УДК 004.021

С.И. БЕЗДЕНЕЖНЫХ

ОЦЕНКА СХОДСТВА ТЕХНОЛОГИЙ С ПРИМЕНЕНИЕМ ТЕХНИКИ СИНТАКСИЧЕСКИХ *m*-ГРАММ

Для оценки сходства технологий предложено использовать методы теории анализа данных. На примере вычисления редакционного расстояния Хэмминга показана работоспособность выбранного подхода. В качестве меры оценки сходства технологий предложено использовать синтаксические m-граммы. Обоснован вид критерия значимости m-граммы в функции оценки.

Ключевые слова: инновации; разработка технологий; теория информации; меры сходства; m-граммы.

Методы выявления аналогий и прецедентов в технических системах широко применяются в экономике. В частности, они используются в методиках оценки стоимости и продолжительности опытно-конструкторских работ по созданию высокотехнологичной продукции [1]. Большинство таких методик основано на использовании данных о стоимости и времени проведения так называемой базовой работы.

Под базовой работой понимается завершенная опытно-конструкторская работа (ОКР), в результате выполнения которой был создан образец, являющийся аналогом (имеющим, обыкновенно, более низкие характеристики) для образца, создаваемого в рамках планируемой ОКР.

Под *образцом-аналогом* понимают образец изделия (отечественный или зарубежный) одного функционального назначения с планируемым к созданию образцом, все или основные характеристики которого в результате выполнения опытно-конструкторской работы планируется изменить.

Основу методического аппарата для определения образца-аналога составляют экспертные методы. В группах изделий с малым количеством возможных альтернатив эта работа не вызывает у экспертов сложностей, однако в больших группах, насчитывающих десятки экземпляров, определение наиболее близкого аналога может вызвать затруднение.

Экспертным методам присущи все достоинства и недостатки экспертных оценок: достоверность и надежность результатов таких исследований сильно зависят от компетентности экспертов, их количества и квалификации. Результаты, полученные разными группами экспертов, несравнимы между собой и, соответственно, не могут использоваться в разных исследованиях.

Другим актуальным вопросом, на который не могут объективно ответить экспертные методы, остается определение степени сходства или различия изделий-аналогов.

Учитывая недостатки существующих методов, в данной статье рассматривается квалиметрический подход к оценке степени сходства изделий и технологий. За отправную точку взята предложенная ранее в [2] идея о том, что технологии можно описать с помощью терминов теории информации. Из чего следует, что для анализа технологий можно использовать методы теории анализа данных [3].

Понятие *технология* далее трактуется как совокупность документированных знаний для создания технических изделий и систем [4]. Она декомпозируется на составные технологии, каждая из которых может быть представлена набором технологических альтернатив. Далее используются следующие понятия:

алфавит G_t в момент времени *t* – это конечное множество доступных технологий $G = \{\tau_1, \tau_2, \dots, \tau_K\}$, которые могут быть использованы при создании изделия;

слово T – конечная последовательность $\{\tau_1, \tau_2, \dots, \tau_N\}, T \subset G_t$ суммарной длины *N*, представляющая технологическую модель изделия.

1. Оценка сходства технологий с использованием редакционного расстояния Хэмминга

В теории анализа данных для оценки сходства изучаемых объектов часто применяются алгоритмы вычисления *меры сходства строк.* Меры сходства являются разновидностью мер близости и выражаются функциями s от элементов скалярного произведения множества X на себя $(X \times X)$. При этом значения функции $s \in R$ должны удовлетворять требованиям неотрицательности $(s \geqslant 0)$ и симметричности $(s(x, y) = s(y, x))$ [5].

На практике в качестве мер сходства зачастую используются величины из диапазона [0, 1], где 1 – полное сходство, а 0 – полное несходство.

Одной из таких универсальных мер близости многомерных объектов одинаковой размерности является расстояние Хэмминга d_H:

$$d_H(T_1, T_2) = \sum_{m=1}^{N} |a_{1_m} - a_{2_m}|, \quad (1)$$

где T_1, T_2 – многомерные объекты, характеризующиеся N числовыми параметрами.

Смысл меры Хэмминга заключается в том, что в многомерном пространстве признаков два объекта тем ближе, чем по меньшему количеству признаков (параметров) они различаются. В случае двоичного характера признаков в качестве меры сходства можно использовать величину μ_H, определяемую на основе расстояния Хэмминга по следующему соотношению:

$$\mu_H = 1 - \frac{d_H(T_1, T_2)}{N}. \quad (2)$$

Технологические модели являются объектами, составные элементы которых в соответствующих позициях можно рассматривать как признаки нечисловой природы. Для применения меры Хэмминга к таким объектам в выражении (1) числовую операцию $|a_{i_m} - a_{j_m}|$ необходимо заменить двоичной функцией сравнения:

$$\delta(a_{i_m}, a_{j_m}) = \begin{cases} 0, \text{если } a_{i_m} = a_{j_m} \\ 1, \text{если } a_{i_m} \neq a_{j_m} \end{cases}. \quad (3)$$

