновых смесей: от баз данных к интеллектуальным хранилищам с интеграцией модулей искусственного интеллекта / В.Ф. Каблов, А.А. Рыбанов, М.А. Маслова // Каучук и резина-2025: Традиции и новации: материалы XIII Всероссийской конференции. – Москва, 2025. – С. 23-24.
2. Нигматуллин В.Р. Использование методов машинного обучения и искусственного интеллекта в химической технологии. Часть I / В.Р. Нигматуллин, Н.А. Руднев // Нефтегазовое дело, 2019. – № 4. – С. 243-268. – DOI: 10.17122/ogbus-2019-4-243-268
3. Rajan K., et al. Data mining and multivariate analysis in materials science. Molten salts: from fundamentals to applications, Dordrecht, Springer, 2002, p. 89-102.
4. Тихомиров С.Г. Управление процессом вулканизации на основе моделирования и оценки ключевых параметров модели / С.Г. Тихомиров, А.А. Маслов, О.В. Карманова и др. // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2024. – № 4. – С. 22-34.
5. Матвеев Н.А. Автоматизация расчёта рецептуры резиновой смеси при помощи программного обеспечения / Н.А. Матвеев, А.П. Моргунов // Омский научный вестник, 2016. – № 5 (149). – С. 55-58.

6. Рыбанов А.А. Применение алгоритмов поиска ассоциативных правил для определения технологически значимых сочетаний компонентов в рецептурах резиновых смесей / А.А. Рыбанов, В.Ф. Каблов // Информационные и математические технологии в науке и управлении, 2025. – № 3 (39). – С. 177-188.
7. Рыбанов А.А. Интеграция синонимических систем в информационные системы управления рецептами резинотехнических материалов / А.А. Рыбанов, В.Ф. Каблов, М.А. Маслова // Вестник Череповецкого государственного университета, 2025. – № 5 (128). – С. 70-82.
8. Каблов В.Ф. Автоматизированный банк данных нового поколения рецептур и свойств резин / В.Ф. Каблов, А.А. Рыбанов, Н.А. Кейбал // Каучук и резина, 2024. – Т. 83. – № 3. – С. 168-173.
9. Sohangir S., Wang D. Improved sqrt-cosine similarity measurement. Journal of Big Data, 2017, vol. 4, no. 25, pp. 1-13, DOI:10.1186/s40537-017-0083-6.
10. Chen F., Farahat A., Brants T. Multiple similarity measures and source-pair information in story link detection. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004). Boston, MA, USA, 2004, pp. 313-320.
11. Порошкина В.В. Меры подобия в рекомендательных системах / В.В. Порошкина // Аллея науки, 2019. – Т. 1, № 4 (31). – С. 909-913.
12. Кузнецов Л.А. Универсальная технология оценки близости информационных объектов / Л.А. Кузнецов // Информатика и её применения,2014. – Т. 8. – № 2. – С. 130-144.
13. Гороховатская Н.В. Метрическая классификация с использованием метода ближайших соседей на основе описаний в виде множеств признаков: дис. ... канд. техн. наук. – М., 2012. – 145 с.
14. База данных рецептур, ингредиентов и свойств резиновых смесей с учетом технических синонимов: свидетельство о гос. регистрации базы данных № 2025621312 Российская Федерация: № 2025620796: заявл. 12.03.2025: опубл. 24.03.2025 / В.Ф. Каблов, А.А. Рыбанов, Н.А. Кейбал и др.; правообладатель ВПИ (филиал) ВолгГТУ.

Рыбанов Александр Александрович. *Кандидат технических наук, доцент, кафедра «Информатика и технология программирования», Волжский политехнический институт (филиал) Волгоградский государственный технический университет, заведующий кафедрой. AuthorID: 234968, SPIN: 9256-6050, ORCID: 0000-0002-8638-9998, rybanoff@yandex.ru. 404111, Волгоградская область, г. Волжский, пр-кт Ленина, 72.*

Каблов Виктор Федорович. *Доктор технических наук, профессор, кафедра «Химическая технология полимеров и промышленная экология», Волжский политехнический институт (филиал) Волгоградский государственный технический университет, профессор. AuthorID: 115782, SPIN: 1093-9003, ORCID: 0000-0002-2970-6109, vkablov5@gmail.com. 404111, Волгоградская область, г. Волжский, пр-кт Ленина, 72.*