В результате мера сходства μ_H конечных последовательностей $T_i = \{\tau_{i1}, \tau_{i2}, \dots, \tau_{iN}\}$ и $T_j = \{\tau_{j1}, \tau_{j2}, \dots, \tau_{jN}\}$ по Хэммингу вычисляется на основе следующего выражения:

$$\mu_H = 1 - \frac{1}{N} \sum_{m=1}^{N} \delta(a_{i_m}, a_{j_m}). \quad (4)$$

Проиллюстрируем предложенную меру на примере. Для этого рассмотрим множество беспилотных летательных аппаратов (БпЛА), представленных в таблице 1, и сравним несколько моделей между собой.

В соответствии с (4), значения меры сходства моделей БпЛА составят:

μ_H(Модель 1, Модель 2) = 1 – 0,2 * (0 + 0 + 1 + 0 + 0) = 0,8;

μ_H(Модель 7, Модель 8) = 1 – 0,2 * (0 + 0 + 0 + 1 + 0) = 0,8;

μ_H (Модель 1, Модель 10) = 1 – 0,2 * (1 + 1 + 0 + 1 + 1) = 0,2;

μ_H(Модель 2, Модель 3) = 1 – 0,2 * (0 + 0 + 0 + 0 + 0) = 1;

μ_H (Модель 2, Модель 10) = 1 – 0,2 * (1 + 1 + 1 + 1 + 1) = 0.

Из приведенного примера следует, что БпЛА моделей 1 и 2 очень похожи. И действительно, различие у них только в типе навигационной системы. БпЛА моделей 2 и 3 являются полными аналогами. Модели 1 и 10 различаются кардинально – у них сходство только в типе навигационной системы.

Таблица 1 – Множество БпЛА ближнего действия

Изделие	Технологическая модель БпЛА				
	Схема БпЛА	САУ	Навигация	Движитель	Посадка
Модель 1	квадрокоптер (1)	Naza 2M (1)	инерциал. (1)	электр.(1)	верт. (1)
Модель 2	квадрокоптер (1)	Naza 2M (1)	СНС (2)	электр.(1)	верт. (1)
Модель 3	квадрокоптер (1)	Naza 2M (1)	СНС (2)	электр.(1)	верт. (1)
Модель 4	конвертоплан (2)	Ardupilot (2)	СНС (2)	бензин. (2)	верт. (1)
Модель 5	конвертоплан (2)	PixHawk (3)	СНС (2)	электр.(1)	верт. (1)
Модель 6	самолет (3)	Ardupilot (2)	СНС (2)	электр.(1)	парашют (2)
Модель 7	самолет (3)	Ardupilot (2)	СНС (2)	электр.(1)	парашют (2)
Модель 8	самолет (3)	Ardupilot (2)	СНС (2)	бензин. (2)	парашют (2)
Модель 9	самолет (3)	PixHawk (3)	СНС (2)	бензин. (2)	парашют (3)
Модель 10	самолет (3)	PixHawk (3)	инерциал. (1)	водород. (3)	самолет. (3)

2. Оценка сходства технологий с использованием техники *m*-грамм

Приведенный пример упрощен, так как сходство технологий не может определяться только количеством совпадений технологических альтернатив в группах. Почти всегда новое изделие имеет свои технологические решения и полное совпадение технологий в технологических группах для сложных изделий скорее исключение. Тем не менее пример демонстрирует работоспособность предложенного подхода.

Рассмотренное расстояние Хэмминга ограничено исключительно операцией замены, поэтому оно применяется только для объектов одинаковой размерности. Существуют другие разновидности меры измерения расстояния, которые рассчитываются с использованием иного набора допустимых операций редактирования и допускают различную размерность объектов. Наиболее известные из них: расстояние Левенштейна, Дамерау-Левенштейна, наибольшая общая подпоследовательность, сходство Джаро-Винклера.

Общим недостатком перечисленных методов при решении задачи определения сходства технологий является их ориентированность на линейную структуру сравниваемых данных, состоящих из одного алфавита. В то время как модель, приведенная в [4], рассматривает каждую технологию рекуррентно как составную структурированную функционально-технологическо-физическую модель. При этом технологическая составляющая модели включает функционально-технологическо-физические модели составных частей. Таким образом, технологическая модель представляется в виде упорядоченного ориентированного графа (дерева), в котором составные технологии локализуются в технологических группах и фактически формируют обособленный алфавит.

Учитывая эти особенности, для вычисления меры сходства технологий можно прибегнуть к алгоритмам синтаксических *m*-грамм[1] [5-8]. *M*-граммами называют сочетания из *m* смежных символов. Синтаксические *m*-граммы – это *m*-граммы, определяемые путями в деревьях синтаксических зависимостей или деревьях составляющих объектов, а не линейной структурой текста.

Синтаксические *m*-граммы нашли обширное применение в методах выявления плагиата, установления авторства текстов, успешно используются для категоризации текста и языка. В области биоинформатики *m*-граммы используются для поиска генетических последовательностей и определения того, с каких конкретных видов животных собраны образцы ДНК. Кроме того, их используют для создания функций, которые позволяют получать знания из текстовых данных.

В качестве меры сходства строк T_1 и T_2в технике *m*-грамм используется величина μ_m:

$$\mu_m = \frac{2m(T_1,T_2)}{m(T_1)+m(T_2)}, \quad (5)$$

где $m(T_1)$ и $m(T_2)$ – количество m-грамм в строках T_1 и T_2 соответственно; $m(T_1,T_2)$ – количество *m*-грамм, одновременно входящих в строку T_1 и в строку T_2, независимо от позиции расположения.

Смысл меры на основе *m*-грамм заключается в том, что сходство строк символов (объектов) рассматривается в контексте близости их

[1] В литературе используются также обозначения *N*-граммы или *q*-граммы.

лексического значения, обусловленного некоторым ядром (коллокацией) в форме подпоследовательности из *m*-символов, которая:

а) может сдвигаться по позициям ввиду морфологических различий;

б) «размываться» с учётом определённых особенностей произношения, правописания и т.п.

В результате на основе *m*-грамм можно анализировать сходство строк как последовательностей символов (объектов) разной размерности. Кроме того, сходство строк, как уже было отмечено, определяется не по совпадению *m*-грамм, начинающихся в одинаковых позициях, а как одновременное вхождение *m*-грамм в объекты сравнения независимо от их позиций в строках сравнения.

Выбор значений *m* является центральным вопросом при формировании меры μ_m и обычно осуществляется на основе лексических или семантических соображений [5].

Семантические значения текстов определяются отдельными словами и их сочетаниями. Поэтому при анализе текстов на практике ограничиваются анализом только по одному значению *m*, в большинстве случаев по двуграммам (*m* = 2) или реже по триграммам (*m* = 3).

В случае сравнения технологий, когда природа элементов последовательностей является произвольной, выбор значений *m* является неопределённым. В этом случае можно выделить следующие критерии сходства последовательностей:

количественный – сходство тем больше, чем в большем количестве позиций совпадают элементы;

качественный – сходство тем больше, чем больше совпадений элементов в смежных позициях.

Следует отметить, что по смыслу совпадения совокупности смежных элементов одна совпадающая *m*-грамма должна быть отделена от другой совпадающей *m*-граммы минимум одной позицией, в которой элементы сравниваемых последовательностей не совпадают. В противном случае имеет место совпадение одной *m*-граммы, в которой число *m* определяется количеством подряд следующих позиций совпавших элементов.

С точки зрения оценки сходства технологий, номера позиций, с которых начинаются совпадающие *m*-граммы (в начале последовательности, в середине или в конце), не имеют значения. Поэтому совпадения *m*-

граммы, начинающиеся, например, с 1-й позиции или со 2-й позиции, или с 5-й позиции и т.д., рассматриваются как один и тот же случай сходства.

В результате для определённого значения размерности сравниваемых последовательностей можно построить ряд вариантов совпадений элементов, при которых сходство последовательностей должно возрастать в соответствии с ростом количества и качества совпадений. При этом увеличение качества совпадений трактуется как появление хотя бы одной совпавшей *m*-граммы, размерность которой (*m*) на единицу выше самой старшей *m*-граммы в предыдущем варианте. Такой подход к определению сходства последовательностей называется *количественно-качественным с приоритетом количества совпадений* [5].

Количество совпадений элементов определяется суммой произведений значений n_m (количество совпавших *m*-грамм) на число *m* (количество элементов в *m*-грамме) – $\sum_{m=1}^{N} mn_m$. Каждое слагаемое mn_m даёт вклад в общее количество совпадений элементов, реализованное совпадением соответствующих *m*-грамм. Как следует из вышеприведённых критериев близости, сходство последовательностей должно быть тем выше, чем более «старшими» *m*-граммами (с большими значениями *m*) оно реализовано. Тогда одним из подходов к установлению меры сходства последовательностей может быть «взвешивание» слагаемых mn_m в зависимости от размерности *m*-грамм.

В результате получается следующая мера сходства *μ*, зависящая от количества и качества совпадений элементов конечных последовательностей:

$$\mu = \frac{1}{N}\sum_{m=1}^{N} mn_m c_m \qquad (6)$$

где n_m – количество совпадений *m*-грамм; *N* – максимальное количество элементов в сравниваемых последовательностях; c_m – вес значимости совпадения *m*-граммы в сходстве последовательностей, $c_m \leqslant 1$.