UDC 004.8+004.942
DOI:10.25729/ESI.2026.42.2.008

Metric similarity analysis of compositional data for fuzzy search of relevant elastomeric mixture formulations

Alexander A. Rybanov, Victor F. Kablov

Volzhsky Polytechnic Institute (Branch) of Volgograd State Technical University,
Russia, Volzhsky, *rybanoff@yandex.ru*

Abstract. The article addresses the pressing task of developing specialized methods for searching and ranking rubber compound formulations with similar compositions in databases. The aim of the research is to develop and conduct a comparative analysis of similarity metrics adapted for quantitatively assessing the proximity of multi-component compositional data represented as normalized vectors of ingredient weight fractions. The core of the work includes a formal statement of the problem of identifying relevant formulations, which requires maximizing a comprehensive similarity function that considers both qualitative composition (presence of ingredients) and quantitative proportions. Four metrics are proposed and adapted as tools: weighted Jaccard and Dice coefficients, Hellinger similarity, and cosine similarity. A theoretical analysis of their properties is

supplemented by empirical validation on a real industrial database containing 6,096 unique formulations. The scientific novelty of the study lies in the systematic application and adaptation of metric analysis apparatus to the task of searching for analogues for compositional materials science data, as well as in revealing a fundamental clustering of the considered similarity measures. Unlike existing approaches focused on binary representation of composition or property prediction, the presented methodology purposefully solves the problem of precise search by composition and proportions. The obtained results revealed a near-functional equivalence of the weighted Jaccard and Dice coefficients (correlation coefficient r=0.991), forming one cluster of measures sensitive to the full set of components. Hellinger similarity and cosine similarity demonstrated a strong correlation (r=0.883), forming a second cluster of measures focused on assessing structural similarity of proportions, with the Hellinger metric showing increased sensitivity to variations in the fractions of minor ingredients. Based on this, practical recommendations are formulated for the combined use of one metric from each cluster to create effective search systems. The developed metric framework establishes a formal basis for intelligent analogue search, automation of component selection, and reduction of development time for new formulations in the industry.

Keywords: rubber compound formulations, search for analogues, similarity metrics, weighted Jaccard coefficient, Dice coefficient, Hellinger similarity, cosine similarity, compositional data, database, materials science