В [5] доказано, что величина *μ* удовлетворяет требованиям, предъявляемым к мерам сходства. В частности величина, определяемая по формуле (6), является неотрицательной в диапазоне [0, 1] и обладает свойством симметричности.

3. Выбор коэффициента значимости *m*-граммы

Характер поведения меры сходства объектов в формуле (6) определяется коэффициентом значимости c_m. Определяющая коэффициент функция обусловливается спецификой природы анализируемых последовательностей, их элементов, а также особенностями исследовательских задач. Чтобы определить наиболее рациональный вариант этой функции для оценки сходства технологий рассмотрим возможные подходы.

Первый из них – это возрастание коэффициента значимости от соотношения $\frac{m}{N}$. Действительно, чем большую часть последовательностей составляет совпадающая *m*-грамма, тем более значимым должен быть ее вклад в сходство. Таким образом, для определения коэффициента c_m необходимо задать некоторую функцию от $\frac{m}{N}$.

При $c_m = 1$ мера сходства *μ* ступенчато возрастает пропорционально общему количеству совпадений, не различая их разное качество по однограммам, двуграммам, триграммам и т.д. Интересно, что полученные таким образом значения меры сходства будут соответствовать мере Хэмминга.

Простая пропорциональная зависимость весов значимости c_m от $\frac{m}{N}$ даёт нелинейно и немонотонно возрастающую картину повышения сходства последовательностей в зависимости от количества и качества совпадений. Так, в точках перехода «количества в качество» сходство последовательностей уменьшается вопреки росту количества совпавших элементов. Такое поведение меры *μ* в соответствующих случаях отражает приоритет качества совпадений.

На рисунке 1 приведены графики *μ* по другим видам функции c_m, которые также демонстрируют приоритет качества совпадений элементов в сходстве последовательностей.

Другим подходом к установлению весов значимости c_m может быть учёт максимального количества $max_m(N)$ возможных совпадений по конкретной *m*-грамме в рамках определённой размерности сравниваемых последовательностей.

Предполагается, что вес *m*-граммы должен быть тем больше, чем меньше совпадений *m*-грамм может реализоваться в пределах

N-элементов последовательности, т.е. чем меньше $max_m(N)$. Например, вес совпадения двуграммы $c_2(N)$ в последовательности из 3-х или 4-х элементов должен быть выше, чем вес совпадения двуграммы $c_2(N)$ при *N* = 5, *N* = 6, *N* = 7, поскольку при *N* = 3 и *N* = 4 совпадение одной двуграммы реализует весь набор случаев сходства последовательностей по совпадению двуграмм, а при *N* = 5, *N* = 6, *N* = 7 – только один из двух возможных случаев сходства по совпадениям двуграмм.

На рисунке 2 представлены расчёты коэффициента сходства последовательностей из 10 элементов (*N* = 10) при различных видах функции $c_m = f(\frac{1}{max_m(N)})$. Как видно из приведённых графиков, использование величин $max_m(N)$ также реализует принцип приоритетности качества совпадений, но с другой спецификой «переходов» меры сходства при изменениях количества и качества совпадений. Например, при использовании для c_m функций с аргументом $\frac{1}{max_m(N)}$ существенно увеличиваются «броски» меры сходства *μ* в точках при переходе от варианта совпадения по одной *m*-грамме к варианту с $(m+1)$ совпадений *m*-грамм.

Конечные последовательности описания технологии имеют схожую, но не одинаковую размерность. В соответствии с эволюционно-технологической теорией [4] их можно рассматривать как случайные реализации некоторой исходной закономерности следования элементов (общей модели). В результате фрагменты исходной закономерности условно фиксированы по месту и анализ сходства можно вести по совпадению элементов или их совокупностей (*m*-грамм) в одинаковых позициях.

Очевидно, что в случае сравнения технологий одинаковой размерности большее соответствие должны все-таки иметь технологии с большим числом совпадений независимо от их качества. Однако вес *m*-граммы должен быть тем больше, чем меньше совпадений *m*-грамм может реализоваться в пределах технологической модели. То есть из двух технологий-альтернатив с одинаковым количеством элементов и равным количеством совпадающих технологий более близкой должна считаться та, у которой наибольшая длина совпадающей последовательности (выше качество). Одновременно при выборе из альтернатив разной длины и одинаковом количестве совпавших элементов должно быть обеспечено уменьшение критерия соответствия для технологии с большей размерностью технологической модели.

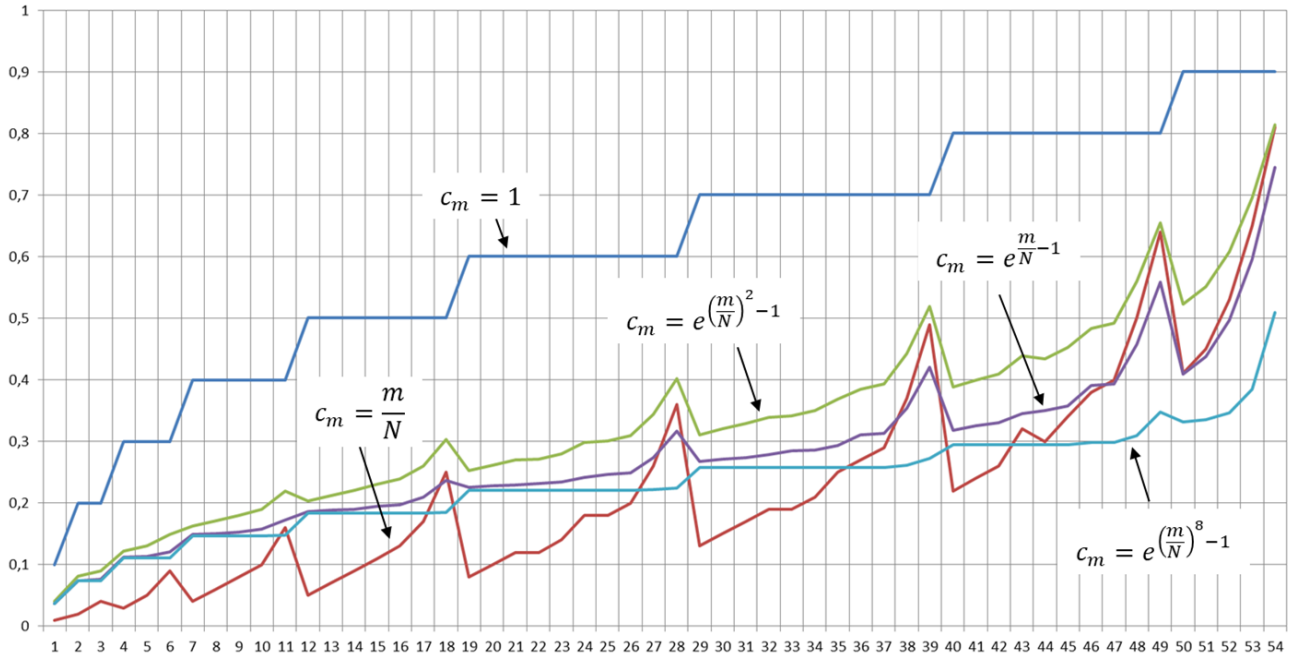

Рисунок 1 – Зависимость коэффициента сходства последовательностей при N = 10 от количества и качества совпадений при различных видах функции c_m

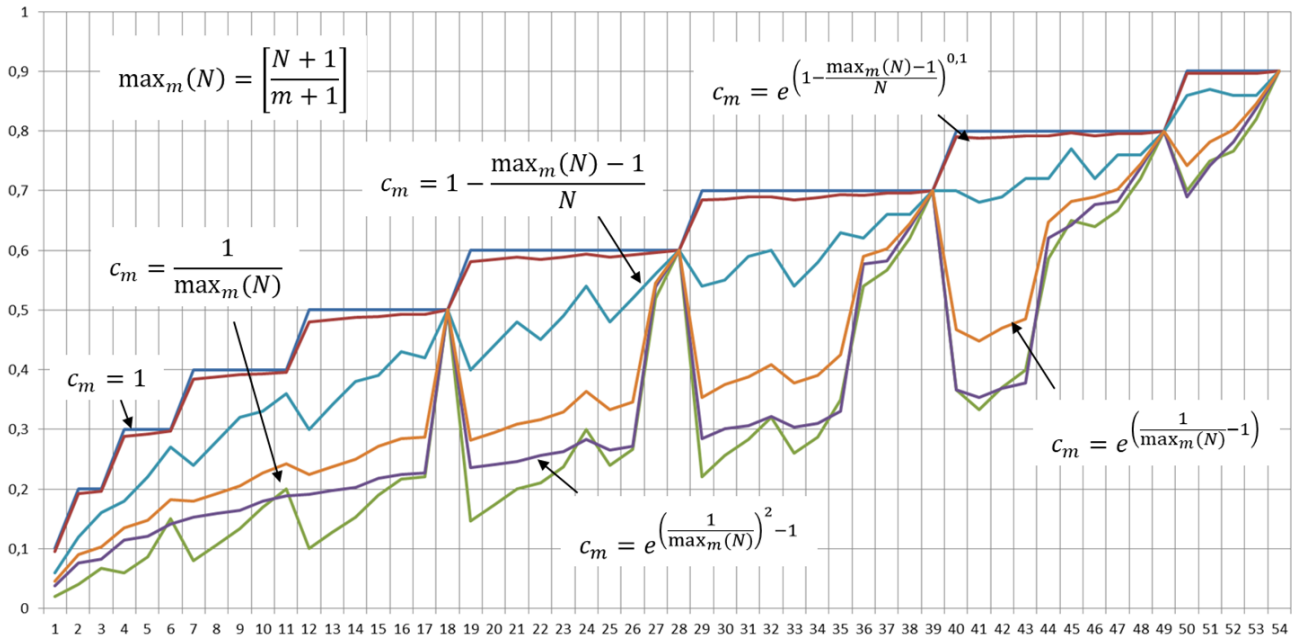

Рисунок 2 – Зависимость коэффициента сходства последовательностей от количества и качества совпадений при различных видах функции c_m (N=10)