References

1. Kablov V.F., Rybanov A.A., Maslova M.A. Evolyutsiya informatsionnykh sistem upravleniya retsepturami rezinnovykh smesey: ot baz dannykh k intellektual'nym khranilishcham s integratsiyey moduley iskusstvennogo intellekta [Evolution of rubber compound recipe management information systems: from databases to intelligent repositories with integration of artificial intelligence modules]. Kauchuk i rezina-2025: Traditsii i novatsii: materialy XIII Vserossiyskoy konferentsii [Rubber and rubber-2025: Traditions and innovations: proceedings of the xiii All-russian conference]. Moscow, 2025, pp. 23-24.
2. Nigmatullin V.R., Rudnev N.A. Ispol'zovaniye metodov mashinnogo obucheniya i iskusstvennogo intellekta v khimicheskoy tekhnologii. Chast' I [Application of machine learning and artificial intelligence methods in chemical technology. Part I]. Neftegazovoye delo [Oil and gas business], 2019, no. 4, pp. 243-268, DOI: 10.17122/ogbus-2019-4-243-268.
3. Rajan K. Data mining and multivariate analysis in materials science. In: M. Gaune-Escard (ed.). Molten Salts: From Fundamentals to Applications. Dordrecht, Springer, 2002, pp. 89-102.
4. Tikhomirov S.G., Maslov A.A., Karmanova O.V. et al. Upravleniye protsessom vulkanizatsii na osnove modelirovaniya i otsenki klyuchevykh parametrov modeli [Control of the vulcanization process based on modeling and evaluation of key model parameters]. Vestnik VGU. Seriya: Sistemnyy analiz i informatsionnyye tekhnologii [Bulletin of VSU. Series: System analysis and information technologies], 2024, no. 4, pp. 22-34.
5. Matveev N.A., Morgunov A.P. Avtomatizatsiya rascheta retseptury rezinnovoy smesi pri pomoshchi programmnogo obespecheniya [Automation of rubber compound recipe calculation using software]. Omskiy nauchnyy vestnik [Omsk scientific bulletin], 2016, no. 5 (149), pp. 55-58.
6. Rybanov A.A., Kablov V.F. Primeneniye algoritmov poiska assotsiativnykh pravil dlya opredeleniya tekhnologicheski znachimykh sochetaniy komponentov v retsepturakh rezinnovykh smesey [Application of association rule mining algorithms for identifying technologically significant combinations of components in rubber compound recipes]. Informatsionnyye i matematicheskiye tekhnologii v nauke i upravlenii [Information and mathematical technologies in science and management], 2025, no. 3 (39), pp. 177-188.
7. Rybanov A.A., Kablov V.F., Maslova M.A. Integratsiya sinonimicheskikh sistem v informatsionnyye sistemy upravleniya retseptami rezinotekhnicheskikh materialov [Integration of synonymic systems into information systems for managing rubber compound recipes]. Vestnik Cherepovetskogo gosudarstvennogo universiteta [Bulletin of Cherepovets state university], 2025, no. 5 (128), pp. 70-82.
8. Kablov V.F., Rybanov A.A., Keybal N.A. Avtomatizirovannyy bank dannykh novogo pokoleniya retseptur i svoystv rezin [Automated new-generation data bank of rubber formulations and properties]. Kauchuk i rezina [Rubber and rubber], 2024, vol. 83, no. 3, pp. 168-173.
9. Sohangir S., Wang D. Improved sqrt-cosine similarity measurement. Journal of Big Data, 2017, vol. 4, no. 25, pp. 1-13, DOI:10.1186/s40537-017-0083-6.
10. Chen F., Farahat A., Brants T. Multiple similarity measures and source-pair information in story link detection. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004). Boston, MA, USA, 2004, pp. 313-320.
11. Poroshkina V.V. Mery podobiya v rekomendatel'nykh sistemakh [Similarity measures in recommender systems]. Alleya nauki [Alley of science], 2019, vol. 1, no. 4 (31), pp. 909-913.

12. Kuznetsov L.A. Universal'naya tekhnologiya otsenki blizosti informatsionnykh ob"yektov [Universal technology for estimating the proximity of information objects]. Informatika i yeye primeneniya [Informatics and its applications], 2014, vol. 8, no. 2, pp. 130-144.
13. Gorokhovatskaya N.V. Metricheskaya klassifikatsiya s ispol'zovaniyem metoda blizhayshikh sosedey na osnove opisaniy v vide mnozhestv priznakov [Metric classification using the nearest neighbor method based on descriptions in the form of feature sets]. PhD dissertation. Moscow, 2012, 145 p.
14. Kablov V.F., Rybanov A.A., Keybal N.A. et al. Baza dannykh retseptur, ingredientov i svoystv rezinnovykh smesey s uchetom tekhnicheskikh sinonimov [Database of recipes, ingredients and properties of rubber compounds taking into account technical synonyms]. State registration certificate of database No. 2025621312 Russian Federation: No. 2025620796: appl. 03.12.2025: publ. 03.24.2025; rightholder VPI (branch) VolgSTU.

Alexander Aleksandrovich Rybanov. *PhD in Engineering, Associate Professor, Head of the Department of "Informatics and Programming Technology", Volzhsky Polytechnic Institute (branch) of Volgograd State Technical University. AuthorID: 234968, SPIN: 9256-6050, ORCID: 0000-0002-8638-9998, rybanoff@yandex.ru. 404111, Volgograd Region, Volzhsky, Lenin Ave, 72.*

Victor Fedorovich Kablov. *Doctor in Engineering, Professor, Department of "Chemical Technology of Polymers and Industrial Ecology", Volzhsky Polytechnic Institute (branch) of Volgograd State Technical University. AuthorID: 115782, SPIN: 1093-9003, ORCID: 0000-0002-2970-6109, vkablov5@gmail.com. 404111, Volgograd Region, Volzhsky, Lenin Ave, 72.*

Статья поступила в редакцию 18.12.2025; одобрена после рецензирования 10.02.2026; принята к публикации 15.05.2026.

The article was submitted 12/18/2025; approved after reviewing 02/10/2026; accepted for publication 05/15/2026.