Таким образом, из рассмотренных видов функций c_m для оценки сходства технологий подходит функция $c_m = 1 - \frac{max_m(N)-1}{N}$, которая учитывает предложенные критерии. В результате выражение оценки сходства технологий (6) будет иметь вид:

$$\mu = \frac{1}{N}\sum_{m=1}^{N} mn_m(1 - \frac{max_m(N)-1}{N}), \quad (7)$$

где n_m – количество совпадений *m*-грамм, *N* – максимальное количество элементов в сравниваемых последовательностях, а $max_m(N)$ – максимальное количество *m*-грамм в последовательности.

В [5] показано, что значение $max_m(N)$ определяется как целая часть отношения:

$$max_m(N) = \left[\frac{N+1}{m+1}\right]. \quad (8)$$

Пересчитаем приведенный ранее пример схожести БпЛА с использованием предложенной метрики *m*-грамм (7):

μ(Модель 1, Модель 2) = 0,2 * (2 * 2 * 0,8) = 0,64;

μ(Модель 7, Модель 8) = 0,2 * ((1 * 3 * 0,6) + (3 * 1 * 1)) = 0,72;

μ(Модель 1, Модель 10) = 0,2 * (1 * 1 * 0,6) = 0,2;

μ(Модель 2, Модель 3) = 0,2 * (5 * 1 * 1) = 1;

μ(Модель 2, Модель 10) = 0,2 * 0 = 0.

Из расчетов видно, что результаты сравнения моделей 1 и 10, моделей 2 и 3, а также моделей 2 и 10 полностью совпали с мерой Хемминга из первого примера. В то время как сравнение моделей 1 и 2, а также моделей 7 и 8 дало отличный от предыдущего результат. Более того, из нового примера видно, что различие моделей в первом сравнении больше, чем во втором, тогда как в прошлый раз они были равны. Эта разница обусловлена наличием качественного отличия: при сравнении моделей 1 и 2 было найдено только две двуграммы, а при сравнении моделей 7 и 8 – одна триграмма и еще одна однограмма.

4. Учет позиции *m*-граммы в иерархии модели технологии

Рассмотренные до этого примеры оперировали линейными структурами. Однако, как уже отмечалось, технологическая модель описывается при помощи упорядоченного ориентированного ацикличного графа

(дерева), в котором составные технологии локализуются в технологических группах. Преимущество техники *m*-грамм состоит как раз в том, что она позволяет отойти от линейной структуры и сравнить объекты, представленные деревьями.

Рассмотрим как это происходит на еще одном примере. Рассчитаем сходство двух объектов, представленных на рисунке 3. Первый из них включает 13 элементов, второй – 12. Осмотр этих орграфов позволяет выявить одну пятиграмму, одну триграмму и одну однограмму. Таким образом, в соответствии с выражением (7), получается:

μ(БпЛА$_1$, БпЛА$_2$) = 0,0833 * ((1 * 1 * 0,5833) + (3 * 1 * 0,8333) + (5 * 1 * 0,9166)) = 0,6389.

Несмотря на то, что результат выглядит вполне правдоподобно, в предложенном выше выражении (7) не учитывается одна важная особенность технологической модели: приоритет совпадения технологической модели высшего уровня перед совпадением составными технологиями низших уровней.

Для учета этой особенности подход к установлению весов значимости c_m должен быть усовершенствован и должен учитывать положение *m*-граммы в иерархии возможных совпадений по конкретной *m*-грамме в рамках иерархии сравниваемых деревьев. Другими словами, совпадение *m*-грамм, включающих элементы верхнего уровня, должно давать больший вклад в меру соответствия, чем совпадение *m*-грамм нижних уровней. Назовем такой подход *количественно-качественным с приоритетом количества совпадений и учетом структуры.*

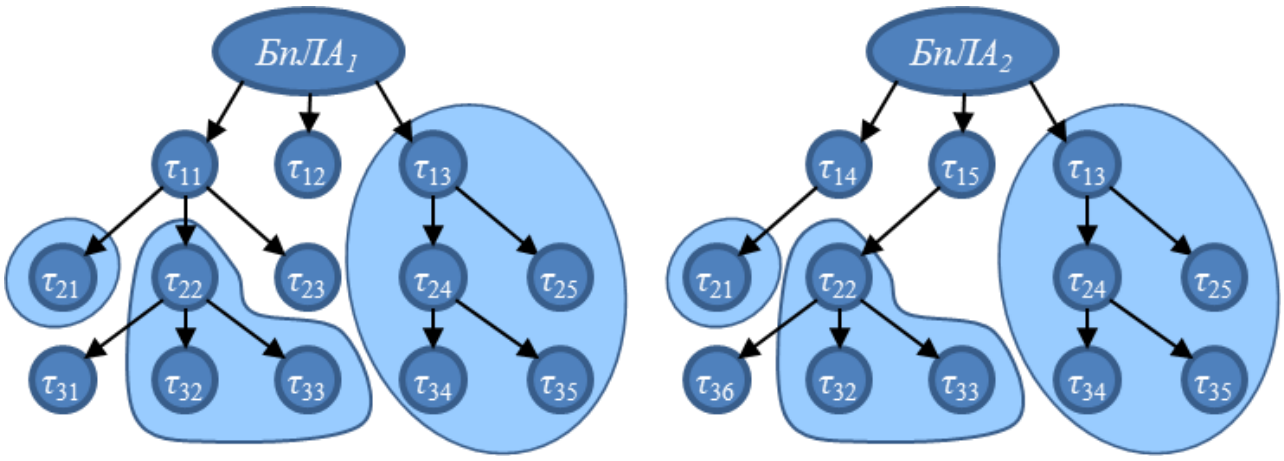

Рисунок 3 – Сравнение двух технологических моделей БпЛА

Основываясь на доводах, приведенных ранее, модифицируем выражение (7), дополнительно поставив возрастание коэффициента c_m в зависимость от соотношения $\frac{1}{l_m}$. То есть чем выше в иерархии дерева стоит *m*-грамма, тем выше должен быть ее вес. При этом глубина дерева значения не имеет.

С учетом возможной разницы глубины деревьев и места локации *m*-граммы в сравниваемых деревьях, при подсчете меры сходства необходимо учитывать узлы с наибольшей высотой в сравниваемых деревьев. Таким образом, выражение (7) приобретает вид:

$$\mu = \frac{1}{N}\sum_{m=1}^{N} mn_m\left(1 - \frac{max_m(N)-1}{N}\right)\frac{1}{l_m}, \qquad (9)$$

где n_m – количество совпадений *m*-грамм; *N* – максимальное количество элементов в сравниваемых последовательностях, $max_m(N)$ – максимальное количество *m*-грамм в последовательности; а l_m – высота *m*-граммы в дереве последовательности.

График на рисунке 4 показывает результаты расчётов значения коэффициента сходства деревьев последовательностей из 10 элементов (*N* = 10) для функции $c_m = 1 - \frac{max_m(N)-1}{N}$ при разной высоте положения *m*-граммы (от 1 до 9). Для удобства график развернут пологой стороной к наблюдателю, при этом чем ближе значения, тем ниже уровень *m*-граммы в дереве технологической модели.

Пересчитаем последний пример с использованием выражения (9):

$$\mu(\text{БпЛА}_1, \text{БпЛА}_2) = 0{,}0833 * ((1 * 1 * 0{,}5833 * 0{,}5) + (3 * 1 * 0{,}8333 * 0{,}5) + (5 * 1 * 0{,}9166 * 1)) = 0{,}5104.$$

Как видно, результат оказался меньше предыдущего. Это обусловлено тем, что две третьи технологий верхнего уровня в сравниваемых образцах не совпали.

Математически предложенный подход допускает взаимное равенство трех и более разных объектов, если в них присутствует баланс совпадения участков технологической модели к их размещению в дереве модели. Однако на практике такую ситуацию можно исключить, если учесть, что каждая технология фактически определяется своей технологической моделью и опирается на свой специфический алфавит. То есть ситуация, когда одни и те же составные технологии порождают разную технологию верхнего уровня, считается некорректной.

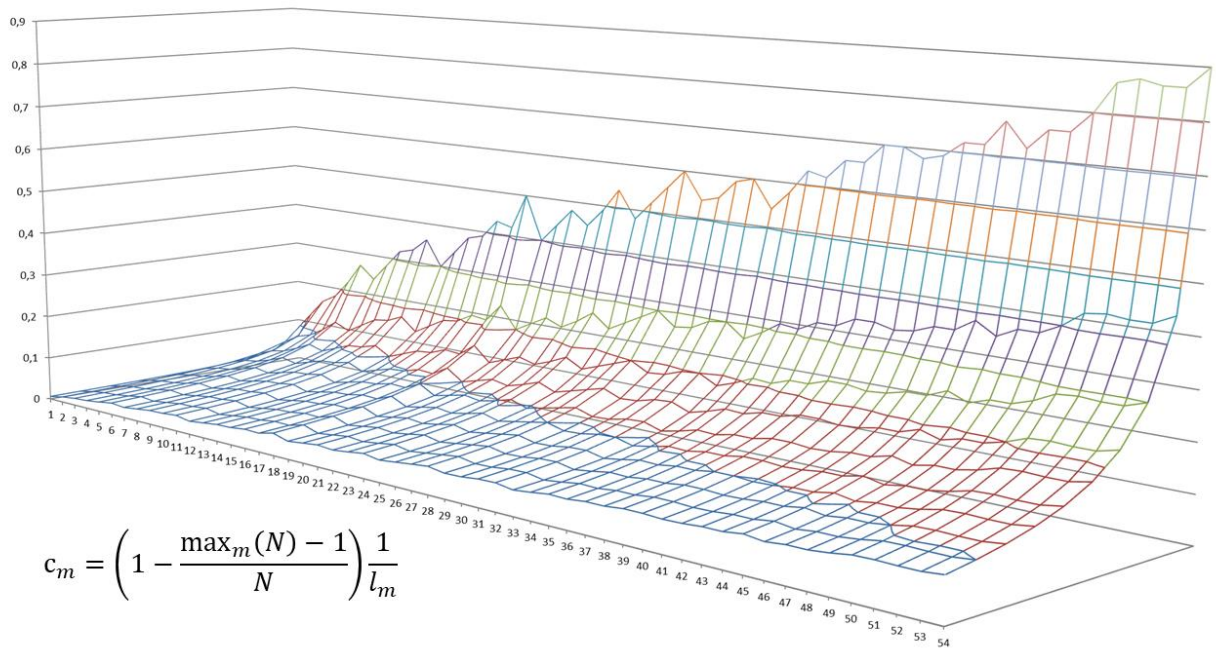

Рисунок 4 – Зависимость коэффициента сходства последовательностей от количества и качества совпадений функции c_m (*N*=10) для *m*-грамм, располагающихся на разной высоте (1…9) дерева технологической модели

Таким образом, расчеты, проведенные с использованием редакционного расстояния Хэмминга, демонстрируют возможность применения информационного подхода к решению задачи определения сходства технологий. Однако ограниченность мер, основанных на вычислении редакционного расстояния, и особенности технологической модели не позволяют применять их в реальных приложениях. Поэтому в качестве меры сходства технологий предложено использовать технику синтаксических *m*-грамм. Для выбранного подхода предложен коэффициент значимости c_m, учитывающий при сравнении количественные, качественные и структурные особенности сравниваемых технологических моделей.

Предложенный подход поможет более объективно ответить на вопрос о степени сходства технологий, осуществлять поиск аналогов независимыми группами исследователей, а также откроет путь к разработке новых методов оценки новизны и инновационности технологий.

Развитием предложенного исследования может быть создание алгоритма выделения *m*-грамм на технологической модели за оптимальное время. Существует ряд эффективных алгоритмов, решающих эту

задачу на линейной последовательности (Лемпеля-Зива-Велча, суффиксный массив, суффиксное дерево и др.), однако ни один из них не подходит для выделения *m*-грамм на дереве.

Кроме того, за последние семь лет существенный шаг вперед сделали методы интеллектуального поиска и определения сходства текстов. Появились такие инструменты как word2vec, GloVe, doc2vec, sent2vec, которые демонстрируют значительные практические результаты при обработке текстов на естественном языке. Основной заложенной в них идеей является представление текста как вектора в многомерном пространстве [9; 10]. Учитывая полученные ими результаты, актуально изучение возможности представления технологии подобным образом.

Список использованных источников

1. Буренок В.М., Лавринов Г.А., Подольский А.Г. Оценка стоимостных показателей высокотехнологичной продукции. М.: Издательская группа «Граница», 2012. 424 с.

2. Безденежных С.И., Брайткрайц С.Г. Информационный подход к оценке сложности и потенциала развития технологии // Вооружение и экономика. 2018. №4(46). С. 8-14.

3. Барсегян А.А. Технологии анализа данных: Data Mining, Text Mining, OLAP. СПб.: БХВ-Петербург, 2007. 384 с.

4. Буренок В.М., Ивлев А.А., Корчак В.Ю. Развитие технологий XXI века: проблемы, планирование, реализация. Тверь: ООО «Купол», 2009.

5. Гайдамакин Н.А. Мера сходства последовательностей одинаковой размерности // Математические структуры и моделирование. 2016. №4(40). С. 5-16.

6. Будников Е.А. Обзор некоторых статистических моделей естественных языков // Машинное обучение и анализ данных. 2011. №2. С. 243-248.

7. Андреева А.Г., Маркина Т.А. Оценка подобия деревьев с помощью вычисления pq-грамм расстояния // Научно-технический вестник информационных технологий, механики и оптики. 2017. №3. С. 490-497.

8. Sidorov G. Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction // International Journal of Computational Linguistics and Applications. 2013. №2. P. 169-188.

9. Николаенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2018. 480 с.

10. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR. 2013